A Computational Approach

To

Identify Covertness and Collusion

in Social Networks

A Dissertation

Submitted to the Faculty of the Graduate School

at

University of Minnesota

by

Pronab Mohanty

In Partial Fulfilment of the Requirements

for the Degree of

Doctor of Philosophy

Professor Jaideep Srivastava

and

Professor David Knoke

October 2020

Abstract

Mathematical and computational interventions in the field of social networks have a fairly recent history. Social networks analysis exists at the intersection of several fields, including social sciences, psychology, organizational behavior, business studies, mathematics, physics, and biology. Studies were often manually facilitated in the last century as the social networks' sizes were typically small. But, the recent emergence of the internet, the world wide web, big data, and numerous platforms of social media have triggered a period of intense academic activities in this field, which is also true in the field of criminology where advances in social network analytics have engendered a flourishing sub-culture that has influenced enforcement techniques spawning new fields such as predictive policing, investigation techniques specifically based on network analytics, and even studies of criminal behavior patterns.

Interest in studying criminal and terrorist networks, generally called covert networks, has peaked after recent attacks by terror organizations. There is a felt necessity of presaging criminal or covert activities well before they erupt into public consciousness. However, recent research has been reactive rather than proactive and has essentially focused on analyzing illegal networks unearthed, and the accent is on disrupting such networks. Relatively little focus has centered on the question of why some networks are termed covert or, indeed, if covertness is innate to all networks., which further leads to the related issues of identifying metrics to measure the characteristics that typify covertness and to detect the presence of covert communities in social networks leveraging the metric so developed.

A further challenge is an increasing emphasis on privacy rights, data protection measures, and exponential growth in encryption

measures, which has placed a ceiling limit on the information obtained on communications. Added to this aspect is the vast volumes of data that need to be processed, requiring commensurately vast use of computational resources, often with very little time.

These aspects have been comprehensively addressed by the dissertation, which has used the ENRON email corpus to identify the employees who had been connected with the financial fraud in some manner. The research seeks to identify covertness within networks without any intrusive analysis or content-based measures, which is necessary given the increasing legal and policy constraints based around privacy, encryption, and general exclusion of personal data from the public domain, and also by reducing the size of the problem. The dissertation also develops specific metrics to define covertness in communications among network entities and defines a separate metric to identify covert entities' clusters with common aims. In the process of defining metrics, the dissertation also seeks to solve the problem of resource-constraints common in law-enforcement agencies by reducing the volume of information to be processed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Ever since John Guare's famous play (and later film) *Six Degrees of Separation*[1] burst into public consciousness, there has been an explosion of public interest in social networks[2]. Alongside the laity, academics and researchers have also invested huge efforts and resources in analyzing social networks. Adding to the momentum has been the exponential growth on the internet and its spin-off social web networks and online chat forums like Facebook, MySpace, Twitter, Tik-Tok, Weibo, Whatsapp, and many others. Social network analysis typically occurs at the intersection of many disciplines right since its inception. Studies into social networks are usually traced to at least three disciplines: psychology, anthropology, and sociology (Knoke and Yang 2008, p *vii*). A parallel wave of interest in this area was generated through the mushrooming of research applications in basic and natural sciences. In the words of Knoke and Yang (2008, p2), "Network analysis became an institutionalized, transdisciplinary perspective whose basic concepts and measures are now widely familiar to researchers from such diverse fields as sociology, anthropology, economics, organization studies, business management, public health, information science, biology, complexity and chaos theory."

[1]Six degrees of separation is the concept that all people are six, or lesser, social links away from each other. It was originally set out as the small-world idea by Stanley Milgrom, a psychologist and popularized in an eponymous playwritten by John Guare.

[2] However, social scientists often suggest that modern social network analysis began with the publication in 1934 of Jacob L. Moreno's pioneering book on sociometry, Who Shall Survive? There are also other defining works on metrics that precede this seminal piece of literature, including those of J.C. Almack in 1922, B. Wellman in 1926,E. Chevaleva-Janovskaja, in 1927, R.M. Hubbard in 1929, E.P. Hagman in 1933. For a more detailed perspective refer to Freeman's brief paper "Some Antecedents of Social Network Analysis", 1996.

1.2 Social Network Analysis in Law Enforcement

One of the later domain additions has been the field of law enforcement, particularly the sub-area called predictive policing. Researchers' interest in this domain has focused on networks formed by terror groups and criminal organizations. The essential research question asked is, “can one predict the outcomes of criminal activities of malfeasant actors in such networks.” The social networks that are the subject of much research and debate are justifiably called ‘covert networks’ due to their inscrutability and resistance to network analysis's standard tools. Current interest in the study of networks indulging in criminal and clandestine activities has their roots in the seminal work done by Claire Sterling, whose book, *The Terror Network*, published in 1981, is considered a standard textbook in police training schools across the world. In her book, Sterling describes the relationships amongSoviet Secret Services and the Palestinian and Irish Republican Army terrorists in the 1970s. However, the one incident that spurred academic researchers to begin applying social network methods to covert networks was the infamous 9/11 Al-Qaida attacks inside the United States. Researchers like Valdis Krebs (2001, 2002) made painstaking efforts in the aftermath to construct meaningful networks of the attackers by piecing together data from newspaper reports, the electronic media, and investigation and prosecution related documents. The data so collected was then pipelined into creating adjacency matrices for making the data compatible with computer programs.

Apart from the 9/11 attacks, other recent threats posed by criminal and transnational clandestine organizations have generated a storm of interest, both academic and political, in covert networks. However, much of the work on such networks to date is based on simulations, theoretical models, and even speculation. Relatively little is based on empirical data, owing to the absence of real-time data (Rodriguez2009; Asal and Rethemyer2006). A crucial lacuna that exists to date in the study of covert networks is the lack of complete data and the opaqueness of the existing channels of transactions between the key players. Indeed, the social analysis of covert networks has come to recognize this key handicap as a structural and defining feature of such networks and has incorporated this feature in research.

1.3 The Concept of Covertness

A key aspect in covert networks studies is how the term *covert* must be quantified to be rendered into some tangible mathematical formula. There are suggestions that covert networks as an entity of research in computer science may be gradated rather than be described as a sharply defined artifact that is different from other social networks. There are differences between ties exhibiting covertness between nodes that, for example, are having extramarital liaisons than between actors who might be trading insider information about a company. And both these instances will vary from covert ties involving terrorists or spies. The costs of exposure vary from embarrassment to legal strictures to summary executions. The formatter thus needs to create these components, incorporating the applicable criteria that follow.

The second question that research needs to answer about covertness in networks is whether this feature applies to all social networks or is confined exclusively to criminal or terrorist networks and other under-the-radar networks labeled as 'covert networks'[3] by observers. A possible answer may be found in the landmark paper on conspiracy networks in the heavy electrical industry of the United States by Baker and Faulkner (1994, p 21) wherein they state thus – "this study is the first quantitative network analysis of intercorporate conspiracies. Most empirical research on organizational action sets has focused on legal activities. Sociologists acknowledge that action sets[4] can conduct illegal acts (Aldrich 1979, pp. 317, 320), but this promising research line has remained virtually unexplored. Our research shows that the study of illegal networks can yield important theoretical and substantive insights into inter-organizational behavior." Thus, in the sociological literature, we know that there are elements (action sets) within overt organizations that might be doing a covert or undesirable activity.

[3]What might be called 'covert' in one country may not be so in another. For example, spy-networks are unquestionably covert in their target countries, but not so in their parent countries.

[4]Illegal interorganizational networks studied here are organizational action sets. An action set is a coalition of organizations assembled for the purpose of carrying out specific activities. Action sets may be short-lived or long lived. Soma are disbanded after success or failure. (Knoke & pappi, 1991, p 510; Knoke & Burleigh, 1989).

Hence, when this study refers to covertness, it does so in the broader context of all social networks, not just the ones labeled as covert or dark or clandestine or by any other synonym. Thus, covertness is treated in this study as an attribute inherent to all social networks irrespective of their labeling. Such an approach has the advantage of measuring the covertness coefficients of communities or subgroups within a network and also acts as a marker for entire networks if the attribute is pervasive across the entire architecture.

1.4 Key Research Questions

As in any new study, this research seeks to answer six basic research questions:

(a) What is the nature of the problem?

(b)Why is the problem important?

(c)Why is the problem challenging? What are the existing methods to address these challenges? What are its limitations?

(d) What are the contributions?

(e) What is their novelty?

(f)How are the contributions better than state of the art?

The first four of these questions are briefly answered in the Introduction itself, and the last question on comparisons with existing methodologies is analyzed throughout this dissertation. Finally, a short comparative narrative is also made out in summary.

1.4.1 Nature of the Problem

The preceding sections have already shed light on the nature of the study's problem. The dissertation briefly seeks to survey what covert networks mean and their chief characteristics and how these properties differ from those of conventional networks. It also takes a detailed look at the existing methodologies in social network analysis and the modifications in these techniques to adapt to covert networks' special characteristics. The dissertation then seeks to answer whether covertness is a feature-based atomic attribute based on which entire subgroups with common aims and objectives within the overall

architecture of a network can be labeled as a collusive covert community. The dissertation also looks at this attribute's design from a standpoint that makes it mathematically malleable to construct other more complex properties and similarity measures. The conception, design, and application of the covertness metric in the manner and form deemed desirable is the prime motivation of this research. A further narrative on this proposed metric's nature is presented in one of the sections following the ones on motivations.

1.4.2 Importance of the Problem

The problem's importance stems from the urgent need for enforcement agencies, security forces, and even businesses to identify covert or clandestine networks, guess the activities' nature, and predict their outcomes within reasonable time constraints resources. This question converges with Krebs's (2002) observations expressed at the very beginning in his network analysis of the 9/11 attacks, "We were all shocked by the tragic events of September 11, 2001. In the non-stop stream of news and analysis, one phrase was constantly repeated and used in many contexts 'terrorist network'. Everyone talked about this concept and described it as amorphous, invisible, resilient, dispersed, and other terms that made it difficult to visualize what this structure looks like." This narrative captures the urgency of the issue and the need to find reasonable answers in the shortest possible time.

1.4.3 Challenges

The primary challenge lies in interpreting real-world issues from a computational or mathematical perspective. Social network analysis is now a full-fledged curriculum in computer science studies. Yet, the nuances of the parent social science domains from where social networks are derived are only just being understood from a mathematical or formulaic framework. It is trivial to expect that the mathematical and computational interpretations will only approximate the sociological constructs. As the research body grows, the gap between mathematical models and the real-world scenarios will continue to narrow. In most computational experiments on real-world social networks models, the results have been very encouraging and point to a bright future in the convergence of these

two very important study branches. But it may well be stated that every sociological issue is a unique one. The formulations to translate it into a computationally meaningful format presents its own unique set of challenges. The key to resolving this problem is to scrutinize each sociological construct from a micro-perspective and then build a computational superstructure on a foundation of smaller constructs. The second key step would be to apply existing computational models to the new problem at hand. Previous results and interpretations will act as viable inputs to the current research. This dissertation has(I have)strived to build micro-models of the sociological problem in hand and has (have) attempted to knit smaller and more basic solutions into a bigger narrative that results in a best-fit answer. In the process, this study has also relied on a wide array of existing mathematical and computational tools adopted in similar real-world problems to bring the proposed solution in convergence with established procedures.

The overall approach to structure the problem is outlined in Figure-1.1. First, I examine whether covertness in social networks can be interpreted in a way that makes sense mathematically. Then I define various sociological structures inherent to the problem from a mathematical standpoint. Lastly, I use the mathematical constructs so derived to evaluate the dataset computationally. It needs to be noted that the study of 'covert' or 'dark' networks (there are many synonyms to these kinds of networks) is an emergent domain even in the field of sociology. The studies have increased in recent times, partly in response to the necessity of the law enforcement machinery, making sense of terror-related attacks and the harmful effects of criminal activities, which are the outcomes of such networks. Bigger financial scams occurring due to collusion and conspiracies of smaller sub-groups which are inseparable parts of bigger and perfectly 'overt' or 'bright'

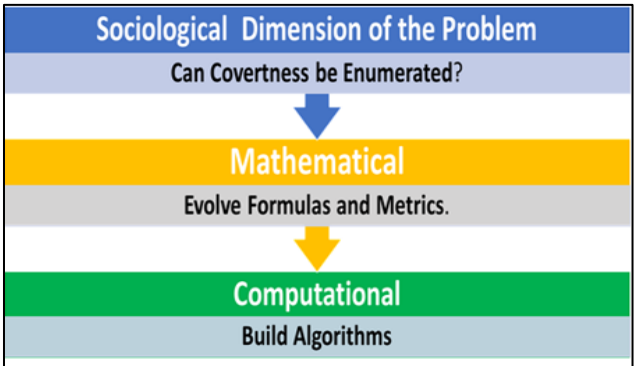


Figure 1.1 The approach to the problem is a layered one. The first and the initial layer comprises concepts drawn from the domain of sociological literature to better study and isolate the research questions. The second layer transforms the issues identified in the sociological domain to a mathematical format. The formulae defined in the second layer act as inputs for a computational modelling of the research problem and ultimately leads to the proposed solution.

commercial organizations (like ENRON, for instance) have added the practitioners in the field of Organization Behavior to the list of specialists interested in detecting and decoding 'covert' interactions within the overall architecture of larger and benign networks.

Examining the challenges posed by the problem, it is easy to see a plethora of them. Sparrow (1991) had highlighted the issues of such networks being incomplete, fuzzy, and dynamic. Add to these several other issues, including a lack of consensus on what defines covertness, interpretation of results from specific datasets remaining confined to those domains alone, the widespread practice of deception, stealth, and camouflage by the constituents of such networks, often trading efficiency for secrecy, the invariable shadow of pre-existing relationships amongst the key actors which remain hidden. In contrast, the network is being observed at the current time and the exchange of communications that defy ordinary social networks' usual structural properties.

There has been increasing interest in unveiling covert networks. Most of the research, however, has focused on *post facto* scenarios, i.e., after some (usually undesirable) event, whose roots lie in the Network, has occurred. Though this methodology helps solve the offense that has occurred as an outcome, it does not predict the possibility of a hostile event occurring by looking at an existing network. A second problem with recent research has been a disproportionate focus on content-based analysis, i.e., tools developed to do semantic and language analyses of information exchanges, and this leads to accurate detections but requires the investigator to know *a priori* which actors to focus on, which rarely happens in real-life law enforcement scenarios. This method is also resource-intensive as the volume of information that needs to be parsed is enormous, and the resources may not be available at short notice. Third, many of the proposed methods depend on networks' structural properties, e.g., geodesic paths, centrality measures, etc., which may either be absent or unreliable due to incomplete information, all hallmarks of covert networks.

Another challenge that the real-world problems of this nature face are the covert subsets' size being searched for. These days, most social networks are massive. The number of malicious actors in such huge and, for the most part, benign networks are minuscule. Even

if the question is centered around networks that may be thought of as wholly covert or criminal in its outlook, the network often forms larger (and 'brighter') networks to carry on with its clandestine activities. The investigator's job is to dig out the fixtures' malfeasant parts without disrupting the activities of the bigger Network in any major way. I answer this part of the problem by proposing a solution that identifies the contours of the 'covert' sub-networks to an optimal extent and leaves the other, more benign parts minimally affected. In other words, this research looks at the proverbial "needle(s) in the haystack" problem. It comes out with an answer that detects the needles while reducing the haystack's size or identifies a comparatively smaller part of the haystack that needs to have a closer look.

A concomitant challenge is to identify commonness amongst the covert sub-groups identified. All nodes (or higher structures) within a network labeled as covert need not be working towards the same aims. There may be different covert sub-groups with differing aims and objectives. In other words, identification of covertness is just a small first step. The more important hurdle is to identify subsets of actors (or communities) within a social network who have common intentions and work towards those objectives in concert somehow. Thus, the riddle is not just finding the 'needle in the haystack' but also sorting out the different size needles.

1.4.4 Key Contributions

This research makes several contributions to the state of knowledge in covert networks.

First, the dissertation takes a detailed look at various abstract sociological concepts that define the problem of complex sociological networks and then distills out those concepts specifically associated with covert social network analysis. These concepts are somewhat more complex than those associated with social network analysis as a whole. The concepts are then articulated as mathematical constructs to enable a later computational evaluation.

Second, this study proposes a simple metric of information confinement, which provides a way around the challenges posed by attempts to hide or encrypt information exchanges or

disguise exchange's real nature. The metrics developed by considering covertness as a property at the most basic network level, i.e., of the tie (or edge) between a pair of nodes, rather than entire sub-networks or cliques/communities. The covertness attribute is based on a measurement of confinement of information in ties between nodes and is used to rank node-pairs (or dyad-pairs as they are termed in this work) based on their covertness measure. This approach to covertness removes much of the subjectivity associated with most studies of covert or dark networks. By focusing on covertness as an attribute rather than as an overall network label, the task of identifying covert entities and communities within a network (including whole networks if the need arises) becomes more objective. It lends itself to precise measurements and mathematical modeling.

Third, this study proposes a metric for measuring collusion between node pairs that have been identified as having high covertness rankings. This metric termed a *collusion index,* enables the construction of covert subgroups within the network with common aims or intentions. This step segregates covert entities' groupings into separate cells collectively covert and whose covertness aligns with a common output or set of similar outputs.

1.4.5 Novelty

The dissertation deviates from the usual studies on covert networks by focusing on covertness as a basic attribute inherent to all social networks to a smaller or larger extent. The covert nature of an entire network is thus an aggregation of this unit coefficient. This study also looks at covertness as a property that clusters covert entities (nodes or pairs of nodes) into distinct communities inside a network. The solution also allows each covert community classification to be based on its constituent entities' common intentions that actuate their covert activities. Classification of this nature allows differential scrutiny of each covert community in the network, which allows specificity of later investigative interventions, instead of indiscriminate lumping together of all covert looking entities in the network (sometimes, the entire network itself) into one investigative bracket which can be both resources consuming and legally unsound.

The solution in the form of a covertness metric proposed in the dissertation minimizes dependence on a content-based analysis or any related intrusive or interceptive methodologies that law enforcement agencies generally adopt. The metric is a direct derivative of easily observed structural or topological properties of the network. The solution developed methods are robust and resistant to the incompleteness of information, which is an inevitable characteristic of such networks. Thus, they can thwart measures of deception or camouflage adopted by the potentially suspect actors. Both the covertness index and the subsequent Jaccard Index based similarity measures are minimalistic and simple to compute and apply. Further discussions on the novelty of the solutions developed in this dissertation and the improvements upon the state-of-the-art methodologies in related research are discussed in the last chapter.

1.4.6 Current Research

Network analytic approaches to covert networks(or criminal or terrorist or dark, the epithets vary, but the meaning remains the same more or less) can be broadly classified into two categories. The first category, which is by far, the most prevalent, is a *post-facto* analysis of the network, i.e., scrutiny of the covert network after an incident (usually illegal or undesirable) has occurred. Such approaches range from detecting the main actors from a centrality standpoint to measuring the network's disrupting actions by the simulated removal of the principals from the covert communities.

Regardless of the mechanism adopted, the crucial underlying assumption inherent to this sort of intervention remains the same; namely, the researcher is well aware that the network is covert and begins his/ her scrutiny with this fact in mind. Based on preliminary assumptions about the nature of the network, the researcher tackles the issue of which actor (or set of actors) are important to the Network, the way the network has evolved, which are the channels or pathways through which the information flows occur, the topological features of the entire network itself, etc.

The second analytical approach, which is comparatively rare in the research literature, is the predictive approach. This type of intervention aims to look at a network and

prognosticate its nature, i.e., whether it is covert or, in any way, predisposed to hide the information flows within itself. Even within this predictive approach, the literature models embed some knowledge about the participant entities' covert nature. A more comprehensive look is offered in the chapter on network analytic interventions in covert networks.

Regarding the methodologies adopted, there are many, ranging from structural analyses based on the topological features of the nodes of the network (e.g., centrality measures, shortest paths, etc.) to the overall architecture of the network (density, core-periphery structure, the assortative missing of the constituent entities, etc.) to graph-partitioning and clustering approaches. Recent research has focused on machine learning algorithms, mostly supervised (and rule-based), and may either have classifiers or clustering mechanisms. Analytical approaches to such networks may also be categorized as either divisive (e.g., partitioning the network into smaller graph structures) or agglomerative (clustering individual nodes into communities).

The solution proposed in this dissertation is patterned more on the second type of approach, i.e., the predictive model. The solution doesn't have any presumptions about the overall nature of the network being studied and begins from the standpoint of zero knowledge about the covert activities of the entities within the network. The metric proposed is based on measuring to what extent interacting entities within the network under observation are confining information exchanges to themselves (i.e., within node pairs or sets of node pairs) and deciding on a minimum threshold value to select such entities or sets of such entities for further scrutiny. In addition to identifying suitable covert candidates, the solution also mines the selected set for patterns of commonness (common intentions) and distills the most 'linked' or cohesive entities. Thus, the result is a set of the most covert looking entities within the network and communities of such entities with a probable common aim; in other words, a 'conspiracy' group. The result is a two-stage approach that is agnostic of the nature of the network and the result of which provides a strong base for later (more intrusive and targeted) interventions.

1.5 Motivations

Krebs (2001), in his landmark analysis of the 9/11 attackers, had been surprised by the sparse nature of the social network he was able to build, based on what was appearing in the news media of the time. In his own words – "I was amazed at how sparse the Network was and how distant many of the hijackers on the same team were from each other. Many pairs of team members were beyond the horizon of observability[5] from each other – many on the same flight were more than two steps away from each other. Keeping cell members distant from each other, and other cells minimize damage to the Network if a cell member is captured or otherwise compromised."

Krebs (2001) says that once the investigators knew who to look at, they quickly found the hijackers' connections and discovered several of the hijackers' *alters*[6]. Being an altar of a terrorist does not necessarily imply guilt – but it does invite suspicion and potential investigation. Krebs (2001) wonders- "The big question remains – why wasn't this attack predicted and prevented? Everyone expects the intelligence community to uncover these covert plots and stop them before they are executed. Occasionally plots are uncovered, and criminal networks are disrupted. But this is very difficult to do. How do you discover a network that focuses on secrecy and stealth?"

The above observations draw attention to the most fundamental aspects of problem-solving in covert networks. In a scenario where information is not readily accessible and the data at hand is inadequate, how does one predict the nature of the actors involved, the nature of their purported activities, the timeline of their payload delivery, etc.? The convoluted nature of the problem is caused by the deception and stealth adopted by the malicious actors to hide their ties. The second aspect is, given the constraints, can such actors be detected and their activities predicted in any meaningful manner before the planned incident happens?

[5]Friedkin (1983) as cited in Krebs (2001).

[6] A social network consists of a focal node which is termed an "ego" and the nodes to which the ego is directly connected are called "alters". In turn, each alter has its own ego network, and all ego networks interlock to form the social network.

After all, Krebs (2001) conducted his analysis *post facto* when the perpetrators had already been identified, and their activities had been laid bare.

1.6 The Proposed Solution

1.6.1 Defining the Nature of Covertness

There has been an outpouring of research in recent years devoted to tackling the menace from ‘covert’ networks or disrupting ‘dark’ networks, identifying key players or actors in ‘terrorist’ networks, etc. In all these studies, there one common strand of thought- that covert or dark or clandestine networks are illegal, or at least their output is unlawful within the scope of what passes for legality in the concerned jurisdiction. This interpretation aspect is dealt with in the chapter's breadth, which surveys the covert networks' literature. Any network that employs secrecy or stealth or a similar narrative to achieve its illicit aims is thus defined as a covert network. Researchers have attempted to pin down these concepts in a more mathematically formalized style, and several formulae are now in existence to study different aspects of covert networks.

However, a key feature that is somewhat overlooked in many of these studies is some common substructure that leads to the expression of covertness as an output of a network or some substantial subset of it. Such a substructure must be, as described earlier, definable as a unit metric, which can be meaningfully summed up to yield a larger structure of covertness, which is detectable. The second property of this unit metric should be that its articulation can be qualified in some manner by the covert intent of the participant nodes (or group of nodes). If a certain set of nodes are covert, the intent behind their covertness is different from another set of nodes within the same network. This aspect should be discernible from the articulation of the unit metric of covertness. To put it more succinctly, all covert sets need not have the same aim, but even for achieving different clandestine objectives, covertness is a required attribute.

1.6.2 Dealing with Dynamicity of Covert Networks

It is well known that social networks are extremely dynamic by nature. Their study presents formidable challenges to the researcher who tries to fit them into some systematic mathematical model. The challenges grow manifold when researchers seek to predict how a network might look in the future, given the dynamics between the actors; this aspect remains the most debated and researched topic in recent social networks studies. A special subset of the study of social networks is the study of covert social networks. As has been discussed earlier, these networks present special problems to the researcher and the investigator alike. The active acts of deception, dissimulation, hiding, and camouflaging of information flow between the actors, and indeed throughout the network add a layer of complexity to the existing problem of social networks' dynamic nature. Add to this the chance that the actors had ties before the network's plotting, and predicting the nature and future shape of a covert network becomes practically unsolvable. However, given the security challenges that the world faces from such networks, an approximate solution is more than welcome.

1.6.3 Uncovering Nuances of Covertness

Covertness as an attribute resists firm mathematical formulation since its nature is abstract. It's seen from the above analyses that a common denominator to all forms of covertness, licit or illicit, is the element of hiding of information in some manner. This hiding is also synonymous with the confinement of information exchanged between entities within a network. This confinement or hiding of information may take the shape of deception, use of covert channels[7], encryption of transactions, lack of sharing of knowledge with third party entities within the networks. These pre-existing ties are not apparent at the time of observation, etc. Most of these actions differ in how they work and how they are used to achieve covertness. The one area of convergence is the hiding of information, i.e., the

[7]In computer security, a **covert channel** is a type of attack that creates a capability to transfer information objects between processes that are not supposed to be allowed to communicate by the computer security policy. Covert channels are defined as channels not intended for information transfer at all. (https://en.wikipedia.org/wiki/Covert_channel accessed on 20.06.2020)

methodologies are different, but the impact is the same. In other words, if we find a methodology to detect the confinement or hiding of information by entities in a network and can also measure how much of the information is dammed up, we can arrive at how exactly these entities are arranging to go about it. Irrespective of the nuances of the techniques the entities are using to bring about covertness in their transactions; the result inevitably is the hiding of information. If we succeed in measuring this quantity, it will not be difficult to reach the causes of what is causing it and, more importantly, causing it.

Second, social networks are seldom completely covert. Many of the covert networks that are so labeled are subsets of bigger networks that may not be defined as covert. Most social networks have covert areas manifest within themselves, and much of this 'dark matter' may not necessarily be illegal or clandestine. Covert areas are perfectly compatible with most businesses and enterprises. For example, a group of senior employees within an organization who confer with each other in private to decide key organizational strategies for a forthcoming market launch will conform to the definition of a covert subgroup in as much as they will not share any information lest it leaks out and benefits their competitors. Such a subgroup's appearance and outcomes are covert but not their aims, anything but clandestine[8]. The larger question that arises is whether covertness as a property is fundamental to every network. The answer should be unqualifiedly affirmative. If so, then how do we measure this property, and more importantly, can we construct basic building blocks of covertness at the most fundamental level of a social network, say at the node or the tie level and then somehow correlate or coalesce the units of covertness into larger covert subgroups within the overall architecture of the network?

Third, this study looks upon covertness as an abstract concept and seeks to meaningfully derive a mathematical formula that describes it as a data point aggregated or disaggregated or factored into calculations. The difference between a sub-structural attribute like density, sparsity, homogeneity, connectedness, and an abstract concept like covertness needs to be emphasized. Any structural attribute is easily observable and measurable, while an abstract attribute like covertness needs to be inferred from such structural attributes after further

[8] Something is clandestine if it is covert and illegal.

analysis, mostly manual. Sub-structures with some specific attributes (like the one described in this study) are more likely to be covert than others whose measure of the same attribute is lesser. It is necessary to corroborate other facts on the ground and perform some deep-packet searches of these groups before concluding their nature. This study aims at presenting a distinct number of sub-structures within a network that stands out from other structures in terms of the overall measure of the covertness attribute. The number of structures selected for further analysis is much smaller than the overall number of sub-structures in the network and presents an affordable scenario. But the caveat remains that the final analysis is always manual.

1.6.4 Covertness as a Universal Attribute

Earlier, there was a discussion about how all organizational networks have covertness manifest within them, even at the most fundamental levels. So, if covertness is considered a universal attribute in all social networks, covert sub-networks or subgroups can be built up as superstructures using the units of covertness as some sort of ‘lego’ bricks. To solve this part of the problem, we need to look at the parallel problem of ‘commonness’ (i.e., common intention) inherent in the larger covert superstructures. To extend the example given earlier, we may consider the second group of employees within the same organization whose aims are not benign and might be involved in trading the enterprise's business secrets. This category of activity is surely risky, illegal, and clandestine and carries the pain of punishment if discovered. It stands to reason that these actors will hide information flows that occur amongst themselves from others and thereby form a covert subgroup entirely different in purpose (illegal) and methodology (deception or camouflaging their exchanges). Thus, when we formulate a technique to measure covertness and then find a way to agglomerate the covert units into bigger structures, we also must devise the means to differentially aggregate units of covertness so that the dark areas within a network fall into independent buckets (or overlapping buckets in case entities repeat themselves across such containers) and form covert subgroups within the network whose methods to hide information may be similar, but whose objectives might be wholly different.

1.6.5 Tie as a Basic Unit of Covertness

While trying to solve these issues, there is yet one more crucial aspect that needs to be considered. This is the question about the nature of the individual actors within the network; given that they may form parts of covert subgroups, is there a possibility that these actors lead some kind of 'double existence,' a sort of Dr. Jekyll and Mr. Hyde type of role? If we go by some of the more seminal studies focused on covert networks, the answer is a surprising "Yes." Let's take the case of some of the plane hijackers who were involved in the 9/11 bombings. Many of them were pursuing innocuous trades (training to be pilots would hardly qualify as something nefarious). They were part of social networks that were overt (the trainee pilot community network, for instance). Hence, there is a likelihood that a node that is designated as a member of a covert subgroup may also be part of other subgroups that may or may not be covert. The solution should thus be able to filter out the covert essence from within the actor or node to allow the node's inclusion in multiple subgroups, covert or otherwise, within the larger ecosystem of the social network. This study brings out the covertness aspect of nodes through their relationships or ties with other nodes. An actor in a network will express its covertness only with another actor who shares its intentions and aims, i.e., the basic unit of a network so far as its covertness content is concerned, is a pair of nodes and the tie between them.

1.6.6 Non-Intrusiveness of the Proposed Metric

Next, the solution should be so devised that it doesn't need any 'intrusive' content, i.e., detecting the nature of any entity or structure (i.e., whether or not its covert and related facts) shouldn't be dependent on what is inside the communication flows. The reasons for this particular facet have been made clear in the earlier passages. Today's world is full of privacy laws, data protection policies and protocols, and encrypted messaging. The surveillance and enforcement agencies typically find it very difficult to bypass these rules and regulations to access the content. Across most open societies wiretapping laws have gotten more and more strict. And the covert actors have become better and better with their

hiding and deception skills. Technology hasn't lagged in this race, and modern-day encryption is extremely hard to beat even with vast computational resources at command.

1.6.7 Minimization of Resource Requirements

The other aspect of difficulty in intercepting and analyzing content is the sheer volumes of data involved. Most communication networks today have typically millions of nodes and the data generated per second is typically in trillions of bytes. The problem of monitoring the content of each exchange of data and coming to any sane conclusion about possible clandestine activities is bewilderingly complex and time and resource consuming. It won't be far off the mark to state that no enforcement agency currently has these resources and brutes processing power on tap continuously. Typically, the time available to detect and stop malicious actors from delivering their payloads is in days, not weeks. The aim should be to meaningfully pare down the numbers of suspect actors to a minimal level while ensuring that the information about at least some part of the covert subgroup is captured for analysis and further dissection. If the set of suspect actors is not large, it will not be difficult to obtain warrants calling for the contents' information exchanges.

1.6.8 Summary of the Proposed Solution

Based on the above discussions, we can summarise the motivations of this research into the following points:

(i) To define the concept of 'covert' or 'covertness.'

(ii) To break down the covertness concept into fundamental covertness units, which can be applied at the network's base layers.

(iii) To devise covertness units in ways that allow individual nodes to 'participate' in more than one subgroup, whether covert or otherwise.

(iv) To achieve the above steps in a not intrusive manner and rely on the social network's existing structural information.

(v) To use the covertness unit as a basic building block to construct larger communities of covertness, which will have some common aims different from the network's overall objectives.

(vi) To detect the covert actors (nodes) or covert communities of nodes and predict their activities with some accuracy.

(vii) To achieve a fair degree of accuracy in detecting and predicting covert communities with easily affordable resources and real-time.

1.7 Problem Definition

This research focuses on the email corpus of the ENRON Company, which went bankrupt following a major financial scam in 2002. The email corpus of ENRON is easily available in the public domain and remains one of the most well researched social network platforms. Researchers have focused on many aspects relating to the ENRON company and the interactions amongst the employees. It's not surprising to note that many of the studies have tried to determine if the employees were complicit in the scandal or had some prior knowledge about the goings-on through various methodologies. This aspect of the ENRON email-based social network makes it useful for comparing the effectiveness of approaches. Most computational approaches are based on analysis of the contents of the emails, or the nature of financial information exchanged amongst the employees. Some have even looked at the incentives derived by certain employees as a means to detect complicities.

Briefly, the ENRON dataset is a corpus of emails collected from the inboxes of 151 of its employees covering a limited period relevant to the investigation of the insider trading scam and other concomitant illegal business practices that led to it. A total of 517,431 email exchanges available from these inboxes, and information about a possible 6568 employee email ids were extracted from the email corpus. Each email id takes the place of a node or vertex in a social network. The half a million or so email exchanges condense to approximately 55,300 unique email pairs, i.e., the number of pairs of email ids that have exchanged at least one mail. Further, scrutiny of the mail exchanges and reports from

contemporary media sources reveals that a certain ENRON employee segment was indicted in the judicial processes that followed multiple investigations. Some of the employees were also examined as witnesses. Based on these sources, this paper has zeroed in on 19 employees who were either indicted or were privy to the proceedings otherwise as witnesses or recipients of the information. There are 43 unique mail pairs from amongst these employees of interest[9]. The challenge is to extract a small enough subset of mail pairs (or edges) from half a million-plus, which will contain a significant number of these edges of interest (EoIs). Notationally, the problem is defined below:

Let's define the ENRON mail corpus as a social network graph G, such that $G = (V, E)$,

where V is the set of all nodes in the graph network

and E is the set of all edges or mail-pairs in the graph network, including those formed when copies of e-mail exchanges between pairs of nodes are marked to other nodes.

The number of edges in the network graph is represented as the cardinality of the set of edges V, i.e., $|V|$

$\&|V| = 6568.$

The number of edges in the network graph is represented as the cardinality of the set of edges E, i.e., $|E|$

$\&|E| = 55{,}300.$

Let's define the set of the employees of ENRON who were part of the scam as a graph G_C, such that $G_C = (V_C, E_C)$,

where V_C is the set of all nodes of interest (NoIs) in the graph network and E_C is the set of all edges of interest (EoIs) in the graph network.

Obviously, $G_C \subset G$ and $V_C \subset V \& E_C \subset E$;

$\&|V_C| = 19$ and $|E_C| = 43$;

[9] The mail pairings between these employees are referred to mostly as Edges of Interest or EoIs throughout this paper. The employees who are of interest are referred to as Nodes of Interest or NoIs.

The ratio of the overall edges of interest (EoI) e_{ij} (i and j are nodes of interest (NoIs) in the graph network) to the set of all edges of the graph G thus comes to:

$$P = |E_C| \;/\; |E| = 43/\; 55{,}300 =$$

The task in hand is now to increase this ratio, i.e., boost the chances of detecting an edge of interest (EoI) in the graph network's set of edges comprising ENRON's mail corpus.

The above problem statement may also be framed in a probabilistic manner:

What is the probability of detecting at least one covert edge from amongst the broad set of edges of the ENRON e-mail network in 20 tries?

Let's define an integer *k*, s.t., *k* = number of tries; here, *k* = 20.
There are 43 covert edges or Edges of Interest (EoIs).
Let's define the number of EoIs as *m*; here, *m* = 43.
The number of edges overall is 55,288 ~ 55,300
Let the total number of edges be defined as *e*; here, *e* = 55,288.
We need to calculate the probability of not getting any covert edges in 20 tries.
Let's define the probability of detecting a covert edge as P_c and not detecting a covert edge as P_{nc}.

The probability of not detecting a covert edge in the first try is(55,300 – 43)/ 55,300.
The probability of not detecting a covert edge in the second try is (55,299 – 43)/ 55,299.
In this manner, the probability of not detecting a covert edge on the 20th try is(55,280-43)/ 55,280.

Notationally,

$$\boldsymbol{P_{nc}} = \prod_{i=0}^{k-1} \frac{((e-i)-m)}{(e-i)}$$

$$\boldsymbol{P_c} = (1 - \boldsymbol{P_{nc}}) = \left(1 - \prod_{i=0}^{k-1} \frac{((e-i)-m)}{(e-i)}\right)$$

Hence, the probability of not getting a covert edge detected in 20 tries (k = 20) comes to:

$$\boldsymbol{P_{nc}} = \frac{(55{,}300 - 43)}{55{,}300} \; x \; \frac{(55{,}299 - 43)}{55{,}299} \; x \ldots . x \; \frac{(55{,}281 - 43)}{55{,}281}$$

$$\boldsymbol{P_{nc} = 0.984560175;}$$

$$\boldsymbol{P_c = (1 - P_{nc}) = 0.015439825}$$

Thus, the probability of detecting at least one covert edge (or Edge of Interest, EoI) from the set of all edges in the ENRON mail network is 0.0154 approximately. The graph plot of the sequence of probabilities of identifying EoIs in k tries with $0<k\leq 500$ is shown in Figure 1.2 below. The vertical line marks the value of the probability of detecting at least one covert edge or an Edge of Interest (EoI) when k = 20, i.e., with 20 tries. The value of Pc goes on increasing monotonically with an increase in the value of k. The probability of detecting a covert edge when k = 500 is 0.33 approximately. As the study progresses, the probability values get enhanced by introducing the proposed solution metrics (i.e., the Covertness Index and then the Collusion Index).

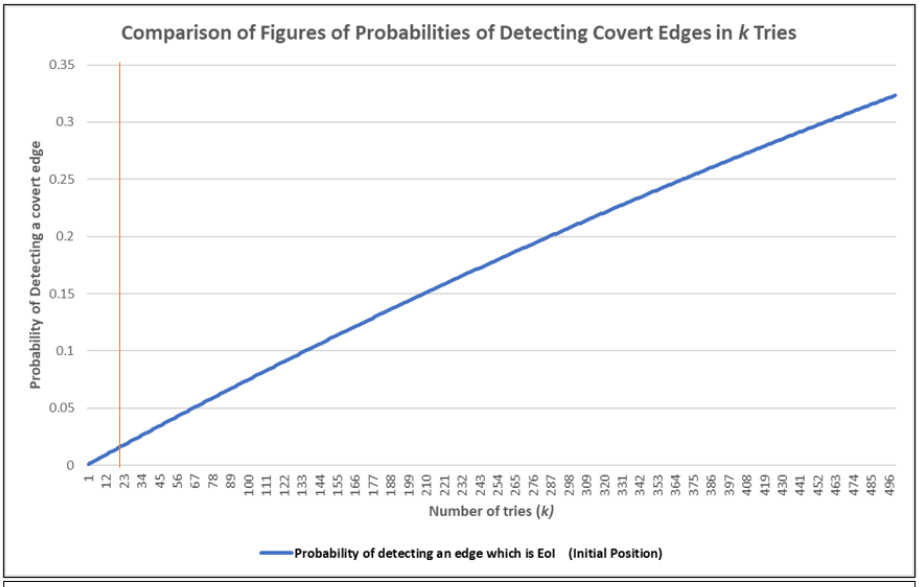


Figure 1.2 Plot showing the probability of identifying at least one covert edge (EoI) from amongst all the edges in the network in *k* tries where the value of *k* varies from 0 to 500. The vertical red line shows the value of probability at *k* = 20.

This research seeks to determine whether the probability of identifying at least one covert edge in a fixed number of attempts can be enhanced using the metrics proposed.

1.8 Outline of the Dissertation

Chapter Two begins with a survey of covert social networks and the various definitions of what best describes covertness from a qualitative standpoint, as discussed in multiple sociological treatises that have dealt with the topic. This survey is a needed primer for this study. It brings out covert networks' unique characteristics instead of ordinary social networks where the priorities and outcomes are entirely different. I then employ select properties of covert networks to develop technical measures and metrics to evaluate such networks and make the covertness attribute a unique one rather than one that applies just to specific networks described as covert, often contextually.

The second chapter of the dissertation also describes various tools and techniques associated with social network analysis (SNA) in general and covert networks. This section discusses the pros and cons of multiple methodologies to detect interest nodes to the investigator and

the network's shape after their deletion. Newer approaches to the old problem of detection of critical nodes are included. This chapter also touches upon the methodologies for linking structurally unrelated nodes and matrix-based approaches to solving community detection problems in social networks.

Chapter 3 discusses the concept of defining a unit for covertness and explores other research in the area which has looked at covertness as an attribute inherent to networks rather than a qualifier for an entire network. In the same part, I propose and derive a formula to mathematically encapsulate the concept of covertness based on the edges (ties) between the nodes of a dyad (pair of nodes considered as a unit). The covertness metric developed in the previous chapter is applied to the email-based network of ENRON. After the covertness values of the edges in the network are estimated using the metric, the edges within the network are then ranked in descending order of the covertness index values defined on their edges. Thresholds based on cut-off values of the covertness index are chosen heuristically, and sets of top-ranked edges are accordingly selected from the overall group of edges in the network. The selected sets of edges are scanned for the presence of edges of interest (EoIs), and the outcomes are measured with the help of metrics like precision and recall to evaluate the efficiency of the covertness index. The results obtained from the covertness index's application are compared with those derived from a hypothetical set of edges of the ENRON mail network. The distribution of the edges of interest is assumed to be uniform. The difference in outcomes is a further testament to the efficacy of the covertness index.

Chapter 4 is a survey of various approaches toward community detection in social networks, emphasizing covert social networks. This chapter discusses the pros and cons of different approaches to the problem. Various solutions existing in the literature are examined, such as graph partitioning, clustering, etc. are discussed and briefs on their usefulness in the current scenario. Finally, the choice of a modified link prediction mechanism to link the edges in the selected set obtained in the previous chapter is justified. The linkage mechanism forms the basis of the similarity index that is developed to translate the sociological concepts of collusion and conspiracy between pairs of nodes in a network into a graph theoretic format. This similarity coefficient, which is termed the “collusion index”

in this dissertation, is used to link pairs of edges from selecting the experimental dataset (i.e., the set of edges chosen based on the specific thresholds of their covertness index values) obtained in the previous section.

Chapter 5 uses the similarity index proposed in the previous chapter. The index is used to link edges extracted from the table of top-ranked covert edges arrived at after applying the covertness index. After the edge-pairs are linked, the pairs are ranked in terms of their scores of similarities, i.e., the more excellent the Jaccard Index score of the edge pair, the higher it scores. As in Chapter 4, sets of edge-pairs are selected based on a heuristically calculated threshold score of similarity. Precision and recall are computed on the selected set of edge-pairs to ascertain how many pairs of edges of interest (EoIs) are present. These results are compared with the corresponding results from a model where the distribution of the pairs of edges of interest is assumed to be uniform. The comparisons reveal the efficacy of using a similarity index to link pairs of edges.

Chapter 6 discusses and analyzes the approaches and methodologies adopted in this study and analyzes the results arrived at. Chapter 7 concludes the dissertation and suggests future steps.

1.9 A Note on Terminology

The study of social networks is an interdisciplinary one involving concepts and inferences from sociology and computer science. A network is a collection of points joined together in pairs by lines in its most basic form. The mathematical construct that is closest to a social network in terms of structure is the graph. A graph is a mathematical representation of a network, and it describes the relationship between lines and points. A graph comprises a set of points and lines between them. The length of the lines and position of the points do not matter. The branch of mathematics that is devoted to the study of graphs is called graph theory. Given the structure of networks, whether they belong in the fields of physical, biological, electrical, economics, or the social sciences, the most optimal way they can be

modeled mathematically is through the lens of graph theory. All computational studies on social networks are predominantly based on the concepts of graph theory.

Thus, many of the terms used in social sciences to describe networks and their structures have their counterparts in graph theory and computer science. From the perspective of sociology, the most basic unit in a social network is an *actor*. The equivalent term in graph theory is the *vertex,* and in computer science, the *node*. Actors have links amongst themselves, called *relations*. Such relations may either be *undirected* or *directed,* where one of the actors is the initiator, and the other is the receiver. The transactions may be mutual, for example, a telephone conversation. The counterpart terminology for a *relation* in graph theory and computer science is the *edge*, also called a *tie.* A pair of actors who share a relationship form a *dyad*. A dyad is termed a *dyadic-pair* or a *node-pair* or just referred to as a pair of nodes in graph theory and computer science. Likewise, a group of three actors is termed a *triad*. These terms have been used interchangeably throughout this dissertation, and there is no loss or alteration of meaning when one term is used in place of the other.

A few terms used in the dissertation are specific to the study. Since the dataset used here is the ENRON email corpus, the mail-ids of the employees are, for the most part, considered as the nodes in the network. In some places, the employees themselves have been referred to as actors in the specific context of their activities, which are under the lens. The mail exchanges between the employee mail-ids constitute the relations between the employees and are referred to as the edges or ties interchangeably. An edge or tie exists between two nodes if at least a single email has been exchanged between them. A single edge or tie is defined as existing between the nodes irrespective of the number of emails exchanged. If there are no mails exchanged at all between a pair of nodes, the edge has by definition a value of 0, and if one or more have been exchanged, the value of the edge is 1. Unusually, pairs of nodes that don't have an email exchange are also considered part of the study. Still, it needs to be noted that in the study of covertness, the material available is often incomplete and needs to be inferred based on further scrutiny.

There are inevitable instances where there may not be any relations existing at the time of the scrutiny. Still, it may have existed earlier, or it might merely be unavailable due to incomplete information. Other significant terms used are *Relationship Sets, Shared Relationship Sets, Neighborhood Relationship Sets,* and *Edge-Vertices.* An edge-vertex is a function defined upon an edge between two nodes and is a list set that stores several pieces of information needed for further processing. Edge vertices are different from edges in that the data stored in an edge-vertex is more significant, and the two terms should not be confused. Another commonly used word in the dissertation is the *edge-pair,* which denotes a pair of edges joining different dyads' nodes. The two metrics developed in the study, the covertness index, and the collusion index, are defined and described in detail during the experiment. Communities within the network, which are generally homogeneous groups, have extrapolated meanings in this study. Here, they mean *covert communities with shared aims* and are referred to interchangeably as *collusion* or *conspiracy* networks.

Chapter 2

Covert Networks & Use of Analytics in Covert Networks

2.1 Overview

Koschade (2006) defines a social network as a finite set of actors and the relation or relations defined on them and describes social network analysis as a mathematical method for connecting the dots that allows the analyst to map and measure complex group. The overall structure rests with the researcher's interpretation of its relationships and dynamics. The nature of a social network lies in the beholder (or researcher), which gives rise to the intriguing possibility that the same organization may be seen to have different network configurations depending on the study priorities and on which aspect of the inter-relationships the researcher may find interesting. This type of conceptualization is at the heart of network multiplexity, the idea that many kinds of relations can coexist among a set of nodes. The researcher has to decide which relational content is essential and which to ignore.

The social network perspective provides a set of methods for analyzing the structure of whole social entities and various theories explaining the patterns observed. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics. The correct interpretation of networks assists in predicting behavior and decision-making within the network.

These features are not essentially different for covert networks. There are suggestions that covert networks function similarly to their overt counterparts (Asal and Rethemeyer, 2008; Crenshaw 2010). A survey of the literature in this regard, however, reveals few comparative studies, with some of the existing work aimed at the similarities and differences between offline and online covert networks (Keegan et al. 2010). There doesn’t seem to be any single definition of what constitutes a covert network. Even within this category, there is a

proliferation of terms to describe networks that are not overt or normal in some sense. Some of the standard descriptions of such networks are “clandestine,” “dark,” “illegal,” “illicit,” “underground,” “criminal,” “terror,” etc. The demarcation of any network as a covert one appears to be context-specific and based on what the study considers overt or “bright networks” instead of dark networks (Raab and Milward, 2003, 419).

Baker and Faulkner (1993) defined covert networks (the illegal network was the authors' precise terminology in their landmark paper) as a network that doesn't often behave like standard social networks. Conspirators don't form many ties outside of their nearest cluster and often minimize the activation of the network's existing relations. Strong ties between prior contacts, which were frequently created years ago in school and training camps, keep the cells linked. Krebs (2002) states that these strong ties remain mostly latent and hidden to outsiders, unlike in typical social networks. Since participants in dark networks intend to commit illicit actions, they deliberately conceal their conspiratorial relations to avoid exposure.

(An example, a sleeper cell that remains inactive until called into action). Freeman and Gill (2013) define covert (or dark) networks as social networks characterized by low visibility, low interactivity between nodes, and high uncertainty about the connections. In a regular social network, strong ties reveal the cluster of network players – it is easy to see who is in the group and who is not. Strong ties may appear to be weak ties in a covert network because of their low activation frequency. Furthermore, the ties of interest typically have multiple qualitative aspects and are likely to change over time. Although connections between individuals change character temporally, such changes typically cannot be measured directly and require some form of estimation. These are fundamentally distinct networks since the actors are “trading efficiency for secrecy” (Fellman & Wright 2004, 5; Krebs 2002). The less active the network, the more difficult it is to discover. Yet, the covert network has a goal to accomplish.

One of the earliest definitions of such networks was given by Erikson (1981, 60:188-210), who defined a secret society as a persisting pattern of the relationship which links

participants in secret activities. This definition covers a wide swath of networks and organizations whose aims are not illegal per se. For reasons that may be political or social, they tend to keep themselves undetected. Examples include the early twentieth century's suffragette movements in western nations, Alcoholics Anonymous, different gay and lesbian interaction networks, etc. The very structure of covert networks is determined by a desire to maximize secrecy and avoid detection (Krebs, 2002; Raab and Milward 2003; Helfstein and Wright 2011). Baker and Faulkner (1993) state that network members must balance the need for secrecy and stealth with frequent and intense task-based communication. The covert network must be active at times – it has goals to accomplish. During these periods of activity and increased connectedness is when it may be most vulnerable to discovery. Thus, the primal instinct of all covert networks of all hues is an overriding need to keep the network secret and undisclosed, although just what needs to be kept confidential and from whom and for what length of time vary widely.

Oliver (2014) observes that much of the literature on covert networks attempt to describe and resolve opposing tensions, the requirement to be secret, and achieve the network aims. The imperative to keep things hidden stems from the severity of network detection, "infiltration" or exposure, ranging from embarrassment to penal punishments, to death (Oliver 2014; Bakker, Raab, and Milward 2012). Secrecy is not easily quantifiable, and this renders the task of measuring the covertness of any network a subjective one. Secrecy may be pervasive across all aspects of the network's structures (e.g., in terror networks and many organized crime networks) or may be limited to any or all of their identities, aims, activities, as in gay and lesbian groups, specific political movements, suffragettes, etc. Oliver (2014) states that the requirements for different secrecy types, for various lengths of time and from different audiences, are likely to produce a range of network structures. She feels that the effects of these facets of secrecy are not well understood.

Simply stated, covertness is not a binary state, and there may be several gradations that will be difficult to quantify in a mathematical sense of the term. Nevertheless, in the wake of several major terrorist incidents and the rapid proliferation of organized criminal networks in recent times, researchers have attempted to identify the commonalities among covert

network security concerns that have led to several studies regarding covert networks' commonalities. Their primary motivation has been security concerns and the pressing need to identify such networks' vulnerabilities to disrupt them. Most of the research in this area has focused on actual events or case studies, like the 9/11 attacks or the Bali bombings. This accent on singular instances combined with differences in approach, data used, methods, context, or even the native discipline has led to significant disagreements about their very nature.

2.2 Distinguishing Traits in Covert Networks

The above analysis brings out features in networks that are responsible for covertness in networks. The preceding discussion explains covertness as something that arises from the general features of social networks. A network's form and the structure of the relationships between its members determine its purposes (Morselli 2009), which is true of covert networks, and the same is true of a covert network as well? However, Sparrow(1991) identifies four distinct traits that point to covertness in a network: (1) Size, (2) Incompleteness, (3) Fuzzy Boundaries, and (4) Dynamic Networks.

Size: Crime related databases are typically enormous, with thousands of nodes, many of which might not be actual members of the network. The fact that such lists of unconnected persons are massive tells us nothing about covert networks' real size, which has been disputed by others who feel dark networks tend to keep themselves small to avoid detection (Bouchard 2007).

Incompleteness: Criminal network data is inevitably incomplete, given the element of secrecy by the actors involved, the biases introduced by the investigative methods, and assumptions. Incompleteness is two-fold: actors can be absent from the database (or inaccurately included), or ties between actors may be unknown (or incorrectly believed to exist). The fragmented nature of these networks tends to get further distorted by agencies' inclination to pay the closest attention to those individuals they were already familiar with

and who may not be the principal players. The nature of incompleteness means that prevalent statistical inference methods can't be successfully applied to predict the entire structure.

Fuzzy Boundaries: In covert networks, it's challenging to decide which individuals to include within the network. Sparrow describes criminal networks as having ambiguous boundaries and possessing complex inter-relationships with other criminal networks. Actors may participate in multiple criminal networks, thus serving as portals that link subgroups. The fuzziness aspect makes it hard to apply conventional centrality measures to study and evaluate illegal or covert networks.

Dynamic: Sparrow observes that covert networks change relatively faster through time than conventional networks, and relationships between actors have different/varied distributions over time. Many of the patterns that define the covertness of a network are transient. This aspect of covert networks is perhaps the most significant for identifying them, particularly if we consider the fact that such networks are likely to be disrupted frequently and need to evolve quickly to survive.

2.3 Ties in Covert Network

Ties may be defined as connections that transfer information and resources between actors in a network. Borgatti (2003) calls this aspect of networks the "flow model" (In other words, ties are channels through which something is transferred from one actor to another). The other significant type is the "bond model," where actors are involved in a collaborative action (conducting a robbery or a terror attack). Ties involve flows of information and other resources, including finance, disease, ideas, issues arising from kinship or friendships, etc. Based on their strength, Granovetter (1973) identifies ties as strong, weak, or absent. He defines the strength of a tie as the "combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie….It is sufficient for the present purpose if most of us can agree, on a rough intuitive

basis, whether a given tie is strong, weak, or absent" (Granovetter, 1973, p.1361). Thus, determining the strength of a tie is a subjective affair, and researchers from different domains are liable to define tie-strength variably.

Sparrow (1991, p.271) believes that the most valuable communications channels to monitor are seldom used and lie outside the relatively dense clique structures. These outer channels correspond to weak ties, or perhaps they are not weak ties to the participants but only appear weak to outside observers who fail to detect the secret link. These are the ties that add most to the efficiency of communication within a covert network. Therefore, urgent or important network signals are more likely to be detected on the weak ties than on the manifestly strong ones. Granovetter (1973) argued that weak ties are most valuable if they are bridges that link otherwise separated subgroups.

Given the imperatives of maintaining secrecy, Oliver (2014) argues that covert nodes may reasonably be assumed to have fewer ties than overt nodes. Keegan et al. (2010), using MMORPG data, have also drawn similar conclusions. However, in a recent study of child exploitation websites, Westlake and colleagues showed the reverse (Keegan et al. 2010). Thus, there may not always be a linear relationship between secrecy and weak ties in a covert network. Still, it may be safely assumed that in most cases, this view is somewhat counter-intuitive to the conventional wisdom that strong ties, kinship, in particular, are crucial to prevent a covert cell from being penetrated by agents trying to disrupt the network. Amore vital hallmark of covert networks is the large number (or proportion) of null ties compared to overt networks. Secrecy is better protected not by many weak ties but by many edges with no ties. Clandestine organizations that use a classic cell structure escape detection and disruption when one member is captured because that member has few connections to others that can be betrayed.

2.4 Pre-existing Ties

Pre-existing ties between the actors are the precursor of covert network structures (Erikson 1981; Harris Hogan 2012; Lauchs, Keast and Yousefpour 2011; Milward and Raab 2006). Such ties increase covert networks' resilience by relying on individuals who can be trusted (Keegan et al. 2010; Lauchs et al. 2011). Thus, given the importance of pre-existing ties in covert networks, studying the existing relations linking actors may be deceptive. Key actors may choose to connect to others in the network rarely to avoid detecting and identifying their roles. Raab and Milward (2003) state that actors in a covert network must continuously evaluate the risks and react accordingly. Many of the ties forging the leading players may have been generated informally and based on trust before the formal network structure evolved. Structural intuition based on a last snapshot of the network can be misleading. Pre-existing ties are said to underpin and help maintain covertness in networks (Oliver, 2014; Crossley, Edwards, Harries, and Stevenson 2012; Edwards and Crossley 2009; Erickson 1981; Sageman 2004).

An excellent example of trusted pre-existing ties leading to a covert network set-up is the 9/11 event. The 19 hijackers appeared to have come from a network that had formed while they were completing terrorist training in Afghanistan. Many were school chums from many years ago, some had lived together for years, and kinship ties related others. Deep trusted ties that were not easily visible to outsiders wove this terror network together. Krebs (2002) states that an accurate picture of a covert network depends on identifying tasks, trust, resources, and strategy/goal related ties amongst the conspirators.

Coleman states (1990) that networks are often the unintended consequences of other activities, which implies that pre-existing ties are not the exclusive progenitor of covert networks. Also, most networks tend to form around shared spaces, events, activities, and covert networks that are also amenable to this principle. Klerks (2001) and Reed (2007) suggest that ties in covert networks are multiplex in nature. The more multiplexed the tie, the closer the relationship between the actors. An example is a network formed by Al Qaeda and its affiliates. The work done by Zimmerman (2013) reveals that these links were multi-

layered and multi-dimensional, ranging from ideology (Islamic fundamentalism) to commercial and even charity. Multiplexity of ties where each dimension's nature may be orthogonal to each other, i.e., information about the nature of one tie, may be independent of information about the nature of another tie. Neither tie can be used to infer anything about the other. This property can make covert networks extraordinarily resilient and resistant to disruption. This property is borne out of how Al Qaida has borne many powerful nations' military might. The US military action in Afghanistan and elsewhere in the Middle East and Africa in the first decade of the millennium virtually annihilated the older network structures of the Al Qaida network as it existed in the nineties. The targeted killing of Osama bin Laden in 2011 capped the effort to cripple the network from launching any meaningful attacks. But the decade since has seen the organization claw back into public imagination through a series of outrageous attacks (but with a lesser impact on the western media). Its affiliates in Iraq, Syria, Kenya, Mali, and Somalia have wrought devastation. Its spiritual comrades-in-arms (not completely subscribing to its philosophies but rather similar in practice and prejudice) have sprung up in areas vacated by the old Al Qaida. Even though some consider the ISIS terror group as a distinct group from the old Al Qaida, there is an increasing trend amongst watchers of Islamist terror groups that ISIS and Al Qaida are joined at the hip in many respects, as explained in excerpts from a recent article in *'The Atlantic'*[10] by Hassan Hassan, co-author of *ISIS: Inside the Army of Terror*.

> "Most historians of the Islamic State agree that the group emerged out of al-Qaeda in Iraq due to the U.S. invasion in 2003. They also agree that it was shaped primarily by a Jordanian jihadist and the eventual head of al-Qaeda in Iraq, Abu Musab al-Zarqawi. The Jordanian had a dark vision: He wished to fuel a civil war between Sunnis and Shiites and establish a caliphate. Although he was killed in 2006, his vision was realized in 2014—the year ISIS overran northern Iraq and eastern Syria. Narratives about the origins of Islamic State ideology often focus on the fact that Zarqawi and Osama bin Laden, both Sunni extremists, diverged on the idea of fighting Shiites and questions of *takfir*, or ex-communication. Such differences, the story goes, were reinforced in Iraq and eventually led to the split between ISIS and al-Qaeda. Based on this set of assumptions, many conclude that Zarqawi must have provided the

[10] https://www.theatlantic.com/ideas/archive/2018/11/isis-origins-anbari-zarqawi/577030/

> intellectual framework for ISIS. Recently, I came to question the conventional wisdom. The groundwork for ISIS was arguably laid long before the invasion. If there was one person responsible for the group's modus operandi, Abdulrahman al-Qaduli, an Iraqi from Nineveh, was better known by his nom de guerre, Abu Ali al-Anbari—not Zarqawi. It was Anbari, Zarqawi's No. 2 in his al-Qaeda years, who defined the Islamic State's radical approach more than any other person; his influence was more systematic, longer-lasting, and more profound than that of Zarqawi. (p.1)"

The views above are a reflection of the fact that though most of the old ties within the Al Qaida network got obliterated with intense action from counterterror groups, some of the ideological moorings that had fostered the network and its objectives remained unaffected, and its "seeds" led to green shoots of ISIS-related terror in the fertile badlands of revolution and anarchy prone border regions of Iraq and Syria, which is ample proof of the fact that resilience of covert networks is a fundamental aspect built into their structures and ties. Repeated counter-terror action in the manner of "mowing the grass" once in a while may reduce resilience in such networks.

2.5 Homophily and Microstructures

Underpinning all networks, overt or covert, are the actors' common objectives and methods. The commonness of objectives, methods, occupations, ideologies, and preferences underpin all networks, overt, or covert. But this trend is marked in covert networks where the actors have incredibly convergent aims and methodologies to achieve them, and participants are often drawn towards these networks based on strong homophilic connections (Harris-Hogan 2012; Freeman and Gill 2013; Everton 2011; Milward and Raab 2006; Reed 2007; von Lampe 2009). Earlier sociological studies (Kossinets &Watts, 2006; McPherson, Smith-Lovin, & Cook, 2001) indicate that a set of facilitators influences the tie-formation process in covert networks. These facilitators may be individual attributes like age, race, gender, type of crime, etc. (Feld, 1982; McPherson et al., 2001; Everton 2011; Thelwall, 2008; Reiss, 1986) of the nodes/actors in the network, or shared

affiliations between actors, like kinship and mutual acquaintances (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006; Kossinets & Watts, 2006).

Other studies indicate that though homophily may play an essential role in the covert network's formative stages, exogenous factors may dilute its effect at a more mature phase. Klerks (2001) looks at organized crime networks and concludes that covert networks tend to become less homogeneous over time. The same conclusion is seen in other works as well (Carrington and van Mastrigt 2013). The actors in a covert network are also aware of attempts to detect their activities using homophily as a tool and may resort to disassortative mixing to avoid detection (Keegan et al. 2010). Thus, homophily is a distinguishing trait in covert networks in their incipient stages, but it becomes less of an identifying characteristic with maturity.

Given the imperatives of a covert network to maintain secrecy and, at the same time, be prepared to deliver results, small cells or microstructures are inevitable in covert networks. Clusters (Memon et al. 2008) and cliques (Gill and Freeman 2013) have been proposed as typical microstructures, often based on homophily around ethnicity and occupation, role/skill specialization, or pre-existing ties (Demiroz and Kapucu 2012; Harris-Hogan 2012; Milward and Raab 2006; Raab and Milward 2003). Memon et al. (2008) showed that many terrorist networks exhibit small-world characteristics, with high clustering coefficients and short path length. But, modeling using idealized network structures did not show clustering (Toth et al. 2013). Helfstein and Wright (2011) used model threat vectors to attack networks to show that covert networks that maximize secrecy do not show clustering. Kirby (2007), using Sageman's data (2004), argues that cliques can be self-starters, progress towards fragmentation and isolation, and generate collective identities for members.

To ensure survival, covert networks avoid developing highly centralized actors. As a counterview, Carley et al. (2002) argue that peripheral individuals act as mediators between sub-networks and spread information. This view is in accord with Granovetter (1973), who had argued that weak ties between actors are often bridges that link subgroups

otherwise separated. Several studies indicate that covert networks cluster around specialized actors, leading to functional differentiation within such networks (Calderoni 2011; Malm and Bichler 2011). This specialization tends to become more accentuated over time (Raab and Milward, 2003). Pedazhur and Perlinger (2006) remarked that terrorist networks with many cliques appear to be more effective than those with fewer cliques, which they have interpreted to mean that the existence of cohesive subgroups within a network had proven to be a predictor of the network's effectiveness.

2.6 Density

Some studies reveal covert networks to be sparse or maximally low density (Krebs, 2002; Demiroz and Kapucu 2012; Gimenez-Salinas Framis 2011; Toth et al. 2013) and loosely organized (Natarajan 2000) with micro-structures (Freeman and Gill 2013). Sparseness is a network trait of covert networks that may result from incomplete information (Sparrow 1991) or the actors' efforts to avoid detection. Such organizations also tend to curtail density of ties to prevent any infiltration; apparently, added action or communication between members will lead to a higher chance of detection by law enforcement officials (Krebs 2002; Morselli et al. 2007; Raab and Milward 2003; Xu and Chen 2008; Demiroz and Kapucu 2012). The US Army's counterinsurgency manual (Petraeus 2007) argues that network density is positively associated with network efficiency, a feature that highlights the secrecy-efficiency trade-off in covert networks, i.e., covert networks tend to maintain minimum density at the cost of efficiency. The issue that comes in the way of this hypothesis is the lack of consensus on what defines 'high' density. Adding to the complexity is the variance of density over time in covert networks. Helfstein and Wright (2011) have shown that density increases over time in covert networks. Other factors affecting density are pre-existing ties, which tend to make networks denser (Krebs, 2002; Raab and Milward 2003) and the actors' levels of skills, with more skilled players leading to sparser networks (Helfstein and Wright 2011).

2.7 Centralization, Core-Periphery, and Poly-Centricity

As discussed earlier, the overriding priority in covert networks is maintaining secrecy, even if this happens at the cost of efficiency in outcomes. One of the aspects of secrecy is to protect the actors who matter in the network. To this end, the general assumption has been that covert networks are decentralized (Enders and Su 2007, 51; Clutterbuck 2008; Keegan et al. 2011, p24; Natarajan 2006). But, Baker and Faulkner (1983) studied three illegal networks and demonstrated that they all had centralized structures. In their study, two of the covert networks, switchgear, and transformers, were decentralized while the conspiracy network regarding the steam turbine generator business had a relatively centralized network structure. Baker and Faulkner (1993) relate this variation to the information processing requirements of the organizations. In their study, organizations that do not need high information processing could decentralize their structure for concealment purposes, while organizations that focus on custom production with high information processing need a relatively centralized formation even if it undermines the secrecy of their relationships. Nevertheless, they add that conspiracy's centralized network was successfully separated into a “small core and a large periphery” (Baker and Faulkner, 1993, p855). In other words, even centralized illegal networks disperse their relationships in a way that lowers risks.

Recent studies have also identified a core-periphery divide in covert networks. An idealized core-periphery structure entails that core nodes are well-connected to other core nodes and peripheral nodes and that peripheral nodes are not well-connected to each other (Rombach, Porter, Fowler & Mucha, 2017, Core-periphery structure in networks (revisited). *SIAM Review*, *59*(3), 619-646). Social networks can be modeled using a mixture of local (node-level, dyad-level, etc.), global (involving the entire structure of the network), and mesoscale (intermediate-scale) perspectives. One of the critical uses of network theory is identifying summary statistics for large networks to develop a framework for analyzing and comparing complex structures. In such efforts, the algorithmic identification of mesoscale network structures makes it possible to discover features that might not be apparent either at the local level of nodes and edges or at the

global level of summary statistics. Networks can be described using a mixture of local, global, and intermediate-scale (mesoscale) perspectives used to identify statistics to define large networks, which leads to the development of a framework for analyzing and comparing complex structures.

The mathematical formulation of mesoscale network structures makes it possible to discover features that might not be apparent either at the local level of nodes and edges or at the global level of summary statistics. In particular, considerable effort has gone into the development of statistical formulation and identification, and investigation of a specific type of mesoscale structure known as community structure in which cohesive (and assortative) groups called "communities" consist of nodes that are connected densely to each other and the connections between nodes in different communities are comparatively sparse. Numerous methods have been developed to detect network communities. Some of these methods allow communities to overlap with each other, and others are mutually exclusive. Investigations of community structure have led to insights in several applications spanning fields and voting networks in political science, friendship networks at universities and other schools, protein-protein interaction networks, mobile telephone networks, and criminal networks. Rombach et al. (2017, p.620) report that "Although (and arguably because) studies of community structure have been very successful, other types of mesoscale structures—often in the form of different "block models"—have received much less attention than they deserve. We consider the type of mesoscale network structure in the present paper known as the core-periphery structure."

The type of mesoscale network structure most studied from the perspective of covert networks is the core-periphery structure. Notably, in the Russian Mafia study, Varese (2013) finds a polycentric structure around a few central actors. Demiroz and Kapucu's survey of Turkey's Ergenekon Terrorist Organization (ETO) network, developed from indictment documents, also shows a network with core and periphery densities 0.735 and 0.441, respectively (Demiroz and Kapucu 2012). Raab and Milward (2003) argue that cores with specialized skills (e.g., finance, strategy, planning) benefit terrorist networks, possibly to increase security for central members. Gimenez-Salinas compared criminal

networks in Spain, which all exhibited core-periphery structures (Gimenez-Salinas Framis 2011). This is quite reasonable for an illegal network, where the numbers of connections are minimal while the smaller groups are fragmented. Also, larger network density is usually lower than in smaller networks setting (Hanneman and Riddle 2005). The figures below, adapted from Demiroz and Kapucu (2012), reflect typical core-periphery formation in a covert network. Several other studies have also independently supported these findings (Demiroz et al.; Varese 2013; Cockbain et al. 2011). Other studies have yielded ambivalent results indicating that covert networks may exhibit both centralized and decentralized structures (Crenshaw 2010).

However, not all networks show this property (Lauchs, Keast, and Yousefpour 2011). Morselli et al. (2007) opine that terrorist networks do not show core-periphery structures, whereas criminal networks do, although they employ only one case study of each to support this claim. The diagrams in Figures 2.1(a) and 2.1(b) of a social network drawn only for illustration are examples of core-periphery structures in a social network. It's interesting to note that the core-periphery structure remains unchanged even when the centrality metrics are changed. In Figure 2.1(a), for instance, the centrality measure chosen is the degree centrality, and a few of the nodes with high scores are concentrated in the "core" area. In contrast, those with low scores populate the network's fringes: the "periphery." When the degree centrality metric is replaced with the betweenness centrality, a similar picture emerges (Figure 2.1 (b)).

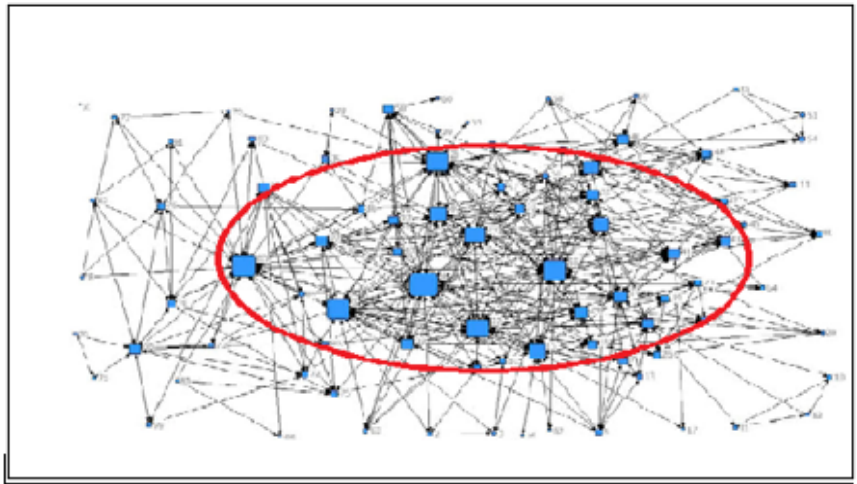


Figure 2.1(a) Core-Periphery structure after degree centrality of nodes are calculated. Node sizes, represented by blue dots are indicative of degree values. The core is demarcated by the red circle. It may be observed that the nodes with high degree are all falling within the circle and those with low scores fall outside it.

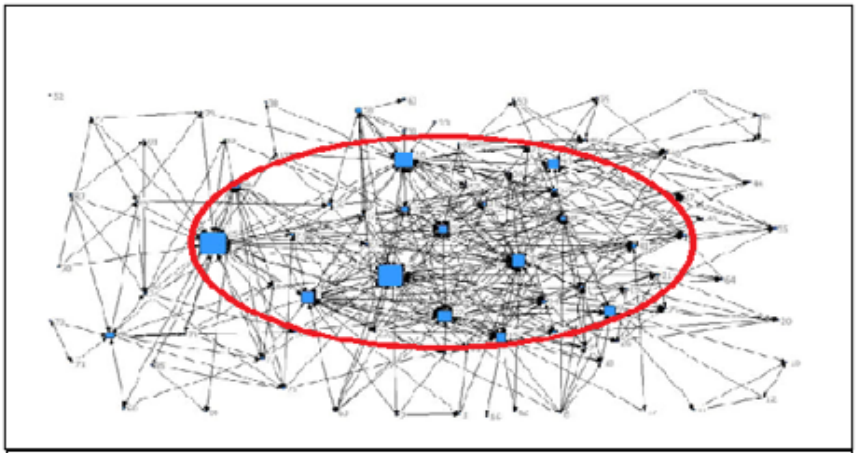


Figure 2.1(b) Core-Periphery structure of the same network after betweenness centrality of nodes are calculated. Node sizes, represented by blue dots are indicative of betweenness scores. The core is demarcated by the red circle. It may be observed that the nodes with high degree are all falling within the circle and those with low scores fall outside it.

2.8 Secrecy Efficiency Trade-Offs

When shaping the structure of their relationships, covert networks necessarily encounter a trade-off between efficiency and safety and prefer the safest path for communication among their members. Ericsson(1981) examined six secret organizations' cases and concluded that one of the essential ways to achieve safety in risky environments is to maintain pre-existing ties. The varying needs for secrecy amongst different covert networks result in significantly different network outcomes and even network structure. Terrorists and organized crime groups have similar concerns concerning secrecy; however, their purposes significantly impact their preference over the safety-efficiency tradeoff. Terrorist organizations aim to create social and political disruption through spectacular events. They tend to remain inert until the most appropriate time comes for an attack (Morselli et al. 2007).

By contrast, organized crime organizations exist primarily for generating monetary gains and aim to maximize their income and try to work most efficiently (Kenney 2007; Morselli 2009). Such organizations tend to prefer efficiency over secrecy because an excessive focus on secrecy results in wastage of time and resources, thereby culminating in a decrease in their gains. For criminals, time is money, while for terrorists, time optimizes opportunity. Enders and Su (2007) discuss two aspects of terror-related networks encountered during intelligence and surveillance efforts. The first approach is the rational-actor model in which terrorist organizations act intending to reach a common purpose. While terrorist organizations are trying to get their destination with limited financial, material, and human resources, they consume certain commodities (e.g., media attention). This model assumes that criminals and terrorists will act rationally to reach the maximum possible output. The second approach focuses on the disruption of networks based on how individuals or cells communicate with each other and how the elimination of individuals or cell units affects the network. The disruption of covert networks is a widely discussed topic with several studies coming out in the last decade or so, specifically to detect key actors or central players and issues resulting in the disruption of these networks (Borgatti

2003, 2006; Carley 2006; Memon et al. 2007; Schwartz and Rouselle 2009; Xu and Chen 2005).

The tradeoff between rational behavior and secrecy related structural priorities forces networks to make choices under external factors, goals, and resources (Baker and Faulkner 1993; Brass et al. 1998; Lindelauf et al. 2009a; 2009b; Raab and Milward 2003). Baker and Faulkner (1993), in their landmark study on conspiracies in the heavy electrical industry in the United States, are amongst the first researchers to have examined the correspondence between secrecy and efficiency methodically. They comment thus –

> "Various practices and organizational devices are used to protect a secret society. Members may conceal the secret society and their involvement in it by limiting face-to-face interaction. Leaders, for example, maybe unknown to ordinary members. Members can increase protection by minimizing the channels of communication. Impersonal communication procedures and decision rules (e.g., the phases-of-the moon system in the switchgear conspiracy) may be used to substitute direct personal communication and negotiation. Organizational buffers can seal off different levels or groups—for example, a graduated division of labor-hierarchy-may separate members of a secret society. Top managers may approve or direct activities but delegate implementation to lower-level operatives. Decentralization limits exposure, making it difficult to uncover an entire network, particularly its leaders. Subversive political movements, for example, are organized into decentralized cells. Secrecy was a paramount consideration in our three price-fixing conspiracies. These criminal networks involved high stakes, major corporations, government buyers, and dozens of corporate managers and executives' careers and reputations, many of whom were pillars of their local communities and elite class members. The conspirators knew their activities were illegal yet continued them despite repeated written directives from the chief executive's office to refrain from meeting with competitors. Given the importance of secrecy, we expect to observe criminal networks that use buffers and other means to maximize concealment. In particular, the need for secrecy should lead conspirators to conceal their activities by creating sparse and decentralized networks. If secrecy were the only consideration, we would expect sparse and decentralized

> communication networks in each of our three conspiracies. But secrecy is not the only consideration. Like participants in legal networks, conspirators have tasks to accomplish, and these tasks must be performed effectively and efficiently. The information must be exchanged quickly and accurately. Problems and disputes must be worked out quickly and smoothly. Most of all, acceptable agreements must be hammered out in time to meet deadlines (e.g., due dates for proposals). Secrecy is critical, but if price-fixing tasks are not performed well, the conspiracy will be a vain and needlessly risky endeavor. Given the need for efficient task performance, what type of communication network is required? The social psychology of small groups and organizational theory agrees that the answer depends on information-processing requirements- the amounts and types of data, knowledge, and intelligence that must be handled to execute a task sequence. Experimental research on small groups has found that simple, routine, unambiguous tasks are performed more efficiently in centralized structures, while difficult, complex, ambiguous tasks are performed more efficiently in decentralized structures. (p.843)"

These remarks cast a clear light on the relationship between secrecy and operational efficiency. Everything depends on the context; if the need for the hour is just to escape surveillance, secrecy becomes paramount. If there are deliverables that can't be delayed, efficiency gains the upper hand. As discussed in the section about centralization, covert networks tend to be decentralized, but internal and external environments' requirements can force centralization in some instances (Raab and Milward 2003).

Interestingly, Helfstein and Wright's (2011) analysis of the six terror networks presents surprising results. They compare networks that prioritize operational secrecy with scale-free networks. The most notable characteristic in a scale-free network is the relative commonness of vertices with a degree that significantly exceeds the average. The highest-degree nodes are often called "hubs" and are thought to serve specific purposes in their networks, although this depends much on the domain. Scale-free networks are typically robust to failure. In such networks, it turns out that smaller ones closely follow the major hubs. In turn, these smaller hubs are followed by other nodes with an even smaller degree

and so on. This hierarchy allows for fault-tolerant behavior. If failures occur at random, and the vast majority of nodes are those with a small degree, the likelihood of a hub being affected is almost negligible. When focusing on operational secrecy, the authors examine the relationship between network structure and its impact on the outcome. Logically, terror networks are expected to form a design that will minimize detection and vulnerabilities. Yet, results show that the networks which are studied did not remain in the most secure state. Helfstein and Wright (2011) argue that when terrorist networks perceive their environment as safe, they act less carefully and have a structure that is more susceptible to disruption. They conclude that risk perception is influential in shaping the network structure of terror groups.

Similar approaches apply graph-theoretic metrics to recognize and understand proposed network structural properties (Lindelauf, Borm, and Hamers (2009a)). They compared covert communication network models to find structures with optimal trade-offs between secrecy to avoid detection and efficiency of information flow to coordinate and control cell members. The model's optimality depends on the assumptions about all cell members' likely exposure if any of the actors are randomly removed. A star graph in which all members communicate only through the commander is an example of an optimal structure for balancing the conflicting objectives if one member's detection also exposes all his links to the other cell members. In contrast, if the probability of exposure is a function of the node centrality in the network, the optimal structures are reinforced rings and reinforced wheel graphs. Many related models making different exposure assumptions and imbalanced secrecy-efficiency trade-offs identify other optimal systems (Lindelauf, Borm, & Hamers, 2009b, 2011). However, many of these graph theoretic models tend to exclude law enforcement agencies that actively seek to detect and disrupt the terror networks.

A color gradient showing the secrecy-efficiency trade-off is shown in Figure 2.2:

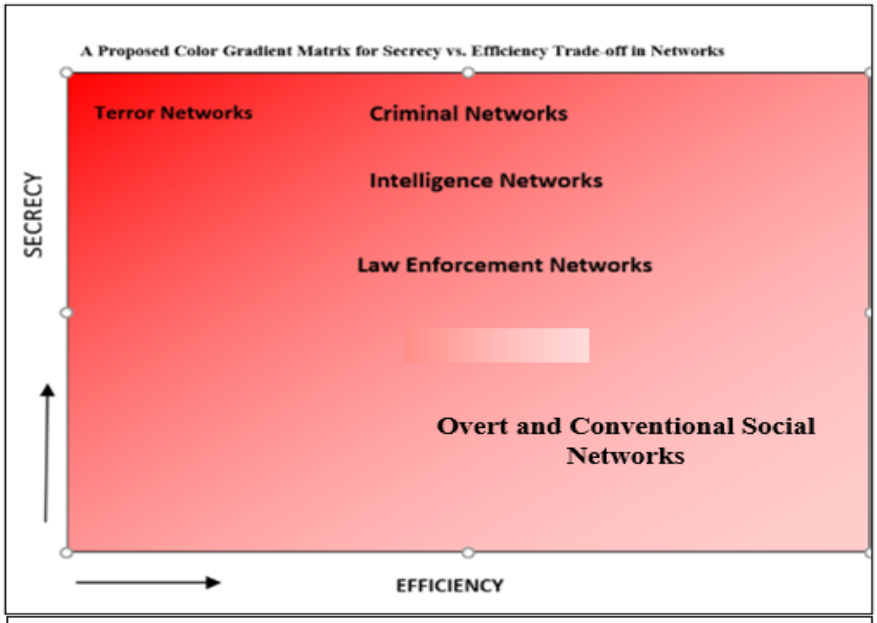


Figure 2.2 A color gradient matrix showing the trade-offs between secrecy and efficiency of different types of social networks. Terrorist networks which need to have secrecy as their paramount requirement and efficiency is not very vital, occur at the top left of the matrix. Other covert networks which have paramount needs for secrecy with varying needs for efficiency occur at the middle of the matrix at differing heights. It's interesting to note that organized criminal networks vary in the trade-off behavior from terrorist network in that they need some quantum of efficiency in order to be functional. Other networks which occur towards the middle of the matrix, include organizations like Intelligence agencies, Police and Law Enforcement departments, all of whom have secrecy requirements but need to deliver on efficiency to some extent as well. Overt or conventional networks, whose need for secrecy is limited and efficiency is the driving concern populate the lower right quadrant of the color coded matrix.

2.9 Path Distance

The average path distance between a node and any random peer in the network is an essential indicator of network safety (Ayling 2009; Kenney 2007; Williams 2001). Large and small networks have different structural characteristics. The size determines how path distance can determine the efficiency of actions taken by network members. Xu and Chen (2008) mention Global Salafi Jihadists as an example of an extensive network with a relatively short mean path distance (2.5 links). Cliques in dark networks would have "denser and stronger relationships with one another" (Xu and Chen 2008, p. 62) but, to avoid detection, cliques tend to be segregated. Krebs's (2002) study of the 9/11 attackers

demonstrates how subgroup networks work. His paper indicates that the 9/11 terrorist network structure had a serpent-like shape (Demiroz et al. 2012). Nodes in the system were connected to a limited number of actors that aimed to lower the risk of detection for the entire network if one or more members within the network were caught.

Krebs (2002, p 46) highlights Bin Laden's statement about this strategy: "Those who were trained to fly didn't know the others. One group of people did not know the other group". He also indicates that the network actors had an average path length distance of 4.75 from each other. This is a relatively long distance of communication in a network of 20 nodes. It is important to note that this path length applied to the specific hijackers (i.e., functional group) involved in the 9/11 attacks. When complementary actors, who were not physically involved in the 9/11 attacks, are added to the network, the average path length decreases from 4.75 to 2.94 (Krebs 2002). One implication is that monitoring the peripheral actors would enable law enforcement more efficiently to reach (i.e., detect) the core actors, reflecting that based on the sensitivity and importance of tasks and actions, security necessities and secrecy precautions may vary.

2.10 Applicability of Social Network Analysis to Covert Networks

As is evident from the previous passages, covert networks' analysis presents challenges distinct from ordinary social network research. Traditional SNA investigates networks where the data are virtually complete. By contrast, covert networks pose difficulties that are not easily handled by conventional social network analysis tools. A known nefarious actor who is being observed is concerned about being followed. So, the actors' regular interactions are not likely to be the interactions of interest. It is expected that the interactions with individuals of interest are the most difficult to watch. So naïve network analysis would place the highest emphasis on the least essential edges and the lowest priority on the most critical edges.

As seen from their defining characteristics, covert networks share many traits with overt networks and are separable only by specific prominent features. All four properties identified by Sparrow (1991) and described earlier in this paper are to be seen in any given social network, possibly to a lesser degree. Many networks about broad social movements or religious organizations are vast in size. Incompleteness and fuzziness can result from inadequate information and research about a network rather than deliberate attempts by the actors to hide linkages. Similarly, most networks will change in time, whether as a result of actors changing their roles or leaving the network or only as a reaction to changing social mores and traditions. One of the challenges in analyzing a covert network is to ascertain how well the traditional concepts of SNA measure. Sparrow (1991) opines that the traits typical of covert networks produce computational difficulties. Such characteristics demand algorithmic complexity and require substantial advances in methods of statistical (or computational) inference and add that "these properties will likely render some of the existing network theory concepts less useful than others. For example, the fuzzy boundaries render precise global network measures (such as radius, diameter, even density) almost meaningless. With the global measures go some, but not all, centrality" (Sparrow 1991, 263).

Freeman (2004) defines four elements that typically define social network analysis: (i) Motivation by a structural intuition based on ties linking social actors, (ii) Research-based on systematic, empirical data, (iii) Utilization of graphic imagery; and (iv) Employment of mathematical and computational models to predict future behavior. Of the four elements, the latter two's performance depends on the first two's completeness and accuracy. Many of the ties binding the leading players are likely forged before the formal network structure evolved, and any conclusions based on the extant structure can yield wrong results. As mentioned, the relations between the key players in the Al Qaeda network were formed in Afghanistan when they fought the Soviets. After Al Qaeda attained notoriety, the network structure that was unveiled by the researchers initially did not reflect the true nature of the principal players. The second element defined by Freeman regarding research based on systematic and empirical data may not be feasible in the case of covert networks where the need to repress leakage of information related to the structure may be paramount. If any

systematic information regarding such networks were found, the network would lose its covertness tag.

In the 9/11 attacks, the research timeframe began when the cell had finalized its intended modus operandi and covertly established itself to undertake its objectives. The deadline ceased after the operation or at the stage when authorities disrupted the cell. Sparrow (1991) suggests looking at the waxing and waning strength of a tie depending upon the time and task instead of looking at the presence or absence of a relation between two individuals. Such observations require longitudinal monitoring of cell members' activities, which only law enforcement has sufficient resources to conduct.

Covert networks generally include several actors who neither have an unusually large number of ties nor are connectors between different parts of the network. Such networks have players situated in a strategic location in terms of their proximity to hubs (or highly active nodes) or large numbers of members. Hence, these members have a high level of access to information and resources. How can these three critical kinds of actors be detected? How can we compare the influence of different actors in the same or different networks? The most common techniques for evaluating the actor's role and power within the network are the traditional measures of centrality. Several prominent, relevant actions that can effectively study violent groups are the degree of centrality, closeness (limited to connected networks and large components), and betweenness. A similar measure was introduced by Brams, Hande, and Ramirez (2006), who developed the concept of influence as a function of the actor's importance within the network. This, in turn, is determined by the number of their ties and their directions relative to the other actors. Baker and Faulkner (1994) suggest looking at archival data to derive relationship data. The data they used to analyze illegal price-fixing networks were mostly court documents and sworn testimony. These data included accounts of observed interpersonal relationships from various witnesses.

The seminal effort to link social network analysis to covert networks through analyzing law enforcement and intelligence efforts and, by extension, to covert networks was perhaps

that of Malcolm Sparrow of the Kennedy School of Government, Harvard University in 1991. There were earlier sporadic efforts, e.g., Coady (1985), Howlett (1980), Davis (1981), but these had focused only on basic network concepts and lacked sophistication). Sparrow looked at the methodologies used by law enforcement agencies akin to those prevalent in the SNA domain. He found that intelligence (and by extension, law enforcement) agencies, despite an evident awareness of the importance of analysis, had remained, for the most part, unsophisticated in their use of analytic tools and concepts. (This has changed over the years, and law enforcement agencies have increasingly adopted sophisticated mechanisms to study crime and criminality patterns, including advanced network analytics). Sparrow observed that agencies typically had plenty of data, much of it computerized, but they had comparatively little capability to extract useful intelligence from it.

According to Sparrow, even the primitive tools (primitive by the standards of existing social network analysis in other domains) that were in use by these agencies were simple forms of SNA. He expressed disappointment over the lack of overlap between the literature of social network analysis and that of law enforcement. He also lamented these agencies' tendency to score individual successes in taking down criminal entities they perceived as crucial without waiting to uncover the entire network or identifying strategic vulnerabilities. The first approach, i.e., to strike before the whole network is exposed, is especially problematic as it is "difficult, dangerous, time-consuming and expensive" (Sparrow 1991, 260).

The efforts to apply SNA to covert networks have been primarily oriented to optimize the process of disruption. In other words, researchers have tried to identify actors or subsets within the covert networks whose removal will likely affect the functioning of the network most adversely. This approach is not without its critics, and many studies indicate diminishing returns from disruption and may allow key actors to develop methods to increase resilience and avoid detection (Tvestovat and Carley 2003; Brafman and Beckstrom 2006; Bouchard 2007; Ayling 2009; Everton 2011; Dujin et al. 2014). Disruption is thought to make covert networks more decentralized and hence, more

difficult to target. In the case of terrorist networks, it only brings about what Sageman (2008) refers to as the “leaderless jihad,” by which he means the evolution of numerous independent and local groups that use Al Qaeda as a franchisee rather than act as an organic part of it.

2.11 Defining Covertness as a Unit Attribute

The previous section broadly encapsulates the types and features of social networks described as dark or covert. This study's primary objectives are to break down the notion of what constitutes a covert network to a more fundamental and functional unit of measurement. The widely varying perceptions about how a covert or dark network might be defined have caused a palpable lack of consensus amongst sociologists and criminologists alike. Plus, the situations where entire networks can be categorized as covert are not very common. Usually, such networks come as parts of much larger networks that might not be covert or representative of any criminal enterprise.

The present research deals with the ENRON enterprise, which was by no means a criminal organization. If anything, it was a widely respected Fortune 50 company with worldwide dealings that were anything but covert. But, the insider trading scam that was budding within the overt confines of the company’s employees network can’t be kept outside the scope of blatantly criminally oriented behavior. Yet, the number of employees who participated in this covert undertaking under a mundane façade is minuscule. Suppose we stick to the convention of labeling an entire network as covert or criminal. In that case, it will entail a severe lack of understanding about the overall network’s function, and many actors who are innocent and in no way connected will be unjustly tagged.

Thus, the aim is to devise a mechanism to decompose the concept of “covert” into a more atomic unit, which will serve as bricks for constructing more uniform covert models instead of wholesale labeling. This ground-up approach towards identifying covertness within a network achieves the twin objectives of avoiding wholesale negative labeling of

a network and pointing out specific parts of an otherwise innocuous network worthy of further scrutiny regarding their outputs objectives. For example, in the case of the 9/11 attackers, if we consider all the contacts of the attackers, including flight trainers, the co-passengers with whom they might have interacted, different types of service providers who they might have liaisoned with during their stay in the United States, then the network size becomes huge (Krebs, 2002). If we add the number of visitors from the Middle East to the United States during that period, the network will have thousands of nodes, if not more. And only a small piece of this network was involved in the planning and execution of networks. Imagine a surveillance agency that was not aware of the conspiracy was scanning the broader network of covert activity; what might it have done with the humongous information that would've come its way.

As discussed in the previous sections, laws on privacy, the widespread use of encrypted messaging systems, and even the simple lack of complete information about a network are formidable barriers to forming any opinions about covertness. There has to be a meaningful way to capture parts of the information in a manner that can be analyzed and dissected on-the-fly to identify at least some portion of a subgroup that might be planning covertly to achieve some wrongful objective. There is a need to add an attribute that can be applied universally on all networks. Furthermore, this universal attribute should be so ubiquitous that it defeats any privacy, encryption, or incompleteness barrier.

To successfully predict covert sub-groups within a network, the proposed attribute should have some key features that may correspond to the properties noted below.

(a) The attribute needs to be based around the easily observable *topological features* of a network akin to some of the more popular centrality measures

(b) The attribute should be a *minimalistic* one that will not require many components or complicated manipulations of input variables; in other words, it shouldn't consume too many computational resources or bandwidth to calculate.

(c) The attribute must be capable of being used for *non-intrusive analysis*, i.e., there should be no need for contents of the information exchanges to be known. It's especially essential to bypass existing data privacy laws, encryption mechanisms, varying policies across countries that allow differential access to information, legal strictures, inadequate information about the network, and covert communication channels not apparent to the surveillance team.

(d) It should be easily *applied* across all networks (node or edge or even higher groupings). That is, the attribute should be of such a nature that it can be combined easily with existing metrics.

(e) The attribute should have the ability to act as a *linkage* mechanism, tying together disparate nodes, edges, triads, or higher group formations based on some formulation of commonness (or collusion as the term is viewed in this study). A corollary of this characteristic is that the attribute should exhibit some form of *structural transcendence*, i.e., there needs to be some tangible structural links between entities that might end up being grouped through its application. This aspect is covered later on in some detail in the section, which discusses the methods of building a *collusion index*.

(f) More importantly, it should allow the investigator to *reduce* the sheer volume of exciting data and likely yield results. That is, the size of the haystack should be reduced considerably. Most networks (even the smaller ones) have tens of thousands of nodes, edges, and other related structures. The resources needed to detect covert activities are usually not commensurate to cope with the massive volumes of data generated; the information immediately available for processing must be manageable and still yield meaningful results.

(g) The attribute should be *dynamic*, i.e., it should be able to change in time and still retain its usefulness despite evolutionary changes in the network structure.

If there are changes in the network's topology, the attribute changes should track the changes and even predict the network's future shape.

The approaches to developing an attribute with all or some of these properties can be categorized into two broad classes. The first approach is node-centric, i.e., the attribute is defined on a node. The value produced by the attribute is assigned to it, much like the popular centrality measures such as degree centrality, closeness centrality, betweenness centrality, and eigenvalue centrality. The second approach is to develop the attribute based on ties or edges or a multiple thereof (i.e., a sub-network). Of these two approaches, the first one, based on nodes, is, by far, the most popular.

2.12 Covertness as a Centrality Measure

Some recent studies on developing node-based covertness indices by sociologists and network scientists, most notably Ovelgonne, Kang, Sawant, and Subrahmanian (2012), have proposed such a node centric covertness centrality measure. Their covertness centrality measure consists of two parts: how “common” a node concerning a centrality measure and how well the node can “communicate” with a user-specified set of vertices.

As they define it, commonness is a measure of how well a node hides in a crowd of similar nodes. Depending on the range of values, the commonness may reflect one of two properties: Optimal Hiding or No Hiding. With *Optimal Hiding*, all nodes are equal vis-a-vis all centrality measures which have been considered. In this case, the nodes are indistinguishable from each other, and so the hiding is optimal. In this case, the commonness has to be 1 for all nodes. With *No Hiding*, if a node is not similar (a similarity measure needs to be defined for this) to any other node concerning any centrality measure, this node's commonness should be 0. Hence, based on the commonness scores, any node in a network will have a value that varies between 0 and 1. A threshold value can be defined to make any set of nodes stand out from the others.

The next characteristic of a node's covertness centrality is its communication potential, which Ovelgonne and colleagues (2012) have defined as an attribute that reflects a node's ability to communicate and cooperate with other nodes to achieve a common objective. The answer as to which communication and cooperation options are essential to achieve that objective rests solely on the researcher and the study's nature. Finally, the node's covertness centrality is calculated based on a combination of its commonness and communication potential. The exact recipe of this combination is again a function of the circumstances where the attribute is sought to be applied.

Memon (2012) has suggested a hybrid approach to covertness centrality by adding weights to ties or edges between nodes in a social network. His study has focused on covert networks in general and on terrorist networks in particular. He adds edge weights to traditional centrality measures like degree centrality, closeness centrality, betweenness centrality, and the shortest path mechanism increase differentiability of identifying the central characters in a terrorist network. This variant of the famous critical node detection problem helps solve problems across different disciplines that deal with networks (or graphs). A further variation of this approach was proposed by Newman (2001c) and Brandes (2001) independently, who used the notion of *Inverted tie strengths* while extending closeness and betweenness centrality, respectively. Thus, the resulting tie weights can be considered costs since weak (and costly) ties have high values, and healthy (and cheap) ties have low values. Hence, the higher a link's weight, the stronger it is, and the less it costs to transmit information along with that link. However, the extra dimension added by using weighted ties or edges has also been applied to node centrality. Hence, it belongs more or less to the node centric approach of detection of covertness.

The second category of covertness centrality applies to edges, triangles, and other graph-like structures in a network with higher dimensions than a node. Newman and Girvan (2002) enunciated the principle of *edge-betweenness* centrality, which they defined as the number of shortest paths that go through an edge in a graph or network. In this manner, each edge in the network can be associated with a score that is an *edge betweenness*

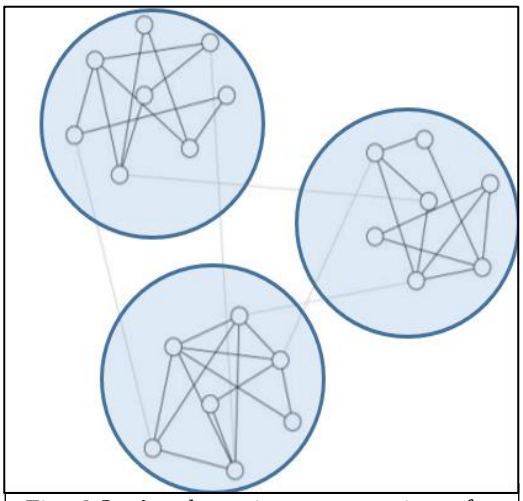


Fig. 2.3. A schematic representation of a network with group structures linked with each other and each group being considered as a nodal unit. In this network there are three communities of densely connected nodes (circles with solid lines), with a much lower density of connections (thinner lines) between them.

Centrality value. This measure is an extrapolation of Freeman's node-betweenness centrality to edges. The concept was used to contrast the hierarchical clustering models with a dendrogram like appearance and had a divisive approach instead. The edge-betweenness values were calculated iteratively for effecting division partitions of the network. This concept, which extends a node centrality measure to an edge, may well be extrapolated to higher graph formations within a network. For instance, if a set of homogeneous communities is identified within the network confines, each community may be treated as a nodal entity, and node-centrality measures can again be applied. In this study, a similar approach is undertaken with a *covertness index* being developed based on information exchanges between a pair of nodes and treating the node-pair (termed as a *dyad*) as a single unit instead of a node. The approach in this paper has been aimed at developing the covertness index as an edge-based property and leveraging this index through a ranking mechanism (higher the value of the index, higher the rank) to prune the network into a more manageable size and then identify community structures as a second step through a separate similarity index. As a further measure, each agglomerative (community or group) is treated as a single nodal unit. The index is then used iteratively until the desired covert communities are identified to the surveillance unit's satisfaction.

2.13 Recent Research

Research into the analysis of covert networks has been relatively extensive in current times. Driving this upsurge has been the recent terror and the consequent desire to analyze these attacks' causes and identify the main actors and structures. During this renewed focus on

terror, interest has also spilled over increasingly into the domain of other networks that are somehow "darker." This category includes organized criminal groups, insurgencies, oppositional political movements (in countries that look upon opposition from an adversarial viewpoint), foreign (and potentially hostile) intelligence networks, etc. Cutting edge techniques in tackling covert networks are increasingly occurring at the intersection of mathematics, game theory, and computer science, especially the newly emergent fields of artificial intelligence and machine learning. These newer technologies prove to be a game-changer in expanding research beyond the necessary case studies of covert networks. Many of the recent studies center around the formulation of tactics to destabilize such networks. Many mathematical constructs and related computer algorithms can fine-tune network parameters to evaluate possible covert structures and then explore the incremental disruption of its ways to incrementally disrupt its activities.

Many of the recent network models are not based on real data but on sophisticated simulations of network functionalities with synthetic datasets. One of the more exciting developments in simulated studies of networks is the concept of a generative adversarial network (GAN)[11], which is a class of machine-learning frameworks in which two neural networks are competing with each other in a game (in the sense of game theory, often but not always in the form of a zero-sum game). This technique learns to generate new data with the same statistics as the training set(Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, and Bengie, 2014). A second approach has been collecting vast quantities of information from open sources, such as television, news reports, social media, and the internet, and then mine these sources for patterns that zero in on the key actors, their relations, and characteristics (Krebs, 2002). These newly evolved techniques have found applications in two main areas: networks of organized criminal syndicates and networks formed by terrorist organizations. Predictably, more recent research has focused on terrorist networks owing to numerous terror-related events around the world since the nineties.

We may broadly divide the current mathematical and computational methodologies and techniques into two broad categories: (1) *Interventionist* and (2) *Predictive*. Interventionist

[11]GAN modeling was developed by Ian Goodfellow and his colleagues in 2014.

strategies comprise, by far, the most predominant narrative in research on covert networks. As the name suggests, the techniques employed in this category study networks after some incidents have occurred (major criminal or terror-related incidents), i.e., interventionist methods are *post-hoc* in their approach. Predictive approaches are challenging to find in current day research literature and have started gaining traction lately. Such practices are considered to be a subset of the broader crime analysis platforms. The main driver behind crime analysis methodologies is that crime is not a one-off random event but takes place in ways that fit into patterns (Brantingham 1981; Tayebi, Ester, Glässer, & Brantingham; Santos 2014; Tayebi, Glässer, Alhajj 2016; Smith, Santos & Roberto 2018). Predictive crime analysis encompasses understanding criminal behavior patterns and the extraction of these patterns to predict crime and interdict it.

2.13.1 Interventionist Methodologies

Interventional procedures require a general understanding of the network being studied, the identities of at least some of the principal actors, the structure (or at least an estimate of it) of the network, and the outcomes that have happened because of the activities of the network. Network analysis of this nature happens only when the analyst is well aware of the network's nature and plies his trade tools on the observable patterns. Interventionist strategies can be further sub-divided into content-based or *interceptive* techniques and *structural* analysis.

2.13.1.1 Interceptive Strategies

Interceptive methods involve wire-tapping of phones and emails (through legal or extra-legal means), contents of messages exchanged on web-based platforms like WhatsApp, Facebook, Weibo, etc. (if these messages are available in unencrypted format), forensically extracted from the cell-phones and computers of persons of interest, etc. Occasionally, such information may be retrieved from suspects through questioning (both legal and extra-legal). The contents then are made to undergo semantic analysis through computational and manual means. Terminologies of interest (words like “terror,” “bomb,” explosive,”

“hawala” are good examples) or any alias terms used in place of these (codes) are extracted, and the mails or messages or voice recordings containing such terms are culled out for more in-depth analysis. This sort of methodology is the most popular amongst law enforcement analysts. It offers the richest and most accessible “pickings” to identify suspects and their plans and activities. Though easy on computational terms, such methods are difficult to implement in the face of increasing legal hurdles. Existing wire-tapping policies in countries that enjoy more democratic space are difficult to obtain, and attempts to get these through less savory methods entail stiff legal costs. Coming in the way of retrieval of content-based data is the increasing use of sophisticated encryption (WhatsApp uses a 256-bit encryption protocol to encrypt its messages), and the possession of an encrypted message is as good as not having it since breaking the code will take an immense amount of effort and time.

2.13.1.2 Structural Strategies

Structural mechanisms to analyze covert network behavior are less direct than interventionist techniques and are more subtle in practice. Structural methods exploit the graph-like structure of social networks and such networks' tendency to broadly follow graph theory's established paradigms. There should be a reasonable inference about the actors of interest, the ties amongst them, the immediate historical context of the network's evolution, etc. Structural analysis may also encompass non-content based knowledge such as meta-data, tie structures, the centrality of certain actors, etc. The peculiar to covert network structures are enumerated at length in the previous sections, and these structures are given the most significant attention in this approach. Interventionist structural approached are often a hybrid of content and non-content based techniques, including the determination of shortest paths between the actors of interest, i.e., actors who might have participated in the covert activity (Malm, Kinney & Pollard 2008; Magalingam, Davis &Rao 2015), security-efficiency trade-offs (Everton & Cunningham 2015), variants of network modularity (Newman 2006; Ferrara, De Meo, Fiumara& Baumgartner 2014), etc.

2.13.1.3 Disrupting Covert Networks

A subset of the structural approach is the methodology of disrupting covert or dark networks by removing nodes deemed crucial to their functioning (Everton 2008; Roberts & Everton 2011; Everton 2012; Everton & Cunningham 2013; Roberts & Everton 2016; Cunningham, Everton & Murphy 2016). Farley (2003, 2007,2009) applied the mathematical theory of ordered sets to calculate the threshold at which a terrorist group ceases to be actionable when key participants are removed. Farley's model rejects the idea of modeling a terrorist network as a pure graph and then destabilizing it by removing key players (Carley, Lee & Krackhardt, 2001) and instead assumes a hierarchical cell structure of leaders' followers. The technique involves searching for the network's *cutset*[12], the network actors whose removal breaks all vertical chains of command linking the commanders to the actual operatives. Keller, Desouza, & Lin (2010) expand upon this concept by defining a four-level strategy to disrupt terrorist (and dark) networks, namely, targeting leadership, targeting grassroots workers or foot soldiers, targeting specific geographies, and targeting random structures within the system. One of the drawbacks of disruption strategies is that they do not factor into such networks' resilience. Knoke (2015, p.5) feels that the "mathematical model is moot for real terrorist groups that are not structured as hierarchical communication networks," which is a fair reflection of the sheer diversity of structures of such networks in the real world.

2.13.2 Predictive Methodologies

Predictive crime analysis has two main functions: strategic and tactical. Strategic analysis is about examining long-term crime trends. Tactical based research concentrates on short-term and immediate problems to investigate the relationship between suspects and crime incidents. Predictive techniques use network analysis to identify potential suspects with the goal of crime prevention—efficient predictive analysis results in policing, which is more proactive and less reactive. One of the principal objectives of forward-looking crime analysis is generating information that can enhance decision making for optimally

[12]A collection of nodes that intersects every maximal chain is called a cutset.

deploying police resources to prevent criminal activity. Tayebi and Glasser (2016) describe predictive policing as a tool for mining information to mitigate crimes. Although neither a "crystal ball" nor a substitute for integrated solutions- "With predictive analysis, this process becomes more efficient and effective using the discovered patterns about crime locations, crime incidents, crime victims, criminals, criminal groups, and criminal networks"(p.2). They add the caveat that "predictive policing methods are neither a substitute for integrated solutions to policing nor equivalent toa crystal ball that can foretell the future" (p.2). They describe predictive networks as having the power to facilitate proactive policing and improve intervention strategies employing efficient use of limited resources. "These methods give law enforcement agencies a set of tools to do more with less" (p.3). Analysts have long realized the importance of analyzing conspiracy networks—networks of offenders who have committed crimes together (Wasserman & Faust, 1993) for designing prevention and interdiction tactics. Thus, one of the priority tasks in predictive network techniques is analyzing the relationships between malfeasant actors to learn the connivance and collaboration patterns. Tayebi and Glasser add a note of caution on its use, stating:

> "Despite the importance of co-offending network analysis for public safety, computational methods for analyzing large-scale networks are rather premature. Contrary to other social networks, concealment of activities and actors' identity is a common characteristic of co-offending networks. Still, the network topology is a primary source of information for predictive tasks" (p.3).

The comparative novelty of the predictive approach is laid bare by Tayebi and Glasser (2016, p.3), who states thus – "To the best of our knowledge, this work is the first comprehensive attempt to use co-offending network analysis in predictive policing suggesting a paradigm shift in the way co-offending network analysis is used for crime reduction and prevention."

Many people think of Wall Street and hedge funds when they think of predictive methodologies used to make futuristic projections in network models. In her acclaimed book, 'The Weapons of Math Destruction,' O'Neil (2016) talks about financial Weapons

of Math Destruction (WMD) and her experiences in the same realm. Still, the examples in her book come from many other facets of life as well: college rankings, employment application screeners, policing and sentencing algorithms, workplace wellness programs, and the many inappropriate ways credit scores reward the rich and punish the poor. As an example of the latter, she shares the galling statistic that "in Florida, adults with clean driving records and poor credit scores paid an average of $1552 more than the same drivers with excellent credit."

She shares stories of people who have been deemed unworthy in some way by an algorithm. She relates the instance of a highly-regarded teacher who is fired due to a low score on an inscrutable teacher assessment tool, the college student who couldn't get a minimum wage job at a grocery store due to his answers on a personality test, the people whose credit card spending limits were lowered because they shopped at certain stores. To add insult to injury, the algorithms that judged them are completely opaque and unassailable. People often have no recourse when the algorithm makes a mistake.

Many algorithms create feedback loops that perpetuate injustice. Crime recidivism models and predictive policing algorithms—programs that send officers to patrol certain locations based on crime data—are rife with the potential for harmful feedback loops. For example, a recidivism model may ask about the person's first encounter with law enforcement. Due to racist policing practices such as stop and frisk, black people are likely to have that first encounter earlier than white people. If the model takes this measure into account, it will probably deem a black person more likely. But they are harmful even beyond their potential to be racist.

O'Neil writes,

> A person who scores as 'high risk' is likely to be unemployed and to come from a neighborhood where many of his friends and family have had run-ins with the law. Thanks in part to the resulting high score on the evaluation, he gets a longer sentence, locking him away for more years in a prison where he's surrounded by fellow criminals—which raises the likelihood that he'll return to prison. He is finally

> released into the same poor neighborhood, this time with a criminal record, which makes it that much harder to find a job. If he commits another crime, the recidivism model can claim another success. But in fact, the model itself contributes to a toxic cycle and helps to sustain it.

As O'Neil eloquently demonstrates, the problem is that these algorithms are often inherently incapable of comprehending real-world problems and hence, incapable of expressing their solutions as well. In O'Neil's words, "mathematical models should be our tools, not our masters."

There has been a great deal of research on achieving algorithmic accountability and transparency in automated decision-making systems - especially for those used in public governance. However, good accountability in the implementation and use of automated decision-making systems is far from simple. It involves multiple overlapping institutional, technical, and political considerations and becomes all the more complex in the context of machine learning-based rather than rule-based decision systems.

Goldenfein (2019) argues that

> relying on human oversight of automated systems, so-called 'human-in-the-loop' approaches, is entirely deficient. It suggests addressing transparency and accountability during the procurement phase of machine learning systems - during their specification and parameterization - is critical. In a machine learning-based automated decision system, the accountability typically associated with a public official making a decision has already been displaced into the actions and decisions of those creating the system - the bureaucrats and engineers involved in building the relevant models, curating the datasets, and implementing a system. (p.1).

There are many accountability mechanisms available for developers of predictive algorithms and mathematical-statistical models to consider, including new computational transparency mechanisms, fairness, and non-discrimination of decisions. An exercise of this nature proceeding without understanding the complexities and limitations of those accountability and transparency ideas risks disempowering public officials in the face of

increasingly complex machine-led decision making. This dissertation aims at making predictive algorithms in the realm of social networks as transparent and as 'explainable' as possible, and at the same time, desisting from exploiting any 'intrusive' content of communications.

2.14 Recent Developments

Improved computational algorithms, especially those connected with neural learning, allow massive-scale simulations of terrorist networks. A good example is generative adversarial networks or GANs (Goodfellow et al. 2015), which can typically generate networks comprising tens of thousands of nodes. The simulation of such large scale networks allows analysts to study the impact of different counterterror strategies on the resiliency of criminal networks and the capacity to conduct future operations. One of the methods adopted for simulations is based on the concept of agent-based modeling (ABM) where an agent, say, an individual terrorist, or a criminal operative like a drug trafficker, is considered an actor within the network, methods involve "(i) the simulation of automated agent behaviors and interactions in the context of their environments; (ii) the analysis of macro-level patterns resulting from micro-level agent interactions" (Keller, Desouza, &Lin, 2010, p. 1020. An example of ABM network analytics is the StochasticOpponent Modeling Agent (SOMA), which uses textual data extracted from document sources to generate rules explaining a terrorist group's behavior. The SOMA Terror Organization Portal (STOP) features the SOMA Extraction Engine (SEE), the SOMA Adversarial Forecast Engine (SAFE), and the SOMA Analyst Network (SANE) "that allows analysts to find other analysts doing similar work, share findings with them, and let consensus findings emerge."(Sliva, Subrahmanian, Martinez, & Simari, 2008, p.1).

Researchers applied STOP to 25 years of monthly data on the Pakistan-based Lashkar-e-Taiba, also called Lashkar-e-Tayyaba (LeT), a proxy militia cum terror group that acts as a proxy for the Pakistan Armed Forces (Sliva, Subrahmanian, Mannes, & Shakarian, 2011). The data used for their analysis of LeT'scharacteristics is part of the Computational Modeling of Terrorism (CMOT) Project, "which is a specialized codebook for developing

datasets on terrorist and other violent organizations throughout the world. Also, besides to LeT, we have collected data for Jaish-e-Mohammed (JeM), Indian Mujahideen (Mujahideen fi-al Hind), Students Islamic Movement of India, Forces Democratique de Liberation du Rwanda (FDLR), and many others. The CMOT data is an example of a behavioral time-series dataset, a class of relational time-series databases that can be used to describe the context and behavior of an agent or group." (p.1).Their model inferred ten rules from the behavioral time series set that predicted when LeT would likely launch offensives against targets in India. These rules could "provide accurate probabilistic forecasts for both real and hypothetical situations," helping law enforcement agencies in India and make optimal deployment and interdiction decisions (p. 6).

2.15 Cutting-edge Tools

Other popular data-mining, event-forecasting, link-prediction models, and tool-based methods include the random walk based tool CrimeWalker (Tayebi, Jamali, ester, Glasser & Frank 2011); CrimeTracer, a supervised learning framework for co-offense prediction (Tayebi et al. 2016); data-mining based link prediction models (Mahesh, Mahesh, and Vinayababu 2010); strategy equilibrium based models (Arce, Croson, and Eckel, (2011); COPLINK, a user collaborative based commercially available model from M/s Forensic Logic; web content-based detection methodology and disruption strategies(Chaurasia, Dhakar, Tiwari, and Gupta 2012); deterministic similarity-based edge prediction methods with a model-based probabilistic approach (Liben-Nowell and Kleinberg 2007; Hanneke, Fu, and Xing 2010); and the same models applied to the ITERATE[13] dataset(Desmaraisand Cranmer 2013) and prediction models based on hidden Markov models (Petroff, Bond, and Bond 2013).

There are other excellent models with agent simulation frameworks and game-theoretic approaches. Some of the well-known works in this area are the Hats Simulator

[13]The ITERATE data are one of the most comprehensive and commonly used data sets on transnational terrorism proposed by Mickolus, Sandler, Murdock, &Flemming (2008). These data are well suited to research purposes as they cover all transnational terrorist attacks over a 34 year period (1968–2002).

(Cohen2004), Game-theoretic results (Sandler 2008), and Dynamic Network Analysis (Carley 2006). The third area of simulation looks at more detailed terrorist (and by extension, other covert and criminal) activities. One of the models used frequently in this area is the one developed by Pattipati (2006), which uses HMM[14] models of terrorist activities. That work is described further in (Singh 2006).One of the more practical models that integrate one or more of all these formats is the Counter-Terror Social Network Analysis and Intent Recognition or CT-SNAIR developed by Weinstein, Campbell, Delaney&O'Leary (2009) of MIT Laboratories. Carley's (2003) Dynamic Network Analysis (DNA) package, which treats terrorist groups as "complex dynamic networked systems that evolve" (Carley, 2006, p. 1), combines traditional social network analysis of ties between nodes with multiagent modeling to connect nodes, locations, events, tasks, knowledge, resources, and other elements.

Dynamic network analysis is perhaps the first primary "toolkit" that has emerged in the quest for an integrated software platform that combines traditional and labor-intensive social network analysis with networks' computational perspectives. "DNA combines the methods and techniques of SNA and link analysis with multi-agent simulation techniques to afford analysts with a set of techniques and tools for investigating complex and dynamic socio-technical systems" (Diesner & Carley, 2004, p. 1). In other words, it is more than just a toolkit. It resembles a "suite of tools," i.e., an ensemble of computer algorithms for extracting data from texts, mapping networks of words in texts, and forecasting changes."Map analysis systematically extracts and analyzes the links between words in a text to model the author's 'mental map' as networks of words"(Diesner & Carley, 2004, p. 2).DNA leverages network analytic methods to identify actors'neighborhoods, focus on actors who aspire to leadership positions, and mark paths between critical actors. Community properties and regular equivalence measures, such as homophily and link prediction, can be applied to estimate the probability of a tie between two actors where no link is seen. More importantly, it allows researchers and investigators to run virtual network experiments, simulate actors' removal, and then observe the model's consequences without

[14]Hidden Markov Model. The premise behind an HMM is that the true underlying process (represented as a series of Markov chain states) is not directly observable (hidden), but it can be probabilistically inferred through another set of stochastic processes (observed transactions, for example).(Pattipati 2006, p.28).

disturbing the existing network under surveillance. The tool allows analysts to evaluate the effects of alternative interventions by investigative organizations in a measurable manner. The efficacy of DNA was demonstrated with an automatic collection of vast amounts of open-source information about Al-Qaida(Carley 2006). The network model that resulted had a "decidedly cellular structure with 5–12 persons per cell"(p. 5). The analysis done along one time period showed that over a decade or so, the network structure of Al Qaida in Iraq decreased in density and interaction levels. Still, it increased in cohesiveness, leading analysts to infer "a movement to a more distributed and efficient organizational form, possibly with larger cells" (p. 4). A key finding of the tool's demonstration was that removing highly central actors in a network would be less effective than taking out important leaders who were beginning to rise.

A prevalent benchmark model for testing strategies and hypotheses related to covert networks is the Counter-Terror Social Network Analysis and intent Recognition (CT-SNAIR) project. According to Weinstein, Campbell, Delaney,& O'Leary (2009, p.1), the project focuses on developing automated techniques and tools to detect and track dynamically-changing terrorist networks and recognize potential capability intent. In addition to obtaining and working with real data for algorithm development and test, the project has a significant focus on modeling and simulating terrorist attacks based on objective information about past episodes. It describes the development and application of a new terror attack description language (TADL), which is used as a basis for modeling and simulation of terrorist attacks. A simulator based on a hidden Markov model (HMM) structure is used to generate transactions for attack scenarios drawn from real events. The model can generate realistic background noise to enable experiments to estimate performance in the presence of a mix of data both wanted and unwanted. The examples of terror-related attacks included specific examples from the September 2004 bombing of the Australian embassy in Jakarta and a fictitious scenario developed in a prior research project in social network analysis. The project employed the DNA tool as a filtering step to divide the actors into distinct communities before determining intent. Given a set of time-ordered transactions between actors, this step helped reduce noise and enhanced the ability to decide activities within a specific group. It generates random networks with structures and

properties similar to real-world social networks for modeling and simulation purposes. Modeling background traffic is an essential step in developing classifiers that can separate harmless activities from suspicious activity. The algorithm used to recognize simulated potential attack scenarios in clutter is based on support vector machine (SVM) techniques[15]. The model is used to demonstrate performance examples, including the probability of detection versus the possibility of false alarm tradeoffs, for a range of system parameters.

Covert networks are constantly mutating and changing, and this dynamicity will shape emerging trends in the social network analysis of covert networks, including those of terrorism and counterterrorism. Predicting covert networks' outcomes and underlying covert structures in ordinary networks is notoriously imprecise, but some broad trends are visible, which can be exploited through mathematical and computational modeling. In the discussion of terrorist groups' behavior, it was seen that under constant pressure from the counterterror agencies, movements of global jihadism like Al Qaida and ISIS have evolved during the past two decades from centralized hierarchies to networked groups, finally to fragmented or discrete cells. It is common policing knowledge that separate units are more difficult to detect and disrupt, especially *lone-wolf attacks* such as November 9, 2009, Fort Hood shooting, and April 15, 2013, Boston Marathon bombing. Knoke(2015 p.7) explains the emerging trends in terrorism this way:

> "Unstable and failed states increasingly offer sanctuaries for terrorists to assemble, train, plane, and launch operations, such as the September 21, 2013, attack by Al-Shabaab gunmen from Somalia on Westgate Mall in Nairobi, Kenya. Insurgencies and guerrilla wars, flaring across Libya, Mali, Yemen, Sudan, the Sinai, Syria, and

[15]"A support vector machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features which are present) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes optimally. Weinstein et al. (2009, p.11) describe the reason for selecting SVM for recognition as being based upon "multiple considerations. First, at a top level, the use of the simulation models for scenario and clutter should be optimal for recognition. But, using these models for recognition will not create a robust detection system. For instance, in real situations, scenarios can be reordered subject to their dependencies. Using the generation model to detect a rearranged scenario in this case will result in a low detector score and probably a miss by the detector. Therefore, separating the detection framework from the simulation framework is critical in our modeling. Other reasons for choosing an SVM are its flexibility in incorporating multiple feature types, good detector performance, and a well-developed tool set."

other parts of theMiddle East and North Africa, offer training grounds for terrorist organizations and their foot soldiers to acquire arms, weapon skills, and combat experience…..Transnational terror will likely plague the planet into the foreseeable future".

2.16 Recommendations

Knoke (2015, p.9) believes that scholars in the interdisciplinary field of terrorism study too often trail behind event-driven trends in transnational terrorism. To get ahead of the curve, researchers must look beyond investigating recent incidents to understanding broader contexts and longer-range perspectives. He outlines some "key issues and opportunities for future network research" which include (extrapolated to include all covert or dark networks which have more or less similar orientations in so far as computational modeling is concerned):

(1) Rigorous comparative analyses of "four historical waves of modern terrorism"[16] for clues about the present and future waves. "Comparing each wave's long-term network dynamics will yield important contrasts and insights into their similar and unique trajectories" (p.9).

(2) "Build more comprehensive, cohesive, and integrated theoretical models capable of explaining the formation, structure, and consequences of terrorist networks. Analytic models of network dynamics must explicate the interpersonal processes by which people are recruited to clandestine organizations, trained in nefarious skills, allocated to organizational positions, and assigned roles in terror operations" (p.10).

(3) Interdisciplinary research and close collaboration between experts in different domains are crucial. "Elements for building social network theories of covert networks should be drawn from diverse social science disciplines,

[16] The four-wave theory of terrorism was proposed by Rapoport, (2001, 2015). These four waves comprise an anarchist wave, beginning in late-19th century following the repressive policies of the Russian tsarist system, followed by anti-colonialist, New Left, and the finally the contemporary jihadist wave. Each wave lasts for approximately a generation according to Rapoport.

encompassing psychological, sociology, geographic, political, economic, and related paradigms. Connecting these elements necessitates close collaborations among substantive experts" (p.9).

(4) Data for computational intervention should be sourced from laboratory experiments rather than collecting inaccessible and dangerous field observation data. "Researchers will construct theoretically based models of interdependent terrorist and counterterror networks comprising both computer programs and human subjects" (p.9).

(5) Develop new methods of measuring network relations among terrorists. Besides improving the accuracy of automated text analysis techniques, other data sources such as photographic, video, audio recordings, biometric authentication, face, and voice-recognition software should be adapted and integrated to generate a new relationship model.

(6) "Create large, high-quality relational datasets to test social network theories of terrorism" (p.9). Sources should include secondary data from public documents, material from the Internet, and cyberspace, including communication networks linking thousands of crime and terrorism oriented websites. A caveat mentioned by Knoke (2015) is that"quality assurance will necessitate such automated routines be supplemented by painstaking hands-on correction of gaps and errors" (p.9).

2.17 Special Features of SNA in Covert Networks

A detailed description of covert social networks, their varied nomenclature, categories, and characteristics was made in the previous chapter. Covert networks are a unique subset of social networks, and such networks are by no means uniform, as we saw earlier. The study of social networks is drawn from a diverse variety of sociological fields such as friendships and conflicts, an example of which is the famous (and well-studied) Zachary's Karate club (Zachary, 1977), the worldwide web, various social media, and instant messaging platforms on the internet like Facebook, Instagram, Twitter, mySpace, WhatsApp, Tumblr, etc., the comity of nations in a world grouping like NATO, SEATO, the United Nations, etc.

Included in this set are the terrorist, criminal, and hybrid terror-criminal networks. Social network analysis and the systematic study of social networks from mathematical and computational perspectives have been game-changers. Newer advances in mathematical modeling, improvements in computational abilities on a logarithmic scale in recent years, and the availability of more modern and more revolutionary algorithmic techniques like artificial intelligence and machine learning have leveraged social network analytic studies massively.

Knoke (1993b) offers a broad insight into the advantages of a structured approach to social network analysis. In his words, "Network concepts and principles offer a perceptive theoretical framework to explain public policy phenomena. By gluing together, several levels of analysis—personal, organizational, systemic—the network approach gives a comprehensive account of political activity and its consequences that surpasses other more piecemeal explanations" (Knoke 1993b, p.164). Knoke (1993a) explains the importance of network analysis in current research on "community power structures and national political elites" which he expresses as increasingly incorporating social network concepts, principles, and methodologies and points to social network analysts using such techniques seek to uncover the various mechanisms underlying the "cleavages and coalitions among state managers, political parties, corporations, interest groups, social movements, mass publics, class segments, and other social formations."The diverse array employed by researchers in the field include "combining reputational, positional, and decision-making measures, researchers delineate the networks of communication ties and resource exchanges, which shape collective actions that attempt to influence the outcomes of political controversies"(Knoke 1993b).

As a particular subset of the studies on social networks, covert network studies derive many of their usual social network analytics methodologies. But, ordinary analytics are not trained to cope with covert networks' most important aspects: confinement of information, hiding of intent, and active deception. We may add encryption and covert channel communications to this list. Given this backdrop, computational studies have come up with a variety of methodologies to deal with it. All of them have many similarities with social network

analytics and also crucial differences. This chapter begins with the usual tools and techniques used to evaluate social networks and segues to the variations introduced to adapt to covert network studies. Many variants have much to do with coverts networks' main characteristics, i.e., size, incompleteness, dynamicity, and fuzzy boundaries. Other derivatives of these main properties, like core-periphery structures, pre-existing ties, sparsity, etc. are also touched. There is a section on the methodologies explicitly used for detecting the employees who had a role to play in the ENRON insider trading scandal. These studies on ENRON fall within the overall scope of covert network analysis.

2.18 Sociological versus Computational Perspectives

Before delving into discussions on social network analysis tools and covert network analysis, it's worth reviewing some of the differences between sociological and computational aspects of social network analysis. Sociological studies of networks have long preceded the computational approaches to the subject. A significant constraint that hindered these studies was the lack of computational resources to account for large networks with hundreds of nodes. Consequently, early computational interventions centered around small-sized networks (Zachary's Karate club network had only 34 nodes), and a few ventured into ones that were only slightly larger. With increasing computational power and complexity, however, the ability to deal with massive real-world networks with millions of nodes also evolved. Social network analytics have acquired remarkable sophistication and specificity (to the network being studied).

Despite the recent explosion of interest in computer science, the subtlety and nuance associated with a sociological outlook of networks have yet to find an exact echo in the computational area. There is an undeniable dichotomy in social network studies when approached from computational and sociological perspectives. Both approaches typically converse past each other. The divergence is especially stark when individual node related differences are considered. Social network researchers rarely account for the possible effect of individual differences on network structure (Mehra et al. 2001). Researchers looking at

unique characteristics are often wedded to conceptualizations of individuals as independent entities rather than the relational approach of network analysis. For studies in covert networks, this divergence plays out as a contrast between research studies that focus on criminal behavior and studies that focus on criminal networks' motivations. Covert network analysis is distinctly different from the studies in the domain of conventional network analysis in that the focus is more on the concealment of activities by the actors and even the actual identities of the actors themselves, which may not be apparent immediately; these aspects need to be teased out during the analysis, which adds an extra layer of scrutiny, and this part of the analysis is a fraught one in most circumstances.

Surprisingly, even within a purely sociological approach, social psychological and social network approaches manifest contrasting methodologies (Robins and Kashima 2008). For example, studies in social cognition relate to actions by the actors and individual perceptions, without an overarching idea of how such individual-level effects merge to form an entire social network system, which leads to a disproportionate focus on network topology to the neglect of node-based individual motivations. Such gaps have occupied the interest of social network theorists. (Emirbayer and Goodwin 1994). While recent research has bridged the deficit in many ways, there are indications that a fuller integration of individual node-based action within social structures will require considering both network and individual variables and their possible interaction. In examining several of these factors together, Copeland et al. (2008) concluded that an exclusive group psychological or network structural approach might not adequately explain organizational behavior. They argued for a more unified theoretical approach linking, in particular, identity and network perspectives, which is also what law enforcement agencies consider in investigations, and these agencies have intuitively accorded importance to co-offending[17] entities or communities within networks.

[17]Co-offending networks are groups of offenders who plan covert acts together. More on this topic in chapter 6.

2.19 Social Geometry in Social Networks

The concept that all social networks have an underlying geometry that lends itself to systematic appraisal has been around since Moreno, the sociology journal founder in 1937[18]. Donald Black has enunciated the notion that social networks can be exploited for structure-based (and hence, mathematical and computational) analysis through his twin concepts of social space and social geometry (1976, 1998 & 2004) and its recent extrapolation to terrorist networks (2004).

Black (2004, pp.15-16) States thus, "Pure sociology[19] explains human behavior with its social geometry – its multi-dimensional location and direction in social space…Social space has various dimensions – horizontal (such as degrees of intimacy and integration): vertical (inequality): corporate (involvement of groups): cultural (language and religion): normative (social control)". Black's model allows for the scrutiny of each variable by including multiple dimensions while holding others constant. That is, the theoretical propositions have under a condition of *ceteris paribus*[20], a probabilistic approach characteristic of mathematics and computer science. Further, the inclusion of these dimensions (or vectors as they may be called in computer science parlance) within the same model allows for the possibility of both interaction effects of entities within the dimensions and correlations between them, with any one of them being used to explain any other. Such models lend themselves uses each of the measurements to explain variation in normative behavior, but relational or cultural behavior might also be jointly accountable by the other dimensions. A multi-dimensional sociological approach that breaks down a nuanced real-world structure into dimensions and variables allows easy mathematical interventions. Not surprisingly, increasing sophistication in networks' sociological modeling has rubbed off on mathematical and computational methodologies in recent times.

---

[18]There are examples of structural approaches to networks in sociology even prior to Moreno. For a more detailed history please refer to Freeman's "Some Antecedents of Social Network Analysis" (1996).

[19] Pure sociology projects human behavior as social life—a concept that does not exist in the mind, is not explainable by the aims of actions. It is distinguished from other sociological models by what is absent from it: psychology, teleology, and even people as such.

[20]*Ceterisparibus*is a Latin phrase meaning all other things being equalwhich is used for manipulating a single variable holding all other variables constant in a multi-variate system.

Another significant contribution of Black (2004) is his work on terrorist and criminal networks to which he extrapolates his constructs of social space and social geometry mentioned above. Violence, which may be considered as a logical outcome of a terrorist network, is described by him as an "unpredictable outburst or unexplainable explosion, but it arises with geometric precision…violence occurs when the social geometry of a conflict – the conflict structure is violent…..It is unpredictable and unexplainable only if we seek its origins in individuals' characteristics (such as their beliefs or frustrations) or the characteristics of societies, communities, or other collectivities (such as their cultural values or inequality). But violent individuals and violent collectivities do not exist: No individual or collectivity is violent in all settings. Neither individualistic nor collectivistic theories predict and explain precisely when and how violence occurs (see Black 1995: 852-58; 2002d:1-3). **Violence occurs when the social geometry of a conflict – the conflict structure – is violent**. Every form of violence has its structure, whether a beating structure, dueling structure, lynching structure, feuding structure, genocide structure--or terrorism"(p.3). In other words, individuals and their behavior patterns don't count in so far as network-based outcomes are concerned, and this is indeed a most interesting way of looking at network outputs since we have a theory that inextricably relates results (including illegal ones) to the structures that have produced them. The actors manning different perches within these structures do not count; preferably, when they are participants within the structure, the results will invariably be the ones the structure is built for.

Adverse or illegal outcomes like terrorism or lynching are described as relational distance, inequality, and functional independence. All these entities are social distances[21] (Senechal de La Roche 1996: p120), which we may interpret in graph-theoretic terms as edges or ties or multiples thereof, like geodesic distances between nodes or communities (shortest paths). Black (2004) builds upon this argument and articulates a mathematical (more or less) relationship between acts of terrorism and various social distance types:

[21]A social distance is defined by Black as a separation between social locations, such as wealth (economic distance); authority (hierarchical distance); integration (radial distance); culture (cultural distance); intimacy (relational distance); organization (organizational distance); and activities (functional distance). (Black 1976, 2000:348, n.13)

> "Terrorism crosses other social distances as well-other vertical distances (such as radial distance, a difference in social integration); organizational distance (a difference in the capacity for corporate action); and another kind of functional distance (a difference in social activity, such as modes of livelihood). In other words, pure terrorism strikes across very long distances and along diverse dimensions of social space-cultural, relational, economic, hierarchical, functional, etc. Accordingly: *Pure terrorism arises intercollectively and upwardly across long distances in multidimensional social space*. So travel the bullets, bombs, and other weapons of terrorists. And the greater the social distances, the greater their destructivity"(p.19). Sociological constructs of this nature paved the way towards constructing a mathematical model of covert networks that includes terrorist, criminal, clandestine networks of various types.

Black(2004) argues that each form of social control has a corresponding social geometry. Suicide, gossip, avoidance, and all other forms of social control, including "terrorism" labeling, is isomorphic with its geometric configuration. The theoretical goal is to identify each behavior's geometry. By doing so, we are stating the social conditions under which that behavior occurs. Afterward, the geometry can be subject to empirical testing to assess its validity. A recent application of social geometry concerns "terrorism labeling" of certain collective and violent actions (Boches 2020). Boches extends Black's theory of behavior geometry by finding the behavior's geographical location, direction, and distance along five dimensions of social space. The *vertical* dimension comprises wealth and its distribution; the horizontal dimension comprises intimacy, integration, and interdependence; the symbolic dimension comprises culture; the corporate dimension comprises organization, and the normative dimension comprises social control. Boches illustrates the concept of location and distance by citing an example of offending among modern street gangs:

> "This feuding occurs laterally across moderate distances in the horizontal dimension, at moderate elevations in the corporate dimension, and at a vertical location significantly below third-party settlement agents. If this geometry begins to change, the associated violence will also start to look different. For instance, holding all else constant, lowering the location of gang violence along the corporate dimension – in

other words, changing the disputants from moderately organized groups to individuals – decreases the probability that the violence will be reciprocal" (p.151).

2.20 Importance of Structure in Evaluating Covert Networks

Black's enunciation of the structural underpinnings of network behavior initially made for his theory of law(1976), and later for terrorism (2004), was a benchmark for sociologists, who started employing the twin structural concepts of social geometry and social distance to explain different forms of violent (and covert) activities, including terrorism, lynching, rioting, vigilantism (Senechal de La Roche, 1996), genocide (Campbell, 2015), domestic violence(Tucker, 1999), while others analyzed the social control of specific behaviors such as suicide (Campbell and Manning, 2019; Tucker, 2015) and homicide (Cooney, 2009). and therapy(Tucker, 1999).

My dissertation mirrors this approach to some extent because of the emphasis on network structures, particularly on actors or nodes. It has been a moot presumption that once an actor forms a relationship with another actor, the relationship's nature dictates the tie's outcome. If the relationship or tie has come about in a situation where both actors forming the tie are circumstanced to produce a particular type of output, say, opaque, and covert, the tie structure becomes defining for the future transaction outputs. So long as the tie exists, the nature of the outcomes will not vary by any great degree. It needs to be mentioned here that this dissertation does, to some extent, acknowledges a defining role to the actors. The dataset under study, i.e., the ENRON email set, has a few employees planning insider trading before its collapse. It's entirely possible that if other employees had been transferred through some design to the posts held by those who were part of the conspiracy, the outcomes might have been different.

The theory that covert structures, rather than individual actors, are responsible for covert outcomes finds an echo in the paper on conspiracy in the heavy electrical industry in the

United States by Baker and Faulkner (1994). The authors elegantly expressed the reasons why such conspiracies became pervasive in this industry:

> "Collusive agreements in the heavy electrical equipment industry go back to the 1880s, but the price-fixing "schemes of the 1950s were given special impetus when repeated episodes of price warfare proved incompatible with top management demands for higher profits"(Scherer 1980, p. 170). Top executives imposed unrealistic profit objectives in an industry characterized by chronic overcapacity, increasing foreign competition, and stagnating demand (Ohio Valley 1965, p. 939). To cope, managers decided to conspire rather than compete. Their elaborate conspiracy involved as many as 40 manufacturers and included more than 20 product lines, with total annual sales over $2 billion. The conspiracy was pervasive and long-lasting; it became, insiders said, a"**way of life**" (U.S. Senate Committee on the judiciary 1961, pp. 16879-84 [henceforwardKefauver Committee])".

I highlighted the phrase "way of life" in the specific context of structures within networks that have evolved in the face of particular circumstances and practices to possess a life of their own. It is trivial to conclude from the observations of Baker and Faulkner (1994) that any player that entered the heavy electrical industry in the specific role of a vendor was prone to this sort of covert activity propelled by its participation in the network structure which had already evolved through decades of reactions to the existing practices. This detailed analysis of the concept of social spaces and geometry serves to underpin the importance of placing our focus on the structures, and especially, the ties between actors rather than on the actors themselves when studying covert networks, my approach, therefore, runs counts to most computational analyses in social network research.

2.21 Causal Modelling of Network Structures

The models presented by Black (2004) and Baker and Faulkner (1994) above point to the evolution of structures within networks around an environment that prevails for some length of time. Such structures have a personality of their own and influence the course of the

outcomes substantially to the extent of subsuming the roles of actors populating these structures' vertices, which is true for all networks. This is even more true for structures that generate violence (and by extension covertness, as there can be no pre-planning of violence without accompanying secrecy). Thus, there is pressing need to identify structures that produce covert results, whether they are networks as a whole or sub-structures within the network. And as such structures arise in response to external variables, there is also a need to carefully observe which factors can generate such structures and if these factors are prevailing in the context of the network's functioning. A multi-modeling approach can be envisaged in these circumstances, where each model will be the product of slight variations in the causal factors. These models may then be compared to the structures being studied to ascertain how closely they fit. The closest fit may then be considered for predicting the outcome of the structure in question. There are many popular mathematical methods for comparing models and their outputs; models in this regard may be both deterministic or stochastic. (For an excellent introduction to the topic of mathematical modeling, please refer to https://people.maths.bris.ac.uk/~madjl/course_text.pdf.)

Thus, the focus shifts from the shape of the structures understudy to the input variables that molded their morphologies in the first place; in other words, a causal analytic approach. There are many instances of causal analyses in the sociological literature. One of the more popular ones proposed by Ross (1993) about terror networks; this study is a generic one and can safely extrapolate to fit other types of clandestine networks. Typical structures like core-periphery, poly-centricity, centralization, sparseness, homophily, microstructures, etc. in covert networks have been discussed in the previous chapter. There is variance even among these specialized structures exclusive to covert networks, depending on the nature of their covert behavior, i.e., terrorist or criminal or counterintelligence. Identifying the external factors that lead to the development of these typical structures' requires close study. Such studies may be termed as the causal methodology, which focuses on the prevalent environmental variables that result in structures that have covert outputs.

Ross (1993) studied the causes of terrorism and stressed the structural factors that cause this phenomenon to arise. He describes his reasons thus – "Specifically, there is some dispute

over which method is the best way to understand the causes of terrorism, the quality of the analyzes, and the hidden agendas of causation studies. Moreover, none of the causes identified are mutually exclusive; all approaches to studying the causes of terrorism borrow concepts from each other. For example, analyzes of specific causes often derive their processes from case studies. Case studies use concepts found in studies of specific causes and attempt to develop comprehensive theories to develop their postulates from case studies and analyze individual causes. Thus, over the years, several causes for the occurrence of terrorism have been presented. This literature's most prominent causes fall into three categories: structural, psychological, and rational choice. In general, structural theories posit that terrorism can be found in the environment and a society's political, cultural, social, and economic structure. Psychological theories try to explain why individuals join terrorist organizations, terrorist group dynamics; and, how participants (i.e., terrorists, victims, and audiences) affect the commission of terrorist acts. Finally, rational choice theories attempt to explain participation in terrorist organizations and the choice of terrorist actions due to the participants' cost-benefit calculations. Of the three, the structural causes are the easiest to test but have rarely been integrated into a comprehensive causal model that would serve as the foundation for testing" (Ross 1993, pp.317-318).

Ross defines his aim to construct what he terms as a tentative model of the structural causes of terrorism to help researchers develop and test the relative importance of previously identified causes and their interactions to determine the scope, intensity, and amount of terrorism. He further defines his motivations to create a test-model that would act as a benchmark for other researchers. This model will allow researchers just to observe if the outcomes are along predicted lines. He states, "A causal model using the structural variables of terrorism would specify the dominant processes by which this form of political behavior occurs. This model does not preclude the possibility that psychological and rational choice theories could not adequately explain the final decision that individual terrorists or groups make to engage in terrorism. It is only a practical research strategy". (Ross, 1993, p.318). This model is a development on the one suggested by Hopple (1982), who indicated that a causal model of terrorism should be created and prescribed which type of variables may be included and theorized that a feasible first iteration of the model of terrorism would have

two independent variables, internal (intra-societal) and external (interstate and systemic); and one dependent variable, transnational terrorism. Hopple argued for restricting the model to one of the categories of terrorism then recognized by researchers.

Ross (1993) opines that the structural causes are ideal starting points for a model of the kind recommended by Hopple because structural variables are easier to actuate and measure than psychological or rational choice variables, which are abstract and do not lend themselves easily to mathematical articulation. Further, Ross (1993) thinks that the accuracy of such a causal model's predictive capability is directly proportional to the specificity of the variables and their interactions that cause terrorism. He also hypothesizes that the higher the number and intensity of structural causes of terrorism (the independent variables), the higher the number of terrorist acts perpetrated by any particular terrorist or terrorist organization (the dependent variable). In his words, "If these variables are causally related, then the systematic elimination or lessening of them should lead to a decrease in terrorism. Knowledge of this kind would be useful to actors involved in counterterrorism measures" (Ross 1993, p.318).

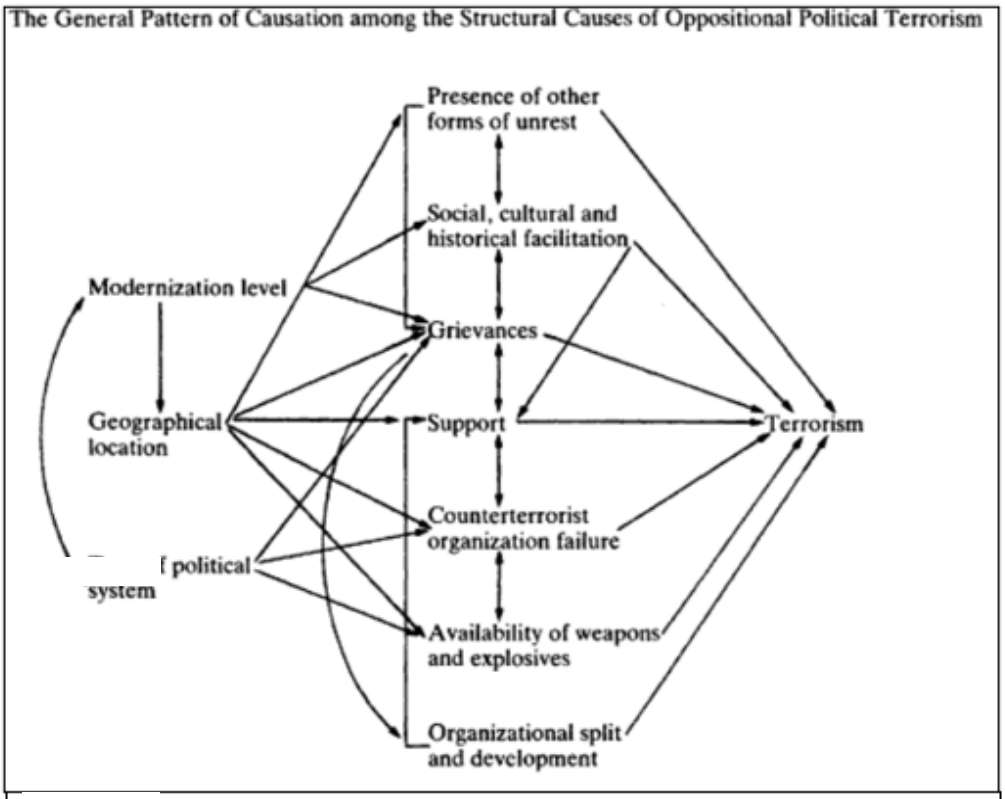


Figure 2.4 Ross's (1993) illustration of Structural causes of Terrorism in opposition to political environment. the structural causes are three layered – the three permissive causes Geographical Location, Type of Political System, and Level of Modernization and are shown at the extreme left of the illustration. The seven precipitants namely, Social, Cultural, and Historical Facilitation, Organizational Split and Development, Presence of Other Forms of Unrest, Support, Counterterrorist Organization Failure, Availability of Weapons and Explosives, and Grievances. are shown in the middle layer. At the extreme right is the outcome, namely terrorism.

In simple terms, the study of causality leads to a better understanding of covert structures, leading to better outcomes prediction. Ross' causal model of terrorism is represented in Figure 2.4 (Ross 1993, p.321). In his model, Geographical Location, Type of Political System, and Level of Modernization are the *permissive* reasons for terrorism. The remaining cause, i.e., Social, Cultural, and Historical Facilitation, Organizational Split and Development, Presence of Other Forms of Unrest, Support, Counterterrorist Organization Failure, Availability of Weapons and Explosives, and Grievances are the *precipitant* causes. This modeling exercise can be generalized for any network structure, as Figure 2.5 shows:

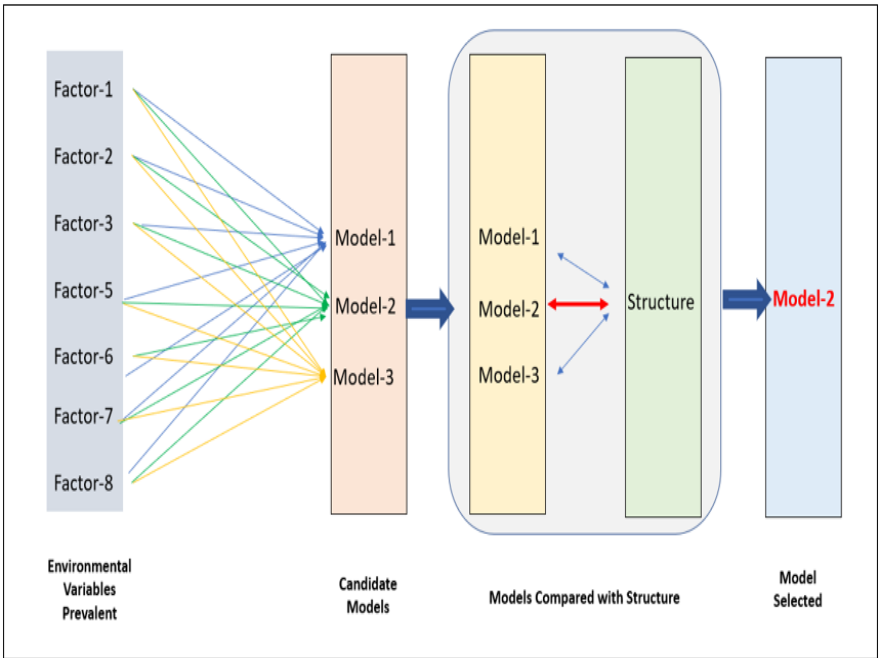


Figure 2.5 A Flow Diagram showing the construction of a causal model of terrorism. There are several factors and environmental variables that may lead to the development of terrorism as an '*oppositional*' structure. To build viable models, the analyst needs to work out various permutations and combinations of the input factors and the results will be competing models (referred to as candidate models in the figure and shown the second vertical layer). Each of the candidate models can then be '*fitted*' with the actual structure of the terrorist network available and the analyst may decide which model fits the best (third vertical layer in the diagram). This model can then be chosen (final vertical layer) for experimentation. The principles enunciated here for building models for terrorist networks may also be extrapolated for other covert networks or covert community structures within conventional networks.

2.22 Domains, Action Sets, and Opposition Networks

Many of the social networks' outcomes are related to groups or communities within the networks themselves, which is true of both bright and dark networks without exception. As discussed extensively earlier, the structures within networks arise from environmental variables, both external and internal. Once formed, these structures tend to be more or less static as far as their outputs are concerned. The presence of groups or communities within networks is similar to *policy domains* and *action sets* in groups of organizations in a particular field (political, social, business, etc.). A policy domain is a component of the political system organized around substantive issues (Burstein 1991, p.327). Burstein (1991) defines policy domains in three sets of characteristics-*substantive* or *functional*,

organizational, and *cultural*. Substantive issues that define an environment are seen as sharing inherent substantive factors which arguably have a certain logic and coherence and which influence how they are framed and dealt with; for instance, fields such as energy, health, transportation, or agriculture, for example, and most specific issues qualify in this definition of a domain. Organizational characteristics define policy domains assets of organizations concerned about substantive problems, which take each other into account as they formulate policy options and work for their adoption. This sociological analysis places less emphasis on the qualities innate to policy domains. Instead, domains are seen as primarily constructed socially by those active in politics (Laumann & Knoke 1987). A third way of characterizing domains is through *cultural* constructs around which organizations and individuals orient their actions. The *cultural* basis of policy domains determines which policy options the domains adopt and which other organizations to deal with. Cultural constructs are strongly influenced by cultural theories about how society works. Domains are significant to the analysis of social networks from the perspective of communities and cliques that are inevitable parts of networks and are also instrumental in the networks' evolution, predicting their outcomes, etc. The use of domain analysis in covert networks is especially crucial given their tendency of sub-centrality and poly-centricity to attempt to disrupt their functions by law enforcement and other legal agencies. Domains are also crucial from an inter-organizational point of view in as much as they determine how networks interact amongst themselves.

In their study of conspiracies, Baker and Faulkner (1994, p.8) described the interaction of criminal inter-organizational networks in price-fixing cartels in the US heavy electrical industry as *organizational action sets*. The concept of action sets was introduced in Chapter 1. This concept deserves a revisit, given its importance in comprehending how covert communities are generated and how they evolve during the network's evolution. Action sets may be defined in the context of the policy domain, the basic unit in organizational state analysis, defined as any part of the system whose constituents are identified "by specifying a substantively defined criterion of mutual relevance or common orientation ... concerned with formulating, advocating, and selecting courses of action" (Knoke and Laumann 1982, p. 256).

A policy domain's members comprise organizations and events whose interests and actions must be taken into account by other members. According to Knoke and Pappi (1991):

> "Every domain encompasses a diversity of controversial policy matters, interest groups, and public authorities, each seeking to influence the ultimate decisions about matters of importance to them and their constituencies…each policy domain also develops a logically coherent substantive or functional basis for framing its policies and that its participants usually construct a common culture about how society works or should work" (p.510).

Domains are seldom permanent, and their participants may choose different domains depending on the issues involved and how important these issues are the participants. Domains are thus decided by the problems and less by participation. Knoke and Pappi (1991) go on to say:

> "the fluid nature of the national policy domain fights is captured by four nested analytic concepts: the event public, the collective actor, the action set, and the opposition network. These constructs rely heavily on network principles in which the basic elements are actors linked by exchange ties of specific forms and content. An *event public* consists of all domain organizations that express interest in a particular policy event, regardless of which outcomes they prefer. (Some members may have no preferred outcome.). A *collective actor*[22] consists of three or more formal organizations within an event public that communicate about policy matters and that favor the same outcome. An *action set*, which is even more restricted, consists of those collective actor organizations that consciously coordinate activities on a particular event. Group cohesion is the essential feature of an action set: All members prefer the same outcome for the event, are directly or indirectly linked in a communication network, and collaborate in lobbying and other activities to influence policy…an *opposition network* is the pattern of overlapping memberships among the

[22] Collective Actor is a concept introduced by Laumann and Marsden (1979) who proposed it as an alternative to the existing political theories and to provide a theoretical rationale for characterizing oppositional structures in political elites. The concept of a collective actor is treated in their paper as an elementary analytic unit in the study of conflict structures in elite systems.

collective actors or action sets that form around a set of policy events within a specific domain during some period. The structure of a domain's opposition network is a function of the degree to which its collective actors' or action sets' members coincide".(p.510)

To recap, domains, actions sets, and identify community groups in covert networks or covert communities within networks that are open but susceptible to the budding of conspiracy sub-networks within their architecture (The ENRON dataset under study is an excellent example of this). Of particular interest are the concept of action sets described as assemblages of organizations brought together to carry out specific activities and consist of the same participant actors across events. Such sets tend to be short-lived or long-lived, depending on how durable their common aims are in different domains and events (Aldrich 1979; Knoke and Pappi, 1991; Knoke and Burleigh 1989). That many of these action sets tend to dissipate after success or failure is a construct which has a lot of similarity to *collusive*(please see *ibid* p.234 for a detailed discussion on collusion) communities with *common intentions*(please see *ibid* p.248 for a specific definition of common intention) and objectives in a covert network and contributes significantly to the study of *community evolution*(Spiliopoulou and Aggarwal 2011) as opposed to the evolution of the networks as a whole.

2.23 Structure and Topology of Social Networks

The most basic and atomic unit that can be identified in a network is the node or actor. In the definition provided by Knoke and Yang (2008):

"Actors may be individual natural persons or collectives, such as informal groups and formal organizations. Common examples of individual actors include children on a playground, high school students, employees in a corporate work team, a nursing home staff, and terrorists operating in a covert cell. Collective actors might be firms competing in an industry, voluntary associations raising funds for charities, political parties holding seats in a parliament, or nations signing a military alliance. Sometimes network actors

encompass mixed types, such as an organizational field comprising the suppliers, producers, customers, and government regulators of health care. (pp.6-7)" The counterpart of an actor in graph theory is a node or vertex.

A network is defined by the relation/link between the nodes in it, as given in the examples above. There can be particular relations between a single set of nodes in a network. For instance, in a product network, the connection could be based on "similarity" or "brought together" in a product set. Similarly, there can be unique/distinct relations between multiple groups of nodes, such as user-product networks. These types of networks are heterogeneous networks. When the network comprises two sets of nodes, it is called a two-mode network. Some examples of two-mode networks include user-product networks (Amazon, eBay, etc.), membership or affiliation networks (actor-movies (IMDB), user-group (YouTube), user-channel (YouTube), user-project (GitHub), user-organization, etc.), user-preference networks (Pinterest, Instagram, Twitter), citation networks, user-stock investment. These two-mode networks can be transformed into single-mode networks between a single set of nodes as in the examples given above and then analyzed. However, two-mode networks can also be analyzed using the methods of Borgatti and Everett (1997) and Latapy et al. (2008).

Apart from social networks, numerous data networks are also formed between objects other than social entities, like sensors, products, words/texts, brain neurons, proteins, genes, geographical locations, predators and preys, webpages, etc. Though the social network analysis measures were primarily designed to analyze social networks, they can also be employed to analyze data networks like these. Common tasks of SNA involve the identification of the most influential, prestigious, or central actors, using statistical measures; detection of hubs and authorities, using link analysis algorithms; discovery of communities, using community detection techniques, and understanding of how information propagates in the network, using diffusion algorithms. These tasks are instrumental in extracting knowledge from networks and, consequently, in the process of problem-solving. Due to such tasks' appealing nature and the high potential opened by such types of analyses, social network analytics has become a popular approach in a wide range of fields, from biology to business. For instance, some companies use social network

analysis to maximize their products' positive word-of-mouth by targeting the customers with higher network value (those with greater influence and support) (Domingos and Richardson, 2001; Richardson and Domingos, 2002; 3 Leskovec et al., 2007). According to these profiles, other companies, such as those operating in the mobile telecommunications sector, apply social network analytic techniques to their call center networks and use them to identify customers" profiles and recommend personalized mobile phone tariffs. These companies also use network analytics to churn out predictions, e.g., to detect customers who may switch to another mobile operator by detecting changes in phone contacts (Dasgupta et al., 2008; Wei and Chiu, 2002).

Another interesting application emerges from the domain of fraud detection. For instance, social network analysis can be applied to organizational communications (e.g., Enron company dataset) to analyze the frequency and direction of formal/informal email communication, revealing communication patterns among employees and managers. These patterns can help identify people engaged in fraudulent activities, thus promoting more efficient forms of acting towards eradicating crime (Xu and Chen, 2005; Shetty and Adibi, 2004).

Although the origins of network studies go back several decades, recent years have witnessed impressive advances in network-related fields, mainly due to the growing interest in social networks (Wasserman and Faust, 1994; Abraham et al., 2009; Charu Aggarwal, 2011; Furht, 2010; Zafarani et al., 2014; Barabási, 2016; Aggarwal, 2009). Social networks have thus become a hot topic and a focus of considerable academic attention. Increasing the use of mathematical devices and algorithmic interventions has widenedSNA techniques' potential for studying several categories of problems. My research aims to provide a general and concise overview of the essentials of SNA to lay down the basis of the approach developed in the course of this study.

To conclude about the nature of the Network, there are several ways. Social network analysis can be described as a tiered structure, starting with node or vertex level properties and progressing to characteristics that define the entire network. Robins (2009) described

social networks' analysis as a multi-tier mechanism, reflected in Figure 2.6below. Although many levels are shown in a hierarchical (pyramidal) manner, each level's properties are not necessarily limited to that level and may be applicable across the tiers. Each of the properties can be dynamically linked to any other, cutting across the levels depicted in Figure 2.6.

The results of these interactions can be complex and may lead to potentially new constructs that can then be plowed back into the model or simulation. Different combinations of variables and parameters can also be made by allotting differing weights to each factor, according to need. This simulation's overarching impression is that it is flexible and dynamic and may be applicable across domains and from longitudinal and non-longitudinal (time-based or otherwise) perspectives. Within such a multilevel and multi-dynamic network ecosystem, the network perspective comes into its own. At this point, researchers, analysts, simulators, and investigators need to bear in mind the possibilities of designing their research. Covert networks tend to be highly dynamic and reactive to external stimuli, and this approach may be considered ideal in modeling these networks.

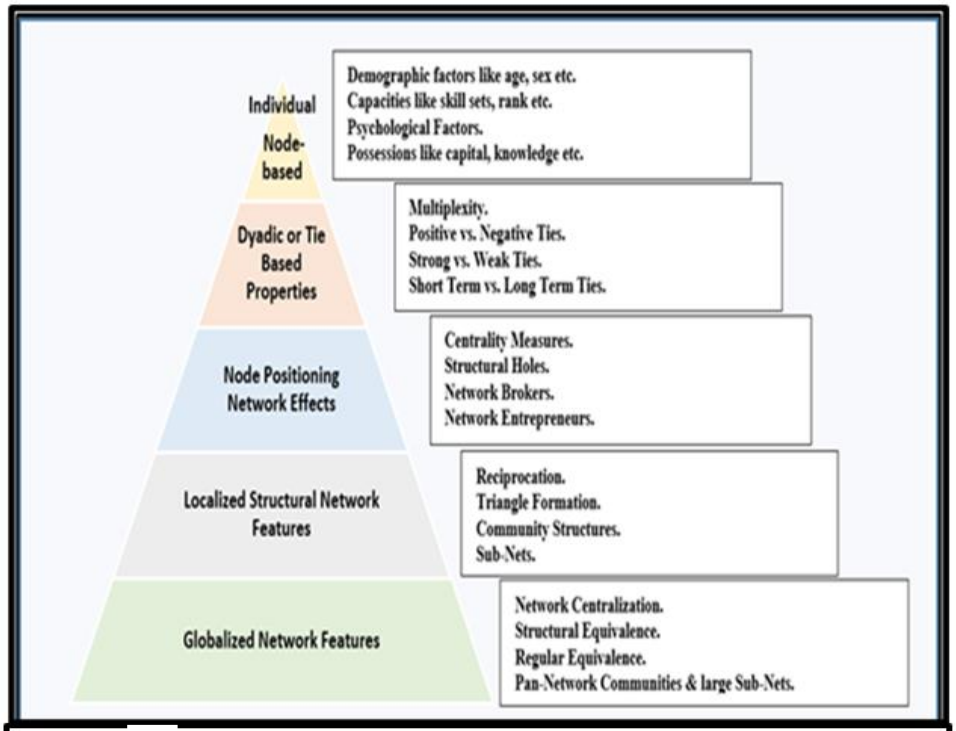


Figure 2.6 The Pyramidal structure above describes five different levels of features and metrics, both individual (node or edge based) and relational, relevant to research into network-based social systems as envisaged by Robins (2009). It needs to be clarified that the levels are not hierarchical, rather, all or any of them might exist, and not independently of each other, i.e. some of these features may combine with each other.

Based on the figure above, of the five levels, the first relates explicitly to individual factors that are not conceptually related to network constructs. In contrast, the remaining four are different relational effects levels, from the dyadic to the global. Robins (2009) considers how some of these individual and relational factors may interact with each other. The interplay of these constructs approximates a sociological model translated into a more mathematical one, i.e., the abstract dynamics of a social network decompose into measurable parts.

2.24 Mathematical Representation of Social Networks

In computational terms, a social network consists of a finite set of vertices and the relations, or ties, defined on them (Wasserman and Faust, 1994). The established relationships can

be personal or professional, ranging from casual acquaintance to close familial bonds. Besides social relations, links can also represent the flow of information/goods/money, interactions, and similarities. Graphs usually represent the structure of such networks. Therefore, networks are often regarded as equivalent to graphs. A graph is composed of two fundamental units: vertices and edges. A pair of vertices define every edge, also called its endpoints. According to the application field, vertices can represent various individual entities (e.g., people, organizations, countries, papers, products, plants, and animals). In turn, an edge is a line that connects two vertices. It can, analogously, represent numerous kinds of relationships between individual entities (e.g., communication, cooperation, friendship, kinship, acquaintances, and trade). Edges may be directed or undirected, depending on if the nature of the relation is asymmetric or symmetric.

Formally, a graph G consists of a non-empty set V of vertices and a set E of edges, being defined as G = (V, E). According to Diestel (1990), the order of a graph G is given by the total number of vertices n or, mathematically, $|V| = n$. Analogously, the graph G''s size is the total number of edges $|E| = m$. The maximum number of edges in an undirected graph is $n(n-1)$, and for undirected graphs, the figure is $n(n-1)/2$.

In the existing literature, two main types of data structures are used to represent graphs: *list structures* and *matrix structures*. These structures are appropriate for storing graphs in computers to analyze them further using automatic tools. List structures, such as incidence lists and adjacency lists, are suitable for storing sparse graphs since they reduce the required storage space. Matrix structures are used to represent full matrices. They include incidence matrices, adjacency matrices (also termed *sociomatrices* or *sociograms*), Laplacian matrices (which contain both adjacency and degree information), and distance matrices (i.e., adjacency matrices whose entries are the lengths of the shortest paths between pairs of nodes).

Several types of graphs can be used to model different kinds of social networks. For instance, graphs can be classified according to the direction of their links, which leads us to the differentiation between undirected and directed graphs. Undirected graphs (or

undirected networks) are graphs whose edges connect unordered pairs of vertices. In other words, each edge of the graph connects concomitantly two vertices.

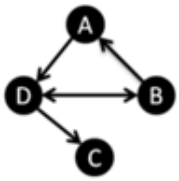
A stricter type of graph is the so-called directed graph (or directed network). Directed graphs, sometimes referred to as *di-graphs*, can be straightforwardly defined as graphs whose edges (also called *arcs or ties*)have an orientation assigned, so the order of the vertices they link matters. Formally, in a directed graph, if E_{12} is an arc and v_1 and v_2 are vertices such that $e_{12} = (v_1, v_2)$, then e_{12} is said to join v_1 to v_2. As the first vertex,$v_1$is called the initial vertex or *tail*, and the second vertex $v_2$is called the terminal vertex or simply *head*. Graphically, directed edges are depicted by arrows, indicating the direction of the linkage. Directed graphs can be either *cyclic*, i.e., graphs containing closed loops of edges or *ring* structures, or *acyclic* (e.g., trees).

The second significant type is *undirected* graphs. A typical example is Facebook, since, in this social network, the established friendship tie is mutual or reciprocal (e.g., if someone accepts a friend request from a given person), it is implicitly assumed that s/he and the other person are friends). By contrast, Twitter is an example of a directed graph since others can follow a person without necessarily following them. In this case, the tie between a pair of individuals is directed, with the tail being the follower and the head being followed, meaning that a one-way relationship is established.

Edges joining nodes can also be assigned weights or values depending on what such weights may represent (e.g., number of emails exchanged in a mailing network). Such networks are called *weighted* networks or graphs. Unless explicitly stated, graphs are considered unweighted, i.e., all existing edges are assigned a value of 1. Unweighted graphs are binary since edges are either present or absent (edges have 0). On the other hand, weighted graphs are richer graphs since each edge has associated a weight providing the user with more information about the social network, such as the strength of the connection of the pair of vertices it joins.

According to Granovetter (1973, 1995), in social networks, *the weight of a tie is generally a function of duration, emotional intensity, frequency of interaction, intimacy, and services exchange*. Therefore, strong ties usually represent close friends, and weak ties represent acquaintances. In other kinds of networks, the meaning of a tie's weight can vary, depending on the context; for instance, a tie can define the number of seats among airports, the number of exchanged products, etc. For undirected and unweighted graphs, adjacency matrices are necessarily binary (as a consequence of being unweighted) and symmetric (as a consequence of being undirected, meaning that $e_{ij} = e_{ji}$), with $e_{ij} = 1$ representing the presence of an edge between vertices i and j, and $e_{ij} = 0$ representing the absence of an edge between vertex pair (i, j). For directed and weighted graphs, such matrices' entries take values from interval $[0, \max(w)]$ and are non-symmetric. In both cases, the matrix has non-negative values. Figure 2.7 below illustrates how a graph can be represented by both an edge list and an adjacency matrix.

Vertex	A	B	C	D
A	-	0	0	1
B	1	-	0	1
C	0	0	-	0
D	0	1	1	-



Vertex	Vertex
A	D
B	A
B	D
D	B
D	C

Figure 2.7 A directed and unweighted graph G represented by means of an adjacency matrix (left-side of the Figure) and an adjacency list (right-side of the Figure).

2.25 Statistical Measures in a Social Network

As has been described, social networks are graph-like structures with nodes (or vertices) and an edge connecting them. Nodes and edges are the basic building blocks of the entire network. At the secondary level, we have two nodes and the edge connecting them, called

dyads. At a slightly higher level, we have three nodes forming a triangular relationship called a triad. Higher-level structures are also discernible in networks, such as quartets or rectangles, cliques (completely interconnected nodes), trees, communities, etc. Many measures have been developed for evaluating the relationships within higher structures and measuring their homogeneity and uniformity. This section introduces some of the statistical measures that my work uses to assess network structures. The measures can be divided according to the level of analysis: node-level or network-level. At the node level, centrality measures provide information about a node or vertex's position within the network's overall structure and identify its key players. Network-level measures provide more compact information and assess the network's overall structure, giving insights into the essential properties underlying the social phenomena.

2.25.1 Node-level Statistical Measures

Studying how individuals interact in the network context helps understand the overall behavior of the social systems that generated those networks, which is usually the final goal of such analysis. Centrality is a general measure of how important an actor or node's position is within the social network's overall structure. *Prestige*, a closely synonymous term used in the sociological literature, adds dimension to this notion, referring to the "extent to which a social actor in a network "receives" or "serves as the object" of relations sent by others in the network. The sender-receiver or source-target distinction strongly emphasizes inequalities in control over resources, as well as authority and deference accompanying such inequalities" (Knoke and Song, 2008, p69).

Centrality can be computed using several metrics; the most widely of which are *degree*, *betweenness*, *closeness,* and *eigenvector* centrality. The first three were proposed by Freeman (1978) in a groundbreaking approach towards social network analysis. These were initially designed for unweighted networks. Recently, Brin and Page (2012) came up with extensions to weighted networks. The fourth metric - *eigenvector* centrality - was later proposed by Bonacich (1987) and had its foundations in spectral graph theory. It became wildly popular after being used as the basis of the well-known Google's *PageRank*

algorithm. These measures determine the relative importance of a node within the network, showing how the relationships are concentrated in a few individuals and, therefore, giving an idea about their social power. Higher centrality values measures are associated with powerful actors in the network since their central position offers them several advantages, such as easier and quicker access to other actors in the network (useful for accessing resources such as information) and the ability to exert control over the flow between the different actors (Freeman, 1978). These central actors are also called "focal points" by Freeman. Many of these node-level metrics (e.g., degree, betweenness, and closeness) are *normalized* in some manner to perform comparisons of networks with different orders and sizes. These metrics have been normalized in this study, as well.

2.25.1.1 Degree Centrality

The degree centrality or valency of a node v, usually denoted as k_v, is a measure of the immediate adjacency and involvement of the node in the network and is computed as the number of edges incident on a given node or, similarly, as the number of neighbors of node v. The neighborhood N_v is thus defined by the set of nodes that are directly connected to v. The degree can be computed in at least two different ways: one, based on the adjacency matrix and the second, based on the neighborhood of a node. The equations below present each of these alternatives for undirected networks.

$$\boldsymbol{k_i = \sum_{j=1}^{n} a_{ij}} \qquad \boldsymbol{0 < k_i < n}$$

where a_{ij} is the entry of the i-th row and j-th column of the *adjacency* matrix

$$\boldsymbol{k_v = |N_v|} \qquad \boldsymbol{0 < k_v < n}$$

Where $|\mathbf{N_v}|$ is the neighborhood of node v. Despite its simplicity, the degree is an effective measure to assess the importance and influence of an actor in a social network. Yet, it has some limitations. The main one is that it does not take into consideration the global structure of the network.

In weighted networks, *strength* is the equivalent of degree and is computed as the sum of the weights of the edges adjacent to a given node, as expressed by the next equation.

$$k_i^w = \sum_{j=1}^{n} a_{ji}^w$$

There have been significant research efforts in studying the degree distribution of several networks, making it possible to classify a network based on this distribution. For instance, Barabási and Albert (1999) and Barabási and Bonabeau (2003) observed that most real networks follow a *power-law or Pareto distribution*[23]. The networks possessing this quality are known as *scale-free*[24], an expression coined by the same researchers (i.e., Barabasi and Bonabeau in their landmark 1999 paper). Other common functional forms are exponential (e.g., railways and power grids networks) and power-laws with exponential cut-offs (e.g., networks of movie actors and some collaboration networks).

2.25.1.2 Betweenness Centrality

Node *betweenness b* measures how closely a node lies between other nodes in the network and can be computed as the percentage of shortest paths that pass through the node. The formula is presented in the equation below. Nodes with high *betweenness* occupy critical roles in the network structure. They usually have a network position that allows them to work as an interface between tightly-knit groups, being vital elements in the connection between different network regions. From the social networks perspective, "interactions between two nonadjacent actors might depend on other actors in the set of actors, especially the actors who lie on the paths between the two" (Wasserman and Faust, 1994)

[23]Power law distributions describe those networks where the distribution of the degree of the nodes or vertices is highly skewedto the right with a large majority of vertices having a low degree and a small number having a high degree.

[24]Scale-free networks are those follow network whose degree distribution follows a power law, at least asymptotically. That is, the fraction P(k) of nodes in the network having k connections to other nodes goes for large values of k as $P(k) = k^{-\gamma}$, where γ is a parameter whose value varies between 2 and 3 typically, there might be exceptions though.

stresses the importance of the fair value of *betweenness*. These actors are also called gatekeepers since they tend to control the flow of information between communities:

$$\boldsymbol{b_v} = \sum_{\boldsymbol{s,t \in V(G) \setminus v}} \frac{\boldsymbol{\sigma_{st}(v)}}{\boldsymbol{\sigma_{st}}} \ ,$$

where σ_{st} denotes the number of shortest paths between vertices *s* and *t* (usually $\sigma_{st} = 1$) and σ_{st} (*v*) expresses the number of shortest paths passing through node *v*. This quantity can also be computed for edges. The *betweenness* of an edge is commonly defined as the number of shortest paths between nodes that run along a given edge of the network. It is quite useful in social network analysis since it allows discovering bridges and local bridges, which are, by definition, edges with high betweenness. In the context of social network analysis, bridges are connections outside an individual's circle of acquaintances. These connections are of great interest to individuals seeking to access new information and resources since they ease information diffusion across entire communities (Kossinets and Watts, 2006). However, situations like these are relatively rare in real-world scenarios. Even if they happen, the advantages they confer are usually temporary due to such edges' temporal instability. A more common and realistic situation is local bridges. The following equation indicates how this measure can be computed:

$$\boldsymbol{b_e} = \sum_{\boldsymbol{u,v \in V(G)}} \frac{\boldsymbol{\sigma_{uv}(e)}}{\boldsymbol{\sigma_{uv}}} \ ,$$

where $\sigma_{uv}(e)$ is the number of shortest paths passing through edge *e*, and the sum indicates that this fraction needs to be computed for every pair of nodes *u* and *v* in the network.

2.25.1.3 Closeness Centrality

Closeness centrality C_v is a rough measure of a node or vertex's overall position in the network, giving an idea about how long it will take to reach other nodes from a given starting node. Formally, it is the mean length of all shortest paths from one node to all other network nodes. Due to its definition, this measure is usually only computed for nodes within the network's largest component, using the equation presented in the equation below. In the social network context, *closeness* is a measure of reachability that measures how fast a given actor can reach everyone in the network.

$$\boldsymbol{C_v = \frac{n-1}{\sum_{u \in V(G) \setminus v} d(u,v)}} \ ,$$

where n is the number of nodes within the network, (u,v) is a dyad or a pair of nodes whose centrality is being measured, and d is the path between nodes u and v.

2.25.1.4 Eigenvalue Centrality

This metric is based on the assignment of a relative score to each node and measures how well a given node is connected to other well-connected nodes. This score is provided by the first *eigenvector* of the *adjacency matrix*. The basic idea behind *eigenvector* centrality is that an actor's power and status are recursively defined by the power and quality of his/her *alters*. *Alters* is a term frequently used in the selfish approach of social networks analysis. It refers to the nodes (or actors in sociological terminology) directly connected to a specific node or actor, called the *ego* (the node in focus). In other words, the *eigenvector* centrality of a given node i is proportional to the sum of i's neighbors' centralities, and this is the assumption behind the *eigenvector* centrality formula, which is as follows:

$$\boldsymbol{x_i = \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j}$$

Where x_i /x_j denotes the centrality of node *i* /*j*, a_{ij} represents an entry of the *adjacency matrix* **A** ((a_{ij}) = 1, if nodes *an edge connects i and j* and (a_{ij}) = 0 if there is no connection) and λ denotes the largest eigenvalue of **A**. *Eigenvector* centrality is a more elaborate version of *degree* centrality, with the difference being that it assumes that not all connections have the same importance by taking into account not only the quantity but especially the quality of these connections.

In theory, *eigenvector* centrality can be calculated for either undirected or *directed* networks. It works best, however, for the *undirected* case. In the directed case, other complications arise. First of all, a *directed* network has an adjacency matrix that is, in general, asymmetric. This means that it has two sets of *eigenvectors*, the left *eigenvectors*, the right eigenvectors, and two leading *eigenvectors*. So which of the two should be used to define the centrality?

In most cases, the correct answer is to use the right *eigenvector*. The reason is that centrality in *directed* networks is usually bestowed by other nodes pointing towards the node in focus, rather than by that node pointing to other nodes. For instance, on the World Wide Web, the number and stature of web pages that point to a page can give a reasonable indication of how significant or useful the page is.

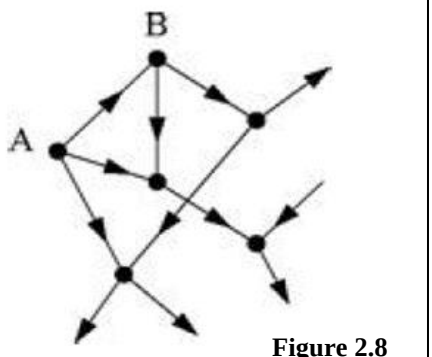


Figure 2.8

While the fact that the page might point to other important pages is of no significance, anyone can set up a page that points to a thousand others, but that does not make the page meaningful. Similar considerations also apply to citation networks and other-directed networks. Thus, the correct definition of eigenvector centrality for a vertex *i* in a directed network makes it proportional to the vertices' centralities that point to *i*, making eigenvector centrality of little use in *undirected* networks, such as the one this paper has focused on.

However, there are still problems with *eigenvector* centrality on *directed* networks. Consider Figure 2.8; Vertex A in this figure is connected to the rest of the network but has only outgoing edges and no incoming ones. Such a vertex will always have *eigenvector* centrality zero because there are no terms in the sum in the *eigenvector* equation above, which might not seem to be a problem: perhaps a vertex that no one points to *should* have centrality zero. Consider vertex B, which has one ingoing edge, but that edge originates at vertex A. Hence, B also has centrality zero because the one term in its sum in the above equation is zero. Taking this argument further, we see that a vertex may be pointed to by others that are pointed to by many more vertices, and so on through many layers. If the progression ends up at a vertex or vertices with in-degree zero, it is all for nothing—the final value of the centrality will still be zero. In mathematical terms, only vertices that are in a strongly connected component of two or more vertices, or the outgoing portion of such a component, can have non-zero *eigenvector* centrality.

Nevertheless, it is appropriate for vertices with high in-degree to have high centrality in many cases, even if they are not in a strongly connected component or out-component. Web pages with many links, for instance, can reasonably be considered important even if they are not in a strongly connected component. Since*acyclic* networks, such as citation networks, have no strongly connected components of more than one node, all nodes will have centrality zero. This fact renders the standard *eigenvector* centrality completely useless for *acyclic* networks.

2.25.1.5 Other Centrality Measures/Katz Centrality

In addition to the four most popular centrality measures, various others have been proposed, and indeed, many variations of the four measures above exist. One is a possible solution to the eigenvalue problem, which is allotting some centrality to each node regardless of its relative importance within the network. By this device, even nodes or vertices with zero in-degrees manage to have some residual centrality. The nodes pointed to derive secondary advantage from the residual centrality values of zero in-degree nodes. Thus, any node with many low centrality nodes pointing to it will still garner some

importance. This modification was suggested by Katz in 1953 and is termed *Katz centrality*. The equation below defines the measure. The term $\boldsymbol{\beta}$ in the equation is the *residual value* allotted to each node, described above.

$$\boldsymbol{x_i = \frac{1}{\lambda}\sum_{j=1}^{n} a_{ij}x_j + \beta}$$

In other work, Brandes (2008) describes variants of *betweenness* centrality that consider paths only up to a specific path. For example, *n-path centrality* uses a simple counting algorithm to find the n-paths in a binary matrix(Sade, 1989). Lindelauf (2011) discusses the power of using centrality measures based on cooperative game theory, which allows incorporating more parameters than network structure alone into the analysis and argues that game-theoretic centrality measures, or power indices as they are called, are useful in the study of covert networks. He presents a specific centrality measure based on the *Shapley*[25] value that he uses to analyze covert networks.G´omez et al. (2003) presented an entirely different perspective on centrality based on game theory analysis. Borgatti and Everett (2006b) did an exhaustive and comprehensive classification and comparison of centrality measures.

2.25.2 Higher Level Social Network Tools and Measures

Apart from node-based measures, various higher-level structural measures enable an overall grasp of the network's architecture. Such measures may comprise other combinations of the node centrality measures just discussed, or maybe entirely independent.

[25] The Shapley value is a common power index and is used to determine the importance of the players, i.e., to set up a ranking based on all the available information.

2.25.2.1 Dyads

The level higher than nodes in a network is the dyadic network comprising pairs of nodes or vertices. These are being referred to as *dyads* or *node-pairs* without any loss of meaning. There are $(N^2 - N)/2$ dyads in a network of N nodes. In directed social networks, there are $(N^2 - N)$*ordered* dyads. At the dyad level, the key to measurability is the presence of an edge or tie that relates both nodes in the dyad (nodes in a dyad or dyad are termed as-*constituent nodes* in this study). The edge between a pair of constituent nodes is described in terms of its strength, intensity, variability in time, etc. One other characteristic of a tie is if it's direct or indirect via intermediary nodes. Indirect edges are constructed based on the regular equivalence of the constituent nodes rather than structural equivalence. Thus, a pair of disparate nodes can be related through homophily ("birds of a feather stick together") or complementarity ("opposites attract") properties (Knoke, Yang, 2008).

2.25.2.2 Adjacency Matrix

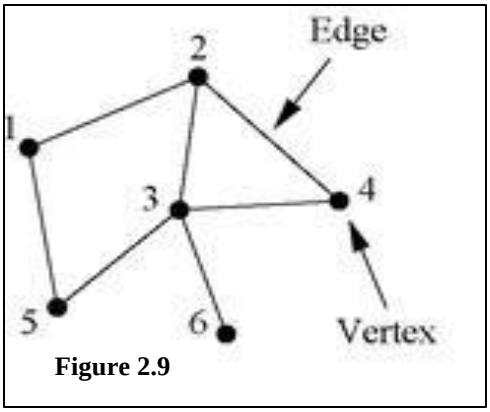


Figure 2.9

The most common way in social network analysis to represent structural ties or edges is through *adjacency matrices*. Newman (2011, p110) describes the construction of a simple adjacency matrix through the figure of a smallish network shown in Figure 2.9. If we denote an edge between nodes *i* and *j* by (*i*,*j*), then the complete network can be specified by giving the value of *n* and a list of all the edges. For example, the network in the figure has *n* = 6vertices and edges (1,2), (1,5), (2,3), (2,4), (3,4), (3,5), and (3,6).Such a specification is called an *edge list*. Edge lists are used to store the structure of networks, but this method of storage representation is cumbersome for very large networks. A better representation of the relationships in a network is the *adjacency matrix*.

The *adjacency matrix* **A** of a simple graph is the matrix with elements $\boldsymbol{A_{ij}}$ such that-

$$A_{ij} = \begin{cases} 1 \; if \; there \; is \; an \; edge \; between \; nodes \; i \; and \; j \\ 0 \; otherwise \end{cases}$$

Based on this formulation, the *adjacency matrix* of the network in Figure 2.9 is shown in Figure 2.10

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Figure **2.10** Adjacency Matrix

One of the points to notice about an *adjacency matrix* is that, first, for a network with no self-edges such as this one, the diagonal matrix elements are all zero, and second that the matrix is symmetric, since if there is an edge between *i* and *j*, then there is an edge between *j* and *i*. One of the significant advantages offered by this implementation is that all attributes in a network, including entities higher in level than nodes, such as dyads, communities, and cliques, can take the place of nodes in the above representation to yield *adjacency matrices* of higher orders. In the current study, a matrix showing the *similarity* of *edge-pairs* akin to an *adjacency matrix* is built using dyads or edges rather than nodes. The adjacency matrix variant is the *incidence matrix,* which describes the relationships between the members of two groups in a network (such networks are termed *bi-partite* networks or graphs, for example, if several individuals are members in several clubs). Unlike adjacency matrices, *incidence matrices* tend to be non-square.

2.25.2.3 Triads

The level higher than dyads is the *triadic relation*, which involves *triple* nodes. A social network of N nodes has $\binom{N}{3}$ Triples. The presence or absence of relations amongst the triads' constituent nodes gives rise to 16 distinct triad types. Triads are generally used in gauging sentiments in a social network, like friendship or rivalry. Triadic analysis can be used for extrapolating relations between a pair of nodes to a third node constituting the triangular relationship. Knoke and Yang (2008, p14) term this property of triads *transitive triadic relations,* which they describe thus, "if A chooses B and B chooses C, does A tend to choose C?" The answer leads to *triad closure* and thence to the crucial question of *pre-existing ties* in social networks, especially covert networks where such ties tend to be latent or invisible to the observer.

2.25.2.4 Community Structures at the Network Level

The highest level of analysis above the microlevels of node, dyad, and triad is the network as a whole. This level may include quadrangular relationships amongst four nodes or polygonal relationships with four or more nodes. Community structures, including dyads and triads, can take the place of nodes or vertices in polygonal relationship patterns to bring out how communities, rather than nodes, interact within a social network. The macrostructures above the three micro levels are essentially community structures into which social networks naturally tend to divide. For instance, the World Wide Web divides into groups of related web pages. The most common community constructs are *cliques, plexes, cores, groups,* and *components*.

Cliques are maximal[26] subsets of the nodes in an undirected network such that an edge connects every member of the set to every other member. A clique's existence is very

[26]Maximal implies that no further nodes can be added to the subset of nodes without destroying the property of complete connectivity of the member nodes with each other.

significant in less dense social networks as it depicts a closely-knit community. Cliques can overlap each other, i.e., member nodes can belong to two or more cliques. Cliques are, however, not real-world constructs, and more often, there are groups within networks, which, while being closely knit, may not exhibit the property of complete connectivity. A more common occurrence is the *k-plex clique*. A *k-plex clique* of size *n* is a maximal subset of *n* nodes within a network such that each node is connected to at least *n-k* other nodes. A related structure is a *k-core clique,* a maximal subset of nodes such that each member node is connected to at least *k* other member nodes. A*k-core cliques* is an *(n-k)-plex clique*. Unlike *k-plex cliques, k-core cliques* are non-overlapping, and if any member node is a member of another *k-core,* both *cores* will fuse into a larger *core*.

Groups are similar to *cliques* and are defined as subsets of nodes with at least as many connections to nodes within the *group* as to nodes outside. A more useful definition was proposed by Radicchi et al. (2004). They defined *groups* as subsets nodes, whose total number of connections with other nodes within the *group* is always higher than the total number of connections they have with nodes outside the *group*. This view of groups offers a powerful way of detecting communities within networks.

Components are maximal subsets of nodes such that each node is reachable by some path from each of the others. A *k-component* is a variant of this concept and is defined as a maximal subset of nodes, connected to the others within the set by at least *k* node independent paths. Sets with values of *k*=2 and *k*=3 are termed as *bicomponents* and *tricomponents,* respectively. Each higher value of *k* implies that the higher value set is a subset of the lower value subset of *k*. Thus, a *tricomponent* is a subset of a *bicomponent,* and both are a subset of a *component*. An interesting spin-off of this community structure is that it can also be looked at as a maximal subset in which if a node needs to be unpaired from another node in the *k-component,* at least *k* links need to be disconnected. The concept is a beneficial one in the disruption strategies of covert networks wherein central actors within the network need to be dislodged to study before their impact on network function, and efficiency can be studied.

The different levels of network analysis discussed above imply that the characteristics at one level cannot be simply deduced from the information available at other levels. For instance, the transitivity of relations is a significant input for friendship formation ("a friend of my friend is my friend"), which is unique at the triadic level but not seen at the node-level or at the dyadic level. Knoke and Yang (2008, p14) make this aspect clear with an illustration:

> "Consider two scientific research communities with roughly similar egocentric, dyadic, and triadic structures in their scientific discussion networks. But, if the first community's complete network is fragmented into several unconnected subgroupings, many scientists may be unable to communicate with others indirectly. If the second community's entire network contains ties that bridge and broker relations among its subgroupings, we could anticipate a more rapid and widespread flow of information and a higher scientific innovation rate. This protean capacity of network analysis to address problems at multiple levels of analysis by encompassing emergent structural relations lies behind its rapid increase in popularity as a framework for theorizing and guiding empirical research."(p.14)

2.26 Social Network Analytics in Covert Networks

2.26.1 Background

As has been discussed, social network analysis is a quantitative methodology to model relationships and behavior amongst networked actors. It focuses on the ties amongst the actors and the implications of these ties on the network structure. The structure or topology of a network can affect, promote, and constrain individual actor behavior (Wasserman & Faust, 1994). The quantitative nature of social network analysis enables the depiction of the network structure and its implications upon collective and individual behavior in a mathematical format. It also aids the identification of actors involved significantly in collusive conduct and gives shape to the determination of communities of actors within

the network. Additionally, the quantitative basis enables the detection of changes over time in network structure, actors' roles, and community formation or dissolution.

It is necessary to note that any social network model is built to a specific context. The data are collected, and the model is constructed in a particular set of real-world questions that the study is attempting to solve. This specificity and the problems associated with it determine the selection of the network analytic techniques. As these analytical techniques are mathematically oriented, they have certain assumptions associated with them. These assumptions crystallize the data requirements to provide inputs to the chosen social network analytic model's mathematical needs. Thus, the conclusions and interpretations derived from the analytical techniques are based on the contextual considerations within which the social network model operates and as well as the data collected at the initial stages. These quantitative aspects of social network analysis enable law enforcement and other governmental agencies to identify key actors or network facilitators and target them or keep them under active surveillance. Removing key actors, detaining leaders of an organized criminal gang, or targeting those with vital skills, such as explosive experts in a terrorist network, may inhibit a network's negative output without large scale arrests or even complete disruption. The network models built based on this type of analysis are dependent upon the correctness of the information sources providing data. When dealing with criminal or terrorist networks, these sources may provide unreliable information, leading to wrong conclusions from the model. The sources may corroborate or disprove reports from other sources, leaving the analysts to decide which data are to be used and which to be discarded.

Social network analysis methodologies have predominantly evolved from research conducted on open networks such as corporations, businesses, governmental organizations, clubs, social groups, and activities where data is available and collected without any surveillance. In contrast, covert organizations considered threats to national security are structured and have mechanisms built into their structures to counter conventional data collection techniques' effectiveness. These unique mechanisms present additional layers of challenges in applying social network analytic techniques to these

networks. A deeper understanding of these organizations and associated mechanisms must address the concomitant implications of modeling such networks. Accordingly, a comparative approach that brings out the similarities and differences between traditional social network analytic approaches and modified approaches to model covert networks has been outlined in the following sections. The following provides an outline comparing traditional social network analytic approaches with their modified versions for modeling covert networks.

There are several studies which have examined various aspects of social network analysis study covert networks, covering the spectrum from terrorist groups to organized criminal networks and insurgencies (Sparrow, 1991; Coles, 2001; Reed, 2006; van der Ressler, 2006; Hulst,2009) with several of these studies introducing new methodologies and algorithms adapted to the specific models (Renfro, 2001; Sterling, 2004; Hamill, 2006; Seder, 2007; Farley, 2007; Geffre, 2007; Herbranson, 2007; Leinart, 2008; Kennedy, 2009;). Acquisition of specialized knowledge of dark networks and their structures, operations, processes, and mechanisms, coupled with the roles of the key actors, has the potential to grant deeper insight into inherent vulnerabilities that are susceptible to exploitation and, more crucially, to make network activities predictable in the future.

2.26.2 Data Collection in Network Studies

The first steps in social network analysis necessarily involve the collection of data about the network. This data is then used to build a visualization of the network. As we saw earlier, all networks have graph-like structures and can easily be reduced into a matrix format. The matrix is a very malleable mathematical construct that can lend itself to various articulations, leading to viable computational models. Social network analysis includes in its folds diverse methodologies to collect data. This section presents a brief synopsis of the data collection procedures followed in conventional social network analysis based primarily on the exhaustive chapter on this subject by Knoke and Yang (2008, pp.15-20). The synopsis includes comments on how data collection procedures that hold true for

ordinary social networks come up short for covert networks. There are brief discussions on various modifications in data collection and collation, which are in practice to cater to covert networks.

2.26.3 Data Distortions in Covert Networks

Imperfections in social networks data stem from various causes. A primary source of error relates to all social network modeling issues, namely, boundary specification. The modeler must decide the inclusion and exclusion of specific data elements and decide heuristically which actors and which relations are to be included in the data set. The issue acquires further complexity when selecting the associated variables to collect on each actor and each relation. Difficulty in obtaining specific nodal or edge information, such as actors' demographic data or the ability to measure relationships' intensities, may drive boundary specification decisions.

The second set of challenges arise in the data collection part of the analysis. An improper data collection design, inherent inaccuracies generated by the specific data providers, or the intentional lack of information characteristic of dark networks – may introduce extraneous, spurious, or inaccurate data. These factors also potentially prevent the comprehensive collection of essential elements that can significantly impact the subsequent analysis and results. In short, SNA is almost guaranteed to be conducted in an environment of imperfect data. (Morris & Deckro, 2018).

2.26.4 Boundary Specifications in Social Networks

It has been discussed in depth that covert networks tend to be incomplete and have fuzzy boundaries. This quality stems from their innate tendency to confine information flows from leaking out. In many instances, the networks observed today may be far removed from their topologies in the past. Thus, data collection during the early stages of

surveillance of a covert network is crucial. Historical data reduces the deficit of information about the network and its constituents and helps foster a sense of caution as new data is evaluated, and conclusions are drawn. Data collection is a necessary first step in the study of covert networks, and all social networks, which by design, are dynamic and fluid.

One of the first questions that a researcher faces while collecting data is setting the limits (of data collection). The answer to this is contextual and may not be any limits (Barnes, 1979 as cited in Knoke & Yang, 2008). Strategies for collecting data have been enumerated by Knoke and Yang (2008, pp.15-20), including the three[27] generic approaches developed by Laumann, Marsden, and Prensky (1989), namely, *positional*, *relational,* and *event-based*. The positional strategy uses actors' attributes, including their memberships in a formal organization or their occupancy of a well-defined position for inclusion in a network.

Positional approaches identify actors similar in status in any organization and who may not be connected through any links. We may equate position with the concept of *regular equivalence* in a graph-theoretic context. This technique runs the risk of producing sets of entirely unrelated actors who might just happen to share some attribute; This is evident from the example of the Cameroonian women who belonged to the same hometowns and who were interviewed by researchers about collective memory choices, ended up being slotted to the same sets based on their hometowns. However, they were often strangers (Knoke & Yang, 2008, p.16).

Relational approaches rely on the actors being interviewed to identify others for inclusion in the group. This strategy is similar to *snowball sampling***,** *reputational* method, *fixed list selection*, *expanding selection,* and *k-core methods*. *Snowball sampling* consists of requesting interviewees to nominate others whom they are related to. In turn, these actors are requested similarly and so on; this method effectively ascertains memberships of

[27]Initially, there were two strategies defined, namely, *realist* and *nominalist* in respect of data collection by Laumann, Marsden and Prensky (1983) which were replaced by the three approach model by them in 1989.

remote populations. *Reputational methods* relate to asking the most well-informed actors about others who need to be included. *Fixedlistselection* uses questionnaires that allow respondents to reply regarding ties with actors chosen by the interviewers themselves. The *k-core* method finds subsets of actors wherein any particular actor has relations with at least *k* other actors. This concept mirrors the concept of the *k-core clique* in graph theory.

Event-Based methods choose actors who are participants in a pre-defined set of activities at particular times and places. Knoke and Yang (2008, p.20) noted examples of Southern Californian beachgoers who visited the beach at least three days in a month. They cautioned that while using this methodology, meticulous attention needs to be paid to categorize activities cleanly. They add a caveat that event-based methods run the risk of missing data by not identifying activities that ought to have been included. A solution would be to observe multiple editions of the set of activities in question to not miss out on any aspect or any actor who is essential and who may not be essential in a particular edition.

The issue of boundary limitation is challenging in complex and dynamic networks. A key hurdle in the study of formal and informal responses within complex problem domains is that they are, by definition, ill-defined, leaving it to the analyst to make choices concerning who is and is not to be considered part of the network. To add to the complexity, networks are social constructs with extreme specificity whose evaluations vary from case to case, depending on whose perspective is being privileged (Nowell et al. 2016; Nowell et al. 2018; Mandell & Keast 2008). For example, network actors may differ from each other in their understanding of who belongs to or within the network. Further, scholars may differ in their perspectives about a network relative to those actors involved in the network (Mandell and Keast 2008; Turrini et al. 2010). Finally, two scholars may study the same social phenomenon yet draw the boundaries of the network differently. Consequently, challenges relating to network boundary determinations are endemic to all network research. They are particularly relevant when applying network analysis to gain insight into social structures that anchor complex domains (Weber and Khademian 2008).

2.26.5 Boundary Selection in Covert Networks

As discussed above, the boundary specification problem involves the inclusion and exclusion of actors and relations. Rule-sets are defined *apriori* to describe an actor's inclusion in the network. Inclusion or exclusion depends on the actors' characteristics, their affiliations, or other specifications. Relation types may also be identified for inclusion in the network from the set of all relations. This triage process entails reducing the overall pool of actors or relations and depends on the analysis being conducted (Wasserman & Faust, 1994). Triage is incredibly challenging in covert network actors. They may deliberately blur the lines between their professional and personal lives, causing difficulties in clearly delineating where lawful transactions and activities end and where illicit operations begin, which has the effect of creating a fuzzy boundary whose exact contours the social network analyst must decide on (Sparrow, 1991). Laumann, Marsden, and Prensky (1983) saw two possible approaches to the boundary specification problem: realist and nominalist. The *realist* approach defines the boundary by assuming "that a social entity exists as a collectively shared subjective awareness of all, or at least most, of the actors who are members" (Laumann et al., 1983, p. 21). For formal organizations with clear membership, this type of selection is effective; however, this assumption makes the boundary fluid when dealing with covert organizations such as terrorist networks. Secondly, it has the potential to create a paradox where actors may consider themselves part of the social network. In contrast, members of the social network may consider the actor excluded from the collective. The reverse of the paradox could just as quickly occur.

The *nominalist* approach, according to Laumann et al. (1983, p. 21), is one where the "analyst self-consciously imposes a conceptual framework constructed to serve his analytic purposes." The social network is thus defined by arbitrary criteria that serve the analyst's lines of inquiry. In opposition to the realist approach, the social network's self-defined boundary is no longer an assumption, but an empirical question of how it compares against the analyst's defined boundary (Laumann et al., 1983). Arbitrary boundary selection by the analyst could significantly distort the results. However, if correctly done, this method could capture the data to precisely analyze the question while concurrently

eliminating extraneous data that could distort the results. The data collected for a social network analysis study includes actors, relations, events, affiliations, or any combination. The inclusion and exclusion rules determine which elements of the four data types are incorporated into the social network model. Various inclusion and exclusion rules can be applied exclusively or in combination to determine which social data elements, expressly which actors, relations, events, or affiliations, are incorporated into the social network model and subsequent analysis (Laumann et al., 1983).

According to Laumann et al. (1983), rules for inclusion and exclusion in actor boundary specification tend to positional, reputational, or a combination of both. Positional rules test actors' attributes for inclusion into the social network. The actor's attributes fulfill a specific position within an organization, which gives the category its name. The second type, a reputational rule, "utilizes the judgments of knowledgeable informants in delimiting participant actors" (Laumann et al., 1983, p. 23). Hybrid rules generated from both types are standard in research (Laumann et al., 1983). Applying these rule categories to real-world problems generates a wide range of options to select inclusion actors into the social network model in question. According to the three rule categories defined by Laumann et al. (1989), an actor's inclusion and exclusion may be based upon membership with particular organizations, positional specification, demographic data or other actor attributes, involvement with specific relation types, event attendance, identification of inclusion by different actors, or a combination of these factors (Kossinets, 2006; Kossinets 2008; Nowell et al. 2018).

If any network's internal transactions are of interest, limiting the network boundaries only to those acknowledged members of the organization may be appropriate to enhance accuracy in representation and interpretation (Marsden, 1990). Examples may include a human trafficking organization in which the actors may have several relations with suppliers, transporters, harborers, and customers. Still, to accurately describe internal processes, the social network may need to be limited to only a subset comprising core members. Positional specifications restrict the network members to those who occupy positions of rank in a formally constituted group. Going by this definition, a terrorist

leadership social network may only include those in command of a single detachment (Kossinets, 2008).

The use of specific actor attributes, including demographic data of the individual, such as gender, age, or rank, enables a reduction in extraneous nodes, which boosts results by limiting the network to significant value actors. A real-world network where this is applicable is an organized crime network based on close family ties, such as the Italian mafia, or specific insurgent organizations based primarily on familial connections (the Haqqani Network in Afghanistan is a good example). The selection of actors may be limited to those who only possess specific relations (Marsden, 1990). A good example is that of terror funding networks, where actors with financial connections may be relevant in determining the network structure, which is easier said than done since routes of financial transactions may include formal (overt) channels like international banking and informal,e.g., hawala channels. These methods are usually based on the individual perspectives of actors pre-identified as belonging to the network; hence, they subject to bias. Kossinets (2008, p. 5) notes that "actors may disagree in their perception of social structure; they may be attributing different weights to certain other actors, relationships or types of relationships."In other words, a paradox can come about in which an actor believes he or she is part of a network, while the other actors do not include him or her as part of the system. Anti-government insurgencies are good real-world examples of this fallacy; individuals within an insurgent network may construe differing thresholds of inclusion activities with the network. Some may view not supporting government forces or donating funds and resources as justification for inclusion, while others may set the threshold higher as in actively fighting, and so forth.

Relational rules, on the other hand, only allow actors to possess specific, defined relationship types into the network model (Laumann et al., 1983; 1989). Relations for a given social network are chosen to represent specific types of actor interaction, and the exclusion of extraneous relationship types condenses the network to represent only required interactions. Incorrectly bounding the relationships in the model removes links that are present in the actual network under investigation.

Events and *affiliations* also provide a basis for inclusion and exclusion rules. An event or specified activity selected by the analyst as relevant to the network allows only those actors and their internecine relationships derived from participating in the event or activity to be included in the social network model (Laumann et al., 1983). Similar boundary specifications can be made applicable to affiliations as well. In some instances, affiliation data between actors is generated by event attendance and can also be derived through membership in multiple organizations and groups. In this category of boundary specifications, actors who attend a particular set of events or are affiliated to the same organizations are included as part of the network (Marsden, 1990). However, Kossinets (2008, p. 5) warns that event attendance "is particularly error-prone and is best described as convenience sampling." To quote an example, such a technique can produce distorted results in the case of terror networks where some of the main actors may not have affiliations to common ideological platforms but might have joined together in some larger cause. By selecting particular affiliations, the analyst runs the risk of missing out on these vital individuals.

Each of the three approaches proposed by Laumann et al. (1989) has strengths and limitations related to (1) its ability to reveal formal and informal institutional norms and structures; (2) its ability to capture isolates and disconnected sub-groups, and (3) its ability to represent social relations over time. Through recognizing these strengths and associated limitations, researchers are better informed to choose the most appropriate strategy (Nowell et al. 2018). To sum up, the analysts' solution is to have complete clarity about a given boundary decision's comparative advantages and disadvantages. Transparency, deliberate consideration of boundary consequences, and evidence of informed choices are all reasonable responses to the challenges of dealing with boundary specification with messy networks in complex problem domains. In 1989, Laumann et al. warned scholars of network research boundary determinations' methodological and theoretical significance. Despite this, decisions concerning how network boundaries are determined are rarely discussed or critically examined, and the extant literature provides little advice to guide a scholars' decision or to consider the consequences of their design choices

2.26.6 Data Collection Procedure

Data collection methods include *single* and *multiple-name generators*, *position* and *resource generators*, *measurement of total personal networks*, and *archival documents*. *Single* and *multiple-name generators* obtain information about respondents' alters in an ego-centric network, name interpreters that obtain information on each alter, and their ego relations. A single-name generator may produce a core set of alters but fails to identify all the contacts which can be elicited using multiple-name generators. A single-name generator uses a single item questionnaire, whereas multiple item questionnaires are used in multiple-name generators. To impose reasonable boundaries on the network size, stringent limits need to be imposed. These constraints are of four types (1) *role* or *content-based* constraints where relationships based on particular roles are used to build networks, (2) *geographical* constraints, which put restrictions on the area, (3) *temporal* constraints, which are based on a window of time (4) *numerical* constraints which limit respondents to naming up-to a certain number of other actors.

Positional generators collect information from respondents about their ties with other actors in specific organizational hierarchy positions. The resultant network size depends on the choice of social positions. Depending on a respondent's origin (relation to someone in the organization), tie-strength can be assigned. That is, if the origin is high, tie-strength,[28]whether strong or weak, is immaterial, and if the origin is low weak ties are beneficial. *Resource generators* typically collect information based on a quantity called *social capital*, which encompasses all forms of assistance or resources that a respondent is receiving and that he or she can enumerate during data collection. A resource generator asks respondents if they know anybody possessing specific resources or skill-sets rather than just their positions. This helps build a network with more layers and is more varied than a positional generator based approach can produce.

[28] A tie is deemed strong if the relation is direct relative, and is termed weak if the respondent's contact is only an acquaintance.

Measuring *total personal network* involves gathering information on all alters known to the ego. This technique invokes name generators, especially "invented" (or designed), to elicit this information, which is done using a *checklist method* in which respondents are given randomly generated names. They are asked if they know anyone by these names. This methodology is used through the *reverse small-world method*(RSW)**.** Researchers create fictitious personalities with invented names, random attributes such as age, sex, gender, membership in organizations, occupation, location, etc. The respondents are then exposed to this list and asked to name alters who could provide a link to these fictitious people. The second step in this process involves querying the respondents about the alters they've named in the first part. These methods have been combined to produce optimal results sets.

Archival documents reveal past information about networks. Such documents are relatively cheap and less resource consuming than the survey-based methodologies discussed above. With the vast majority of the archival data now available in digital format, data-mining tools have become indispensable for extracting useful information from vast quantities of data. The explosive growth in the internet-based archival frameworks and the development of increasingly sophisticated web-crawling search engines have transformed this data retrieval method. However, archival documents often have to be combined with surveys to present a complete picture.

> Archival documents are also a handy way to counter the incompleteness that is a common feature of covert networks; particularly, archival information can firm up the pre-existing ties in a covert network in a significant way. Networks built out of archival information can be extrapolated to current times using sophisticated mathematical tools (Markov chains and Bayesian statistics are good examples) to configure what the network should look like in current times. This model can be used as a null set to compare with the structure we have in hand. Substantial differences can alert us to the possibility of deception by the actors in the network.

Data collection procedures in covert networks are likely to differ significantly from those prescribed for conventional networks described in the section above. In most instances, information about covert networks or covert cells within open networks is obtained after the "deed is done." The terrorist network structure of the 9/11 attackers was constructed by Krebs (2002) painstakingly based on open-source information available from television channels and newspapers. The difficulties Krebs faced while mapping out the network is best described in his own words:

> "I set out to map this network of terrorist cells that had so affected all of our lives. I would be mapping a 'project team' – much like the legal, overt groups I had mapped in countless consulting assignments. **Both overt and covert project teams have tasks to complete, information to share, funding to obtain and administer, schedules to meet, work to coordinate, and objectives. How a typical project team does all of that is easy to map and measure using several sets of ties – task, resource, strategy, and expertise links. I was surprised at the difficulty of this particular effort – both in data definition and discovery**. My data sources were publicly released information reported in major newspapers such as the New York Times, the Wall Street Journal, the Washington Post, and the Los Angeles Times. As I monitored the investigation, it was apparent that the investigators would not be releasing all pertinent network/relationship information and maybe releasing misinformation to fool the enemy. I soon realized that the data was not going to be as complete and accurate as I had grown accustomed to in mapping and measuring organizational networks." (pp. 43-44)."

The portions highlighted in the passage reproduced from Krebs's paper are interesting. Krebs was used to mapping out networks as a part of his work schedule. Even so, all his knowledge and experience in this domain had not prepared him for the challenges that this particular problem posed. His sources were from open channels, which are often incomplete and may include incorrect information based on how the journalists(the primary sources or collectors in this case) have extracted, interpreted, and expressed them. It may be argued that this case was just a "one-off," and plenty of "complete" information is available in most cases. Unfortunately, for covert networks, these characteristics are the

norms and not the exception. Baker and Faulkner's (1993) pioneering works and Erikson (1981) are useful pointers in this direction. Wayne Baker and Robert Faulkner (Baker and Faulkner 1993) studied conspiracies in the heavy electrical industry equipment suppliers have recommended looking at archival data to derive relationship data. The data they used to analyze illegal price-fixing networks were mostly court documents and sworn testimony, including various witnesses' accounts. Krebs (2002) adds that others did not directly observe the hijackers of September 11th in great detail. Bonnie Erickson (Erickson 1981) stresses the importance of "trusted prior contacts" for a secret society's effective functioning. Krebs 92002) mentions that the 19 hijackers "appeared to have come from a network that had formed while they were completing terrorist training in Afghanistan. Many were school chums from many years ago, some had lived together for years, and kinship ties related others. Deeptrusted ties that were not easily visible to outsiders wove this terror network together." (p.44).

The issues faced by analysts while conducting surveillance of covert networks or sub-networks are generally two-fold. The data they receive is "current," i.e., what is existing at that timepoint, shorn of past connections, and the impact of pre-existing ties on the current topology. (A detailed discussion on pre-existing ties has been done earlier in Chapter 2 of *ibid* work). The second challenge is that the data is only obtainable post hoc, i.e., after the incident. As has been stressed repeatedly, such "after the fact" analyses may be useful in investigations to establish guilt or provide a roadmap for future interventions by law enforcement agencies in similar cases, but they are practically useless in predicting imminent offenses or adverse outcomes. For successful prediction modeling of covert networks, data collection procedures need to have radically different properties. This dissertation has already delineated the metrics that the requirements for successful predictive modeling in this context, i.e., accurate detection of the covert networks and the ability to predict possible outcomes within a reasonable timeframe.

These requirements have a significant bearing on data collection for building the network model. Two things to be on the lookout for while collecting data about covert or dark networks are (1) structures from past instances which have led to similar outcomes as well

as isomorphic structures in the network being investigated (2) that data is being collected from multiple sources for corroboration.

Regarding the issue of structures to look for, it should be borne in mind while collecting data for covert networks that violent or adverse outcomes, despite their sudden appearance, are seldom sudden in their build-up. Black (2004) has described the violence as arising from pre-defined structures that have arisen over time in reaction to prevailing political, legal, and cultural conditions. Watts (1999) also addresses the same issue and points out that "networks can affect a system's dynamical behavior in what might be termed an active and a passive sense; and that it is the passive sense…Active implies that the network is a device to be manipulated consciously for an actor's ends; passive implies that the network connections themselves, in concert with blind dynamical rules, determine the global behavior of the system" (p.3). Once formed, the structures are designed to produce the outputs they do, and the characteristics of individual actors staffing different positions in these structures are not that significant. It is good for an analyst to have a library of structures and topologies that have produced outcomes in earlier instances currently being investigated.

Second, the analyst should collect the same or similar data from multiple sources and *corroborate*[29] all the information. Various surveys and sampling mechanisms related to conventional networks were discussed previously. All of these tend to believe the first source that is encountered. It needs to be kept in mind that participants in covert networks will not be forthcoming about their roles, functions, relationships, and even identities. It's best to presume that deception rules their behavior. Obtaining second or even third-hand accounts about the same information is prudent, and bolstering the same through circumstantial evidence from the field is also recommended. Sociological treatises abound on the use of corroboration and triangulation for data collection and rectification purposes. Refitting these methods to the study of dark networks is but a step also.

[29] Corroboration is a legal term that refers to the requirement that any evidence adduced be backed up by at least one other source.

The value of obtaining data from multiple sources is well documented in research-

> "qualitative analyses of text that are often supplemented with other sources of information to satisfy the principle of *triangulation*[30] and increase trust in the validity of the study's conclusions. It would not be uncommon, for example, to analyze transcribed interviews along with observational field notes and documents authored by the respondents themselves. The purpose of multiple sources of data is corroboration and converging evidence." (https:\\www.sagepub.com/sites/default/files/upm-binaries/43144_12.pdf, p.350).

Flick (1992, 2018) addresses validating results through triangulation with other results by enunciating the idea of a master reality behind the use of several methods and reducing bias. On this topic, Mathison (1988, p.13) states, "Good research practice obligates the researcher to triangulate, that is, to use multiple methods, data sources, and researchers to enhance the validity of research findings … it is necessary to use multiple methods and sources of data in the execution of a study to withstand critique by colleagues".

2.26.7 Informant Bias

Knoke and Yang (2008, p.35) define informant bias as the discrepancy between self-reported and actual behaviors. Informant bias occurs due to many reasons, including the inability of respondents to report their behavior accurately, tendency to "impose categorical from on noncategorical affiliation patterns," propensity on the part of the sources to "correct their perceptions to maintain a balanced network among their close friends," Other reasons include the belief in informants that they are more central to the scheme of things than others, Bias may also result from the varying abilities of sources to recall events; those have excellent domain knowledge tend to introduce errors by "reporting on nonexistent members." Knoke and Yang (2008, p.36) quotes a study by H. Russell Bernard and his associates who did a comparative analysis of seven sets of paired communication network datasets and found that "about half the informants' self-reports

[30]Triangulation: A method used in qualitative research that involves crosschecking multiple data sources and collection procedures to evaluate the extent to which all evidence converges.

were erroneous in some ways" and concluded that cognitive data about communication couldn't be used as a proxy for the equivalent behavioral data. However, Knoke and Kuklinski (1982) have been cited in Knoke and Yang (2008, p.36) as having challenged the conclusions of Bernard and his colleagues by questioning the "accuracy and unobtrusiveness of the observers," which led to Bernard acknowledging that "facilitative factors" such as expertise of the sources in some field, could lead to a reduction in informant biases.

Distortions may also result from the tendency to forget less prominent players and falsely recall major actors in a network, false recollections of interactions that never occurred, and false recall of persons interacting. Consensus between informants leads to a reduction in bias. Knoke and Yang (2008, p.37-38) remark that "Highly knowledgeable informants produce unbiased data about long-term repeated patterns. They also tend to produce consensus answers to questions, which indicates greater validity[31]." The conclusion that may be drawn from the above is that analysts while collecting data in the field to build network models, must be picky about whom to interview and choose wisely from amongst the pool of respondents who are regarded as knowledgeable and collect information about relationships from sources who have close ties with the others sought to be included. The analysts also need to be aware of sources' predilections to portray themselves as being more central in the network than the average perceptions of the others about these individuals.

2.26.8 Reliability of Data

Reliability as a measure of the extent to which "a particular instrument, when applied repeatedly to the same subject, yields an identical result every time." Reliability measures include interobserver reliability, test-retest reliability, and internal consistency reliability, including split-half reliability and Cronbach's α reliability[32]. A mathematical co-efficient

[31] Validity refers to the extent to which the data collected gives a true measurement / description of the ground realities. Data is only useful if it actually measures what it claims to be measuring.
[32]Cronbach'salpha is a measure of internal consistency, that is, how closely related a set of items are as a group.

frequently used to measure reliability is Jaccard's coefficient (coincidentally, this dissertation uses the same measure to link pairs of related covert edges while identifying covert communities with common intentions). If we go by this scale, reliability values vary from 0.00, indicating nil reliability, to 1.00, showing full reliability. A high level of consensus amongst respondents results in high levels of reliability of the information. Knoke and Yang (2008, p.39) summarize the concept of reliability by stating that "individuals with high reliability tend to have a higher correlation in their self-reports than do individuals with low reliability….though reliability predicts validity, informants with more validity will have more similar responses to one another than informants of low reliability."

They conclude that social network data is substantially different from other types of data. The impacts of informant reliability and validity measures differ significantly from conventional data. This is particularly so in egocentric analyses where high correlations amongst informants in choosing similar alters indicate high reliability. High informant validity is implied by high correlations of an informant's description of their alters' characteristics, such as age, gender, education, and economic status (p.40).

2.26.9 Missing Data

Missing data is a significant source of data distortion in social networks. The impact of missing data on building network models is well illustrated by Knoke and Yang (2008, pp. 41-44). They've explained the impact of missing actors (nodes) and missing relationship information (ties) on the resultant network structure. In the ensuing example reproduced from them, N is the number of nodes or actors in the network, R is the relational response rate, M is the number of actors not responding on relationships with other nodes. The impact is evaluated in terms of both undirected (or non-directed) and directed networks. The relational response rate (R) for egocentric networks is calculated by dividing the number of reported ties by the total number of possible dyadic relations among the alters. For example, if the ego reported about 8 of the ten nondirected relations, then $R = 0.80$, or 80 percent; if the ego failed to report 6 of the 20 directed relations, then R = 0.70 70 percent.

Calculating the response rate for a complete social network is more complicated. A complete network consists of the dyadic relations among all pairs of *N* actors in the network. R is less attenuated for a non-directed network because a report by one member of a dyad suffices when the measure is reliable. For example, to measure friendship between actors A and B, the information provided by either informant could be used to determine whether that relationship is present or absent. That is unless both A's and B's reports about one another are missing, we measure their friendship with a single report. In general, for a complete nondirected network of *N* actors with no alter reports from *M* actors, the response rate for a particular relation is illustrated from the following example:

Let's assume that a network has five actors, labeled A, B, C, D, and E. The ten nondirected dyadic relations among these five actors are AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE. If actor A fails to report its relations, those dyadic ties can be obtained from the other four actors' reports about A. Thus, the relational response rate is 100 percent despite missing reports from one node. When the missing nodes range between 2 and 4 (1 <*M* <N), the relational response rate is $\left(1 - \frac{{}_{2}^{M}C}{{}_{2}^{N}C}\right)$ percent. For example, if three nodes (A, B, and C) do not report their relations with anyone, the response rate is $\left(1 - \frac{{}_{2}^{3}C}{{}_{2}^{5}C}\right) = 0.70$, or 70 percent. Three nondirected relations are missing (AB, AC, and BC), but at least one member reported the other seven dyads. If no actors provide information, both nodal and relational response rates fall to 0 percent

For the above example of a network of five actors, the nodal response rate and the relational response rate for varied numbers of missing nodes are given in Table 2.1:

Number of Missing Nodes	*Nodal Response Rate*	*Relational Response Rate*
0	100%	100%
1	80%	100%
2	60%	90%
3	40%	70%
4	20%	40%
5	0%	0%

Table 2.1 Table showing Nodal and Relational Response Rates corresponding to missing nodes.

The general formula of the impact of missing information on undirected networks is:

$$R\begin{cases} = 100 \text{ percent when } M = 0 \text{ or } M = 1 \\ = \left(1 - \frac{C_M^2}{C_N^2}\right) \times 100 \text{ percent when } 1 < M < N \\ = 0 \text{ percent when } M = N \end{cases}$$

It may be noted that the relational response rates for non-directed networks are always higher than the nodal response rates at every level of missing nodal reports.

For directed networks, where there is an asymmetry in ties (for example, giving advice, trust-based links, financial transactions), the calculations vary from the above in that missing nodes have a more significant impact on the relationship model. If node A is missing from the set, for instance, let's say, ties (AB, AC, AD, and AE) go unreported.
Let's suppose that node A is missing, and ties (AB, AC, AD, and AE) go unreported. Unlike ties in the undirected model explained above, which can be retrieved from the nodes which are not missing (i.e., B, C, D, and E), the same ties in a directed network will not be available as they are directed outwards from A and knowledge about them will only be available with A.

2.26.10 Dealing with Data Distortions

The current dissertation deals with an email-based network, and the ties are undirected. One of the advantages of an email-based network (or, for that matter, any electronic transaction-based network) over other types of social networks is that the cognitive and actual information will always match. There is less dependence on the memory recall or peculiarities of the actors needed to be interviewed or surveyed while building the network model. Even so, we will see in the chapters following that missing information has a significant impact on the ensuing network structure and, consequently, on the prediction of outcomes of the network. In the ENRON email-based corpus made available to the public, only 151 inboxes of employees are available for study out of the thousands of employees who might have worked there. These 151 inboxes have been selected for public viewing based on the investigators' perception of what is interesting and what isn't. With the non-availability of several inboxes, the employees' information with those mail-ids was also lost to the investigators.

Consequently, during the experiments conducted during this study, we shall see that the missing nodes greatly impacted the results. The predictions based on the initial lossy data also tend to be lossy. Please see Table 2.2 for details:

The ENRON dataset used in this study has 151 mail inboxes available in the public domain. From the mails headers contained in these inboxes, information regarding 6429 other employees was obtained. It needs to be noted here that there is no other source of information available regarding these employees. There may likely be even more employees whose mail-ids have not figured in the mails' mail headers in the 151 available inboxes. However, after applying the covertness index metric, which is the first stage of the experiment, the number of employees of interest to the study, i.e., those who were either indicted in the insider trading case or who were in some ways aware of the proceedings reduces from 19 to 16. All three losses occurred in the category of employees who did not have inboxes (reducing that number from 11 to 8), whereas the number of

employees whose inboxes were available remained at 8. In the category of employees who were of no interest to the study, i.e., those who had no role in the scam, the results are even more apparent: their number drops by 74% from 6418 to 1674.

Similar results are observed after the second phase of the experiment is carried out on the database. The number of employees of interest to the study, i.e., those who had either been indicted in the insider trading case or those who were in some way aware of the proceedings, reduces from 16 to 12 (the number of those with inboxes drops by three as does the number without inboxes). However, much more significantly, the number of employees of little interest to the investigation drops even more sharply from 1674 to 621. This is a welcome result since this dissertation's key objective is to define attributes that can quickly weed out actors who are less likely to require attention from investigators. However, it is significant that entities with incomplete information were lost. It indicates that entities with incomplete information are more prone to elimination than those with complete information.

	Employees of Interest		Employees not of Interest	
	Employees with mail inboxes	Employees without mail inboxes	Employees with mail inboxes	Employees without mail inboxes
At the start of the Experiment	8	11	143	6418
After the First Experiment (Application of the Covertness Index)	8 (Loss = 0)	8 (Loss = 3)	106 (Loss = 37)	1674 (Loss = 4744)
After the second part of the experiment (Application of the Collusion Index)	6 (Loss = 2)	6 (Loss = 5)	67 (Loss = 76)	621 (Loss = 5797)

Table2.2 Table showing losses of node related information while conducting the experiment. After the first part of the experiment the number of nodes of interest i.e. employees of ENRON who have had some role (indicted or aware of the details) decreases by 3 i.e. from 19 to 16. But the employees of interest who are not having mail inboxes have lost 3 whereas those who are having inboxes have lost none. The proportion of the loss is greater for nodes with incomplete or nil information. The results are on similar lines for those employees who are not of interest to the study. From this category, those employees who are having inboxes lose 37 or about 26% after the first part of the experiment whereas those who have no inboxes lose a whopping 4744 or 74%. Similarly, after the second part of the experiment, where the edges are clustered into pairs, in the resultant set of employees, those employees who are having mail inboxes available to begin with have smaller losses. The employees who are of interest to the study, i.e. who are in some way connected to the insider trading scandal are down from 8 after the first stage to 6, a loss of 2, whereas the number is down from 8 to 6 in case of those employees who had no inboxes at inception, a net loss of 5 from the start. Similarly, in the set of employees who are outside the scope on interest, those who are having inboxes decrease from 106 after the first part of the experiment to 67 i.e. an overall loss of 65 or about 53% from the overall figure (143) after the second part. In contrast, those employees who had no inboxes to begin with have decreased from 1674 to 621, a loss of 1013 and an overall loss of 5797 or a loss of about 90%. This indicates that entities with incomplete information in the network model tend to be filtered out during the process of scrutiny faster than those entities whose information is complete while building the network model.

2.26.11 Methods to Mitigate Missing Data

Researchers in social network analysis are perpetually searching for strategies and tools to counteract the harmful effects of missing and distorted data. The analysis of covert networks, however, presents conditions in complete contrast to those of conventional SNA. In the words of Kossinets (2008), who studied criminal networks:

> "While data collection quality in an analysis of conventional social relationships (such as 'friendship' or 'advice' networks) may be improved by appropriate research design and cooperation on the part of the participants, the situation in a criminal investigation is exacerbated by the unfortunate fact that criminals seldom cooperate with the law–enforcement agencies. Not infrequently, they engage in a conspiracy to conceal their identities and the structure of criminal organizations. Since investigators typically proceed by expanding the ego-networks of several main suspects, the key actors may be omitted due to ignored or unknown interaction contexts. Actors with false or multiple identities also (deliberately) introduce errors into the criminal group's structural representation. A plausible conjecture is that links may be easier to uncover once we know the primary suspects (via surveillance). However, since we expand the suspects' circle by traversing interactions in certain contexts, missing links are of great importance, too. As a result of the conspiracy, some meetings, telephone conversations, or email exchanges may not be recorded. The consequences are two-fold: first, investigators may be missing certain connections between actors in the main pool of suspects; second, since those connections lead to other potential suspects, truncated ties effectively hinder the course of the investigation. We interpret this type of missing data due to incriminating interaction contexts left outside the analysis scope,"

This is a universal feature of analysis when looking at building models of covert networks. Many investigators resort to approximations and informed guesswork to fill in the gaps. The results vary; if the guesses have been effective, the network model is more or less accurate. If not, erroneous outcomes ensue. There is always a need for the analyst to have

an ear to the ground in such cases and be in close touch with the domain specialists who can point to flaws during the collection of information at initial stages.

One of the more acknowledged methods of collecting data about covert or criminal networks is the one adopted by Krebs (2002). While gathering data in connection with the 9/11 attacks, Krebs realized how tenuous the sources of data were and was amazed at the sheer scale of wrong and misleading information pouring out into the public domain:

> "Once the names of the 19 hijackers were public, discovery about their background and ties seemed to accelerate. From two to six weeks after the event, it appeared that a new relationship or node was added to the network daily. In addition to tracking the newspapers mentioned, I started to search for the terrorists' names using the Google search engine 1. Although I would find information about each of the 19 hijackers, I rarely found information from the search engine that was not reported in the major newspapers I was tracking. Finding information that was not duplicated in one of the prominent newspapers made me suspicious. Several false stories appeared about a cell in Detroit. These stories, originally reported with great fanfare, were proven false within one week. This made me even more cautious about which sources I used to add a link or a node to the network."(p.45).

In his observations, one thing to note is the emphasis on trusting duplicated information in trusted domains like prominent newspapers. This technique approximates the concept of triangulation. His study ended up constructing the terror network in stages, reflected in the network diagrams in his work.

The first model (Figure 2, p.4) that resulted from his data-gathering efforts was based on the "trusted prior contacts" of the hijackers. The model was sparse, reflecting "how distant many of the hijackers on the same team were from each other. Many pairs of team members were beyond the horizon of observability." (p.46)Krebs concluded that "Keeping cell members distant from each other, and from other cells, minimizes damage to the network if a cell member is captured or otherwise compromised." Krebs quotes the mastermind, Usama(Osama) bin Laden, who is recorded having described this strategy on a videotape

that was found in a hastily deserted house in Afghanistan. "*Those who were trained to fly didn't know the others.One group of people did not know the other group."* (p. 46).

After the initial network model took shape, Kreb's next problem was to ascertain how exactly a covert network accomplishes its goals. He solved the problem by using the concept of "transitory short-cuts"[33]in the network. Such short cuts happen when meetings are undertaken to connect disparate parts of the network to coordinate tasks. After the coordination is accomplished, the key actors' cross-ties go dormant until the need for their activity arises again, presumably before the terrorist act is to happen. Krebs reports that "One well-documented meeting of the hijacker network took place in Las Vegas" (p. 10) and he factored the new ties resulting from this meeting and a few others into the construction of a second and larger network model which comprised "trusted prior contacts plus meeting ties (short-cuts)"(Figure 3, p.4).

Krebs's research revealed that the hijackers' network had a "hidden strength – massive redundancy through trusted prior contacts. Through kinship and training/fighting in Afghanistan, the ties forged in school made this network very resilient. These ties were solidly in place as the hijackers made their way to America. While in America, these strong ties were rarely active – used only for planning and coordination. In effect, these strong underlying ties were mostly invisible during their stay in the U.S. It was only after the tragic event that intelligence from Germany (about the Hamburg Cell) and other countries revealed the dense under-layer of this violent network." (p.50). Krebs added these dense "underlayers" to the second model and came up with the third and final model – "the hijackers' network neighborhood" (Figure 3, p.8) which was much denser.

[33] The concept of transitory short-cuts is similar to the concept of short paths proposed by Watts, (1999). In his words – "If the world did not contain many people, then it would not be surprising if they were all closely associated (as in a small town). If most people knew most other people then, once again, it would not be surprising to find that two strangers had an acquaintance in common. If the network were highly centralized—say a star—then an **obvious short path" (p.4)**

It needs to be pointed out that Krebs and many other similar researchers based their studies on the information available after the incident. Though some of this information appeared to be misleading initially, as time passed, the data became more "settled" and was corroborated through multiple sources. However, this doesn't settle the issue of building models of covert structures that lurk without their payloads being delivered. The detection of covert networks or subnetworks through prior or preventive surveillance is almost certainly tougher than post hoc analyses. However, many of the methodologies Krebs adopted are instructive. Looking at past ties in a network and ascertaining how they have evolved through time gives a very accurate window into the stages of a conspiracy's development. For instance, an investigator is smart to ask why a pair of actors who had been chatting incessantly on their cell phones a month ago have stopped conversing with each other all of a sudden and even though they remain in the vicinity of each other. Or, why two actors who were rarely communicating with each other have suddenly started exchanging information; such penetrating and close observation is likely to lead an analyst to a rich trove of hidden information.

Kossinets (2008, p.4) proposed a two-fold strategy to deal with missing data in criminal networks. The first approach develops analytic techniques that capture global statistical tendencies and do not depend on individual interactions. The second is a complementary strategy to develop remedial techniques that minimize the effect of missing data. To achieve this, he adopted a three-step iterative process:

> (1) take a real (large enough) social network or an ensemble of random graphs and assume that network data is complete;
>
> (2) remove a fraction of entities to simulate different sources of error; and
>
> (3) measure network properties and compare them to the "true" values (from the "complete" network).

He has quantified the uncertainty caused by missing network data. He assesses the sensitivity of graph-level metrics such as average vertex degree, clustering coefficient, degree correlation coefficient, size, and mean path length in the largest connected component. Although noting that his results may not be generalizable to all networks, he

concludes that “for forensic (criminological) research, it seems most important that the network of suspects is well-connected so that investigators can start from a few principal actors and “snowball” to the rest of suspects. As we have found that the size of the largest connected component is very sensitive to the omission of actors, an obvious recommendation would be to expand surveillance at the early stages in the investigation.” (p. 26)

2.27 Summary

Knoke’s recommendations are made with terrorist network modeling in mind. Still, the content holds equally well for any other covert or dark network where the main aim is to secrete away information from law enforcement's prying eyes. Indeed, the suggestions are suitable for any social network analysis. With the advent of artificial intelligence and machine learning algorithms, which are essentially computer programs that “learn” and “make inferences” from the training data that is fed into them, it becomes a sine qua non that computational research into social network modeling marches in tandem with domain-specific knowledge and lock-step with inputs from domain specialists. Collaborative relationships should help correct any bias that’s likely to creep in while predicting outcomes.

Although the study by Kossinets (2008) was “specific” to the dataset he studied (as is the case in the dissertation), the learning points are universally applicable across all networks: the errors and distortions that may creep in at the initial stages of network model building can have a profound impact on the analysis later on. There is no good alternative to accurate surveillance. These points need to be ingrained in any covert networks analysis where missing (and misleading) information is the norm and not the exception.

Chapter 3

Design and Application of a Covertness Metric

3.1 Introduction

The previous chapter highlights the challenges of applying traditional social network analysis techniques to covert networks. The contents also show up the substantial differences between traditional social networks and covert networks. One way to tackle this problem is to make out a very case-specific scenario and identify attributes peculiar to the covert network at hand. These attributes can then be metricized in some manner and used for measuring the outcomes. This approach is useful only if the concerned law enforcement (or oversight department in an organization in case the idea is to keep tabs on intra-corporate conspiracies) has many resources. It has a good amount of prior knowledge of what is being scrutinized. All too often, that is not the case, and the surveillance agencies are working against unknown or hidden adversaries in a timeframe that is marked with unknown deadlines. To avoid these pitfalls in conventional covert network analysis, this study proposes a universal metric for measuring all social networks' covertness potential, not just ones that are deemed covert. Covertness as a centrality measure has been highlighted in a few studies before this (Ovelgonne et al. 2012), and the idea has been further exploited here.

3.2 Defining an Edge-Vertex Function

In a typical email network, we have nodes that have previously sent emails to each other or sent and received copies of others' emails. Thus, a node in such a network will have a certain number of emails received and a certain number of emails sent, the total of which we may define as its incidence. A certain portion may be received or sent as copies of emails exchanged between two other nodes of these emails. We may start our study from

the basic sub-network structure in any network; namely, the relationship between two nodes also called a tie or an edge. Both terms are used interchangeably for this study's purposes. In an email network, an edge comprises a pair of nodes which have exchanged emails between them. Each mail so exchanged is termed a *relationship*. For simplicity, as noted earlier, the tie's direction is not considered here.

3.3 Edge as the Basic Network Unit

One of this study's unique contributions is its focus on the edge or the tie between nodes rather than the nodes themselves. I have observed that a disproportionate corpus of studies has revolved around identifying nodes (or players) in networks of interest, whether covert or otherwise. This approach runs the risk of identifying a particular node as wholly covert. In contrast, in real-life situations, it's often observed that the same node may function in a completely overt or transparent manner in many of its dealings with other nodes in the same social network, yet transact with others in opaque covert ways. The same behavioral trend may also be longitudinally (time-dependent) visible as well. For example, many ENRON employees who played pivotal roles in the conspiracy leading to the scam were also functioning in a completely open fashion with many other employees. A quick scrutiny of the emails reveals that even with some of the other employees who eventually became their co-conspirators, older mail exchanges have practically nothing to do with any label that may be tagged with their actions leading to the scam. Thus, it is a key inference that while studying covertness, it's important to keep the context in mind. In other words, what are the prevailing situations and surroundings in which the actions of any node (or player or employee, as in this case) are being judged?

This study treats an edge as a relationship (or a relational tie) between a pair of nodes, referred to as a *dyad*. Wasserman & Faust (1994, p. 18) define dyads as subgraphs of size two consisting of a pair of actors and all ties between them. They define a tie as a linkage or relationship between two actors (nodes) at the basic level. The tie is inherently a property between the dyad's nodes and therefore is not thought of as pertaining simply to

an individual actor. Robins (2009) expands the definition of a dyad further to include information on what he has named "dyadic level factors" and defines a dyad as a "pair of actors and the social relationships between them" (which may at times be non-existent if visible links do not link any two nodes). He mentions that social networks are typically measured at the dyadic level regarding relationships of certain types observed between actors' pairs. Accordingly, dyadic level factors may also be "embodied in the design and method of data collection of a network study." The dyad or the edge-based properties occur typically at tier-2 of the pyramid defined by Robins (2009)[34];

A focus on edges or ties as the basis of observation in a social network has its echoes in previous studies, though not in any way related to its usage in this work. Most methods of social network analysis look at edges exclusively as a relationship attribute between nodes. These approaches consider the node the fundamental unit of analysis and the edge as a means to relate one to the other. But this doesn't mean that edges have never been the center of attention in social network analysis. Wasserman & Faust (1994, p18) focused on dyadic analyses and on the properties of pairwise relationships, such as whether ties are reciprocated or not. Robins (2009) also examined the apparent contradiction between traditional node-based approaches to social network analysis and one that pays heed to relationships. A similar approach is seen in the concept of *edge-betweenness* developed by Girvan and Newman (2004). The betweenness property of an edge forms the center-piece of the proposed algorithm for community detection.

Robins (2009), in his study, critiqued a pure graph-theoretic model built around node based attributes and approaches focused only on node-centric approaches that he has found to be inadequate. He argues that graph-theoretic approaches say nothing about the actors in a network, except that they express and receive ties. Suppose the network topology is considered as a complete representation of the network. In that case, the implication is that any particular qualities of the actors, i.e., the individuals within the network, are irrelevant. Robins (2009) further states that "extreme claims" of interpretation are indeed untenable

[34]Please see the pyramid figure in the earlier section on analytics.

for human social systems, and more crucially, for criminal network studies like the one in focus here.

3.4 Types of Relationships

Wasserman and Faust (1994) expanded upon the definition of a dyadic entity, thus:

> "A dyad is an unordered pair of actors (nodes) and the arcs (ties) that exist between the two actors in the pair. The dyad consisting of actors (nodes)*i* and *j* will be denoted by $D_{ij} = (X_{ij}, X_{ji})$, for $i \neq j$. Dyads are defined for unordered pairs, where the first actor index is less than the second, so that $i < j$. Every pair of actors is then just considered once. There are exactly $({}^{g}C_{2}) = g(g-1)/2$ dyads. However, there are $g(g-1)$ ordered pairs of actors. (p.510)"

Here, the term ordered pair means a directed graph, and the direction of the information flow is also a factor in identifying a dyadic tie. For example, in an undirected graph, a tie between nodes i and j is depicted as e_{ij} and $\forall (i,j) \in E$, and $i \neq j$ E is the set of all edges in graph G, $e_{ij} = e_{ji,}$ i.e., the ties in undirected graphs are reflexive. In a directed graph, however, $e_{ij} \neq e_{ji,}$ i.e., ties are directional and non-reflexive.

Wasserman and Faust (1994) elaborate upon the types or states of relationships between the nodes belonging to a dyad and define the following three types:

1. **M: Mutual relationships**
2. **A: Asymmetric dyads (2 types)**
3. **N: Null dyad**

Wasserman and Faust (1994)[35] define a mutual relationship between node *i* and node *j* as one that exists when $i \rightarrow j$ *and* $j \rightarrow I$ in the dyad, and a mutual relationship is apparent in a

[35]The notation followed in this section broadly follows the notation pattern adopted by Wasserman & Faust (1994).

sociomatrix[36] when both the *(i, j)* and *(j, i)* cells (located symmetrically about the diagonal of *X*) are unity; that is, X_{ij}= 1 and X_{ji}= 1 so that the dyad D_{ij}= (1,1).

The second state defined by them is the asymmetric dyad, which can occur in two ways. Either *i →j or j →i*, but not both ; specifically, *Dij* = (1,0) or (0,1). Thus, in the Adjacency Matrix (sociomatrix), only one of X_{ij}or X_{ji}, which are symmetrically located off the diagonal, will contain a 1. This type of relationship is asymmetric since the relationship is not reciprocated.

The third type of dyad classified by Wasserman and Faust (1994) is the *null dyad*, in which neither of the pair of actors (nodes) has a relationship (or tie) with the other. By default, a dyad that is not asymmetric or mutual must he null. Thus, the cells in the adjacency matrix (sociogram) which correspond to*(i,j)* and *(j, i)*;symmetrically placed within the matrix, are both 0; that is, X_{ij}= X_{ji}= 0, thereby implying that D_{ij}= (0,0)

Figure 3.1 illustrates all four relationship types. Null dyads are important for research and analysis purposes in covert networks where null edges may often signal missing information. It's required not to discard null dyads and explore a potential link through predictive methods.

[36]A tabular representation in matrix form of data collected using a sociometric method to measure interpersonal relationships. Used by Social Scientists and is identical to the term Adjacency Matrix (used by Computer Scientists).

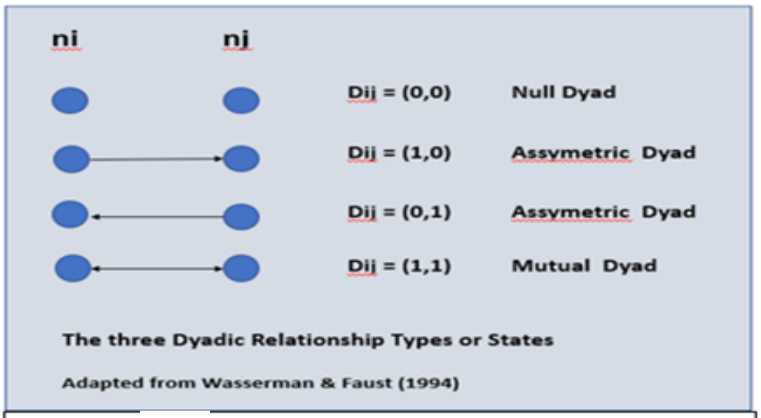


Figure **3.1** Types of ties described by Wasserman & Faust (1994). It is interesting to note that even though there may be apparently no link between two nodes in a dyadic pair, as is seen in the first instance, the pair may still be valuable in a study of covert networks. This is so since there may be pre-existing ties between the two or information exchange over covert channels which are invisible but can be later inferred from a range of analytic and predictive methods. The two asymmetric ties shown next are representative of directed networks and the last one which is described as symmetric is representative of undirected networks.

The dyadic properties indicated above are the center-point of developing an Index of Covertness in this study. A d*yadic tie*[37] is typically the simplest form of social network with only two nodes joined by an *edge* or relationship or tie. A dyadic tie can be directed or undirected. Simply defined, a *dyad* is a pair of actors (nodes) and all the ties between them, and can be in one of three states: *null* (no ties), *asymmetric* (one tie; two possibilities), and *mutual* (two ties)(Wasserman & Pattison, 1996). Why these concepts have been discussed in some detail here becomes apparent when we come to the second half of the approach in this paper, where we seek to *enjoin* pairs of *edges* formed between the *constituent nodes* in a dyad. We'll see that such *edge-pairs* may either be 1 or 0, depending on whether linkage is possible or not. There is no further use of *asymmetric dyads*, as all ties have been considered subsequently in this paper as *undirected*.

[37]For the purposes of this study the term *DyadicTie* is used interchangeably as an *edge* or a *dyad* without any loss of the meaning of this term.

Any social network can be decomposed into a set of dyads or edges. For this study, an *edge* is defined as a tie or relationship that relates two nodes without indicating the direction of the information (i.e., the relationship is undirected). The study assumes that all pairs of nodes in a social network have edges, even those that haven't exchanged any information and can be considered null or absent. This approach allows the construction of the basic substratum upon which we will construct a tangible metric or index for measuring the covertness inherent to the dyads constituting the network; this entity we call an *Edge Vertex,* which will be defined and explained in detail in the coming sections.

3.5 Relationship Set

Let's describe a set *R* comprising all instances of relationships (or emails exchanged) between two network nodes. A *Relationship Set* between nodes ***i*** and ***j*** should read as R_{ij}. Further, each instance of a relationship between nodes ***i*** &***j*** may be read as $R_{ij(m)}$. For example, the first communication between ***i*** &***j*** is read as $R_{ij(1)}$. The second is $R_{ij(2)}$, and so on. Thus, for the *edge* ***i-j***, R_{ij}, there is a set of ***m*** elements **{** $R_{ij(1)}$, $R_{ij(2)}$, $R_{ij(3)}$,..... $R_{ij(m)}$**}** where ***m*** is the total number of instances of communication (emails) between nodes ***i*** &***j***. Each instance has attributed a value of 1 (Values may vary if weights are allotted to each instance depending on its importance, but again for this study's purposes, each instance's weight is universally considered 1 for simplicity). Of the above, some might be copied (or intimated) to other nodes. Suppose node ***a*** is made aware of one such communication, say, the second instance between nodes ***i*** &***j***. The communications will be marked as a binary flow of information between the edge formed between dyad **(*i, j*)**& node ***a*** with 1.

Definition*(**Relationship Set**)Consider a finite set of nodes,* $\boldsymbol{V = \{V_1, V_2, \ldots..., V_i, V_j, \ldots. V_n\}}$ *of n entities constituting a social network. Consider the finite set of all communications exchanged between nodes* $\boldsymbol{V_i}$ *and* $\boldsymbol{V_j}$, $\boldsymbol{R_{ij} = \{R_{ij(1)}, R_{ij(2)}, \ldots., R_{ij(m)}\}}$ *comprising m entities. Then we call* $\boldsymbol{R_{ij}}$ *as the* ***Relationship Set*** *between Nodes* $\boldsymbol{V_i}$ *and* $\boldsymbol{V_j}$ *of the social network, and* $\boldsymbol{r_{ij} = |R_{ij}| = m}$ *is the cardinality of set* $\boldsymbol{R_{ij}}$.

3.6 Shared Relationship Set

I now describe a set R_{ij}' where R_{ij}' is the set of all communication instances between ***i*** &***j,*** which have been associated with at least one node, neither ***i*** nor ***j***.

Thus, R_{ij}' is the set comprising emails that have been copied to at least one node other than nodes ***i*** or ***j***.

For example,

If the first mail exchanged between nodes ***i*** and ***j,*** i.e.$R_{ij(1)}$, is copied to nodes ***a*** and ***k*** and the third instance of communication, $R_{ij(3)}$ is copied to nodes ***c*** & ***d*** then, we consider $R_{ij(1)}' = R_{ij(1)}$,

$R_{ij(2)}' = R_{ij(3)}$etc. and thus,

$R_{ij}' = \{R_{ij(1)}', R_{ij(2)}', \ldots\}$.

Further, let,

r_{ij}'= Number of instances of communications between *i* &*j*, which are known to other nodes, i.e., the number of elements in $R_{ij,}$'i.e. r_{ij}' = Number of elements in set $R_{ij,}'$ i.e., its cardinality.

Definition*(Shared Relationship Set)* *Consider the finite set of all communications exchanged between nodes* V_i *and* V_j, $R_{ij} = \{R_{ij(1)}, R_{ij(2)}, \ldots, R_{ij(m)}\}$ *comprising m entities and consider a finite set* $R_{ij}' = \{R_{ij}'_{(1)}, R_{ij}'_{(2)}, \ldots R_{ij}'_{(k)}\}$ *comprising k entities where* $k \leq m$, *constituting all such mail communications between nodes* V_i *and* V_j *which have been copied to nodes other than* V_i*and* V_j. *Then we call the set* R_{ij}' *as the* ***Shared Relationship Set*** *of nodes* V_i *and* V_j *of the social network, and* $r_{ij}' = |R_{ij}'| = k$ *is the cardinality of set* R_{ij}'.

A *Relationship Set* and a *Shared Relationship Set* can be explained with a simple illustrative example. In Figure 3.2, we have a basic social network consisting of four nodes; ***a, b, c,*** and ***d***. In the example given below, mail exchanges occur between two nodes, namely, ***b*** and ***c***. In the first two instances of email exchanges, no copies are marked.

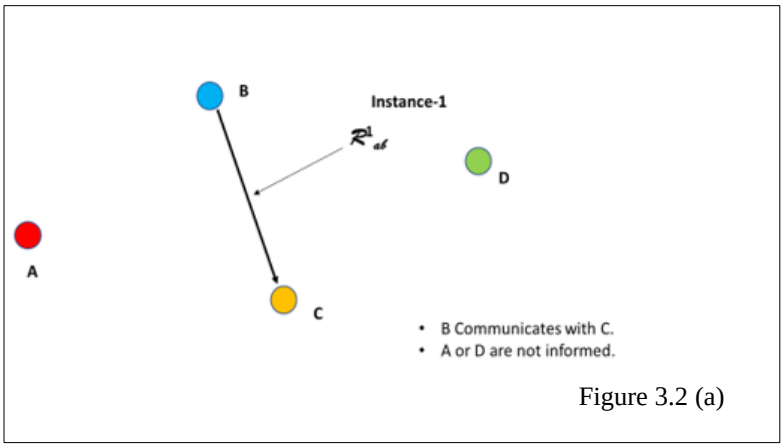


Figure 3.2 (a)

In Figure 3.2(a) above, node ***b*** sends a mail to node ***c***, and this instance is not copied to the other two nodes in the above network.

As per the definitions above, the *Relationship Set* between nodes ***b*** and ***c*** is $R_{bc} = \{R_{bc(1)}\}$ after the first instance of mail exchange.

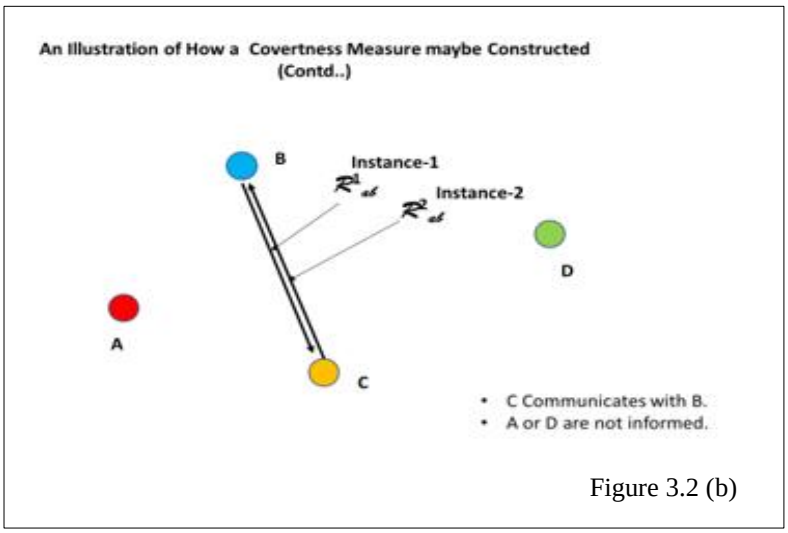


Figure 3.2 (b)

In Figure 3.2(b), node ***b*** sends a mail to ***c***, which is the second instance of a communication exchange between the two nodes, and this instance is also not marked to any of the other two nodes in the network.

Going by the definitions above, the *Relationship Set* between nodes ***b*** and ***c*** becomes to $R_{bc}= \{R_{bc(1)}, R_{bc(2)}\}$ after the second instance of mail exchange.

The third instance of mail exchange happens between nodes ***b*** and ***c***, but this time there is a copy marked to ***c***, shown in Figure 3.2 (c).

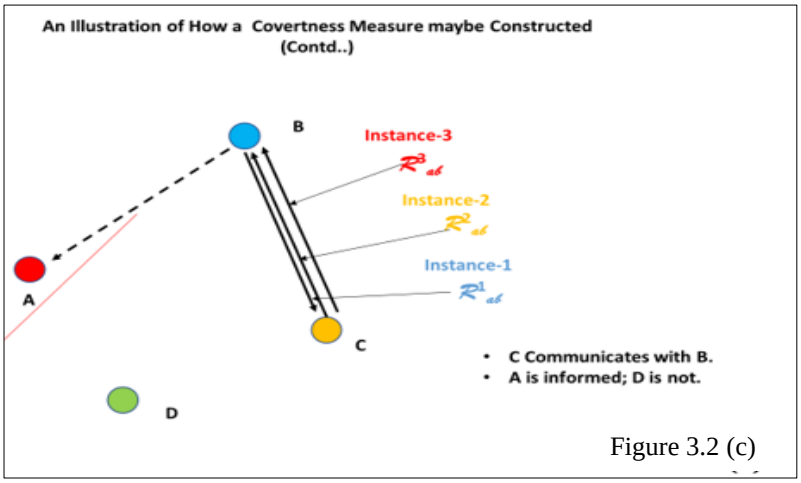


Figure 3.2 (c)

Thus, after the third instance of mail exchange, the *Relationship Set* between nodes ***b*** and ***c*** becomes to $R_{bc}= \{R_{bc(1)}, R_{bc(2)}, R_{bc(3)}\}$. More importantly, since there is an instance of a copied mail, the *Shared Relationship Set* starts to be populated; $R_{bc}' = \{R_{bc(1)}'\}$.

This routine gets iterated each time a mail is exchanged between the two nodes, and both the sets, namely, the *Relationship Set* and the *Shared Relationship Set,* increase in size.

3.7 Neighborhood Relationship Set

We now come to the question of the sets of nodes that received copies of some of the email instances exchanged between the constituent nodes of a dyad, which has already been touched upon while defining the Relationship Set and the Shared Relationship Set. Neither of these entities contains any references about the identities of the nodes which received the copies. This information will be of crucial importance to the second part of this study.

We attempt to find links between dyads based on their common intentions (meaning described later in the paper).

Definition*(Neighborhood Relationship Set)* *Consider the finite set of all communications exchanged between nodes* V_i *and* V_j, $R_{ij} = \{R_{ij(1)}, R_{ij(2)}, \ldots, R_{ij(m)}\}$ *comprising m entities and consider a finite set of finite sets* $N_{ij} = \{\{N_{ij}\}_{(1)}, \{N_{ij}\}_{(2)}, \ldots \{N_{ij}\}_{(m)}\}$ *comprising* $\mathbf{m}$ *entities comprising all mail communications between nodes* V_i *and* V_j*whether or not these emails have been copied to other nodes. If there is an instance* $\mathbf{k}$ *of mail exchange between nodes* V_i *and* V_j*, which is not copied out to other nodes, then* $\{N_{ij}\}_{(k)} = \{\varphi\}$*. Then we call the set* N_{ij} *the* ***Neighborhood Relationship Set*** *of the edge vector* $(E_v)_{is}$ *formed by the pair of nodes* V_i *and* V_j *of the social network and* $|N_{ij}| = |R_{ij}| = m$ *the cardinality of set* N_{ij}*.*

For the sake of notational uniformity, we may represent the finite set N_{ij} as $\Gamma(ij)$subsequently. The symbol Γ represents the dyad's neighborhood in terms of nodes that have received copies of mail instances exchanged between the pair of nodes. Just what a *neighborhood set* consisting of email copy-receiving nodes looks like is shown in Figures 3.3(a) and Figure 3.3(b).

Figure 3.3(a) depicts a fictional email-based social network where the solid lines between the nodes represent mail exchanges and the dotted lines joining nodes represent mail instances. In Figure 3.3(b), the mail exchanges (i.e., the solid lines) have been removed, leaving behind only the dotted lines representing the copies sent out by the pair of nodes (a,b). As can be seen, nodes c,d, and g are the ones with dotted lines joining them to the dyad (a,b), i.e., these are the nodes that have received copies of instances from (a,b). Thus, nodes c, d & g constitute the dyad (a,b). However, recall that the Neighborhood Relationship Set of any dyad (or edge) is a set of sets. Each element of this set is a set that is constituted by the identities of nodes that have received copies of a particular instance of mail exchanged between a and b, i.e., the constituent nodes of the dyad. Hence, nodes c, d, and g in Figures 3.3(a) and 3.3(b) *may belong to different sets* within the overall Neighborhood Relationship Set.

This may be true because these nodes may have received different copies of emails exchanged between ***a*** and ***b***. If one of the instances was marked out to say ***c*** and ***d*** but not ***g***, the set that would be formed will be (***c,d***), and if the copy of another instance of mail is marked out to nodes ***c*** and ***g*** but not ***d***, the set that is formed will be (***c, g***) and so on. Each such set will be a subset of the overall Neighborhood Relationship Set that would be created after considering all the exchanges of emails that have taken place between nodes ***a*** and ***b***. The number of sets within the overall Neighborhood Relationship Set will equal the number of instances of mails exchanged between nodes ***a*** and ***b*** of the dyad (***a,b***), which have been marked out as copies to other nodes within the network.

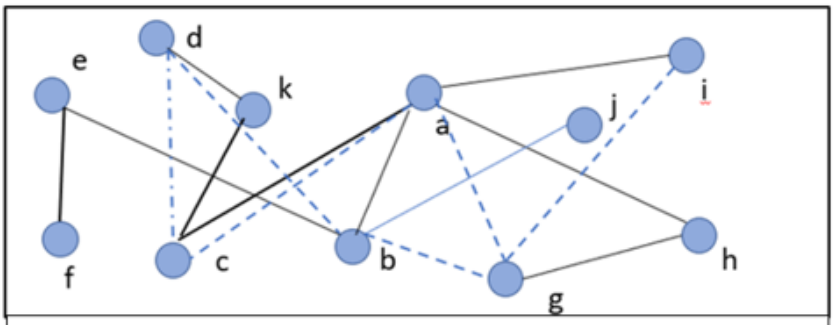


Fig. A mail based social network. The solid lines connecting the nodes represent mail links i.e. the nodes linked together have sent or received mails from each other at least once. The dotted lines represent copies sent by the nodes to nodes outside the dyad pair.

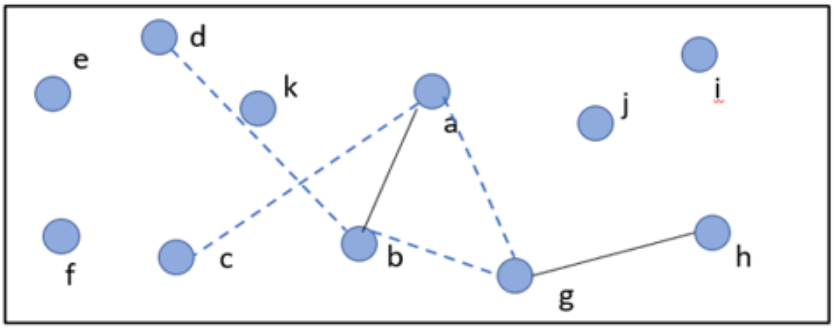


Fig. The solid lines representing the mail links are removed from this figure. The dotted lines which represent copies sent by the node pair (***a, b***) to nodes outside the dyad pair have been retained. The nodes which have links with (***a, b***) represent the *Neighbourhood Shared Relationship Set* ***Γ(ab)*** of the Dyad pair (***a, b***) the dyad pair. Thus, the nodes ***c, d*** and ***g*** belong to this set and can be described to populate ***Γ(ab)***.

Figures 3.3(a) and 3.3(b)

3.8 Edge-Vertex Function Explained

Now that Relationship Set, Shared Relationship Set, and Neighborhood Relationship Set have been defined, we can define an Edge-Vertex function. A brief description of how the concept Edge-Vertex works were given in an earlier section. However, before defining Edge-Vertex, the concept is explained below with illustrative examples and Figures 3.4 (a), 3.4 (b), and 3.4(c):

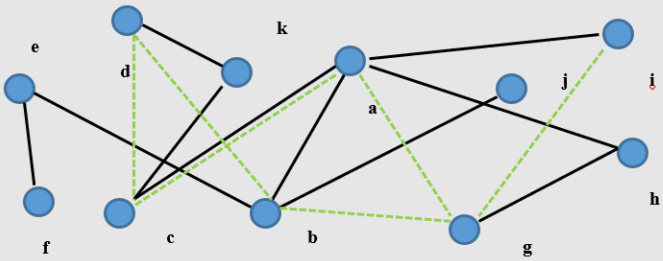
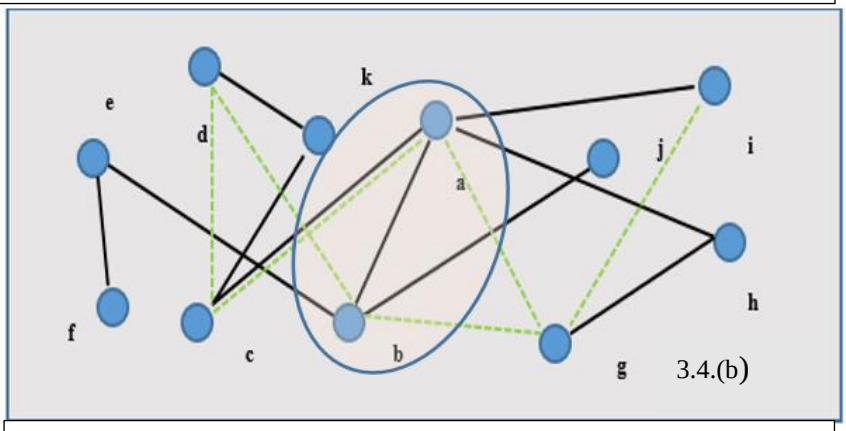


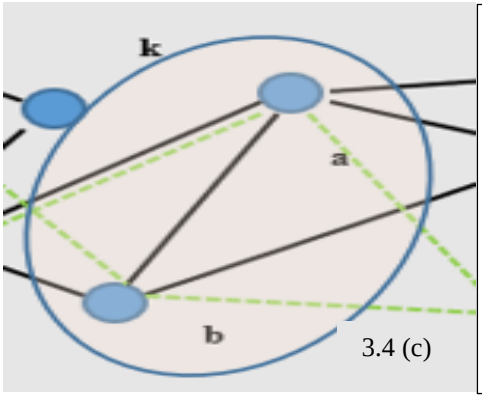
Figure 3.4.(a)

Fig. ?? : The above Social Network is an example of an e-mail exchange community. Each mail-id which is the equivalent of a node in a network is named as an alphabet. The solid lines that link the nodes represent actual exchanges of mail. The dotted lines in green indicate mails copied from one node to the other. The links or edges are not shown as directionally oriented since the study looks at undirected links. That is, it doesn't really matter who has mailed whom.



3.4.(b)

Fig. ?? : The Dyad of interest i.e. the nodes ***a*** and ***b*** have been highlighted in the figure above. There is an Edge E_{ab} between the nodes defining the mail exchange relationship or tie between them. The dotted lines emanating from the pair of nodes are symbolic of the mails which have been copied out from the overall mail exchanges between nodes 'a' and 'b'.



3.4 (c)

The Dyad of interest i.e. the Nodes 'a' and 'b' have been shown in isolation in the figure above. The Edge Eab between nodes 'a' and 'b' has 4 dotted lines emanating out of it which reflects that there are 4 mails copied out.

Figure 3.4 (a) shows that the social network shown is an example of a representative community based on e-mail exchanges. Each mail-id, equivalent to a node in a network, is assigned an alphabet used as its identity. The solid lines that link any pair of nodes represent actual exchanges of mail between them. The dotted green lines between nodes indicate mails copied from one node to the other. The links or edges are not shown as directionally oriented since the study is based on undirected links. That is, it doesn't matter who has mailed whom.

From the figures above, it's clear that if we are to ascribe an index or a value based on the tie between the pair of nodes comprising the dyad, the use of a measure akin to the degree centrality, i.e., the number of edges connected to nodes ***a*** and ***b*** is the most preferable and simple one amongst all the centrality measures enumerated earlier. The measure is adapted to developing a tie based metric and hence is deployed in a modified format that is made clear in the analysis that follows. The way degree centrality is modified is explained below based on the figures shown above.

Figures 3.4(a), 3.4(b), and 3.4(c) shows that node ***a*** has exchanged emails with ***c***, ***b,*** and ***h***; thus, its degree centrality is 3. Similarly, ***b*** is seen to have exchanged emails with ***e*** and ***a,*** and thus, its degree centrality is 2. It needs to be pointed out here that the degree centrality measure does not indicate the total number of emails exchanged by these nodes. It merely reflects the total number of tangible links that a particular node has with other nodes via mail exchanges. For instance, let's suppose that nodes ***a*** and ***c*** have exchanged ten emails, nodes ***a*** and ***b*** have exchanged 20 emails, and nodes ***a*** and ***h*** have exchanged 15 emails. Then, node ***a*** is a part of 45 mail exchanges, which is not reflected by its degree centrality of 3. What is proposed here is to use the volume of mail associated with a particular node (or, more correctly, the volume of mail between a pair of nodes) rather than the node's degree centrality attribute? Thus, the metric used here is a variation of the degree centrality, in that it is calculated around the edge between a pair of nodes rather than a single node.

Thus, the edge between nodes ***a*** and ***b,*** i.e., $e_{ab,}$ belongs to the set ***E*** that consists of all the edges in the mail based social network. The edge e_{ab} between nodes ***a*** and ***b*** is essentially

the mailing link between both these nodes. The edge itself represents the individual instances of mail exchanged, which are the elements that constitute the Relationship Set R_{ab} between the nodes (this set was defined earlier). To obtain the Edge Vertex, the number of emails that have been copied to nodes outside of the nodes constituting the dyad is recorded. Then the ratio between the two, i.e., the ratio between the number of mails copied out by the dyad to the number of emails exchanged between the constituent nodes, becomes an index of the covertness of the tie that links the pair of nodes. (We may also directly calculate and record the fraction obtained from deducting this Overtness Index from 1, which gives us the Covertness Index of the edge.) Finally, the list of all sets of nodes that have received copies of mails is also recorded in a separate Shared Relationship Set.

The Edge-Vertex is thus a list that contains the following five elements connected to a dyad that it's defined on:

-Node Identifiers of the nodes constituting the dyad.

-The cardinality of the Relationship Set between the nodes.

-The cardinality of the Shared Relationship Set between the nodes.

-The Covertness Index of the tie between the constituent nodes.

-The Neighborhood Relationship Set between the nodes.

Given the above facts, we may now define the Edge Vertex as follows:

Definition ***(Edge-Vertex)*** *Consider a finite set of nodes,* $\boldsymbol{V = \{V_1, V_2, \ldots..., V_i, V_j, \ldots. V_n\}}$ *of n entities constituting a social network. Consider further, any pair of nodes* $\boldsymbol{(i,j)}$ *from the set* $\boldsymbol{V}$*. Consider also, the finite Relationship Set* $\boldsymbol{R_{ij}}$ *consisting of the instances of communication between the pair of nodes selected and the Shared Relationship Set* $\boldsymbol{R'_{ij}}$ *between the selected nodes and the Covertness Index* $\boldsymbol{C_{ij}}$ *of the tie between the nodes. Consider finally that the finite Neighborhood Relationship Set* $\boldsymbol{N_{ij}}$ *consists of nodes that have received copies of mail communications from the selected pair of nodes. The Edge-Vertex function* $\boldsymbol{(E_v)_{ij}}$*, when defined on the edge eij between nodes i and j, outputs the list* $\boldsymbol{(E_v)_{ij} = \{V_i, V_j, |R_{ij}|, |R'_{ij}|, C_{ij}, N_{ij}\}}$.

The above definition superficially appears to say the Edge-Vertex function is identical to *E*, the set of all edges in graph *G*, but it isn't. The set *E* comprises all possible ties in a network between distinct nodes. In contrast, the Edge-Vertex between nodes ***i*** and ***j*** is a list-set and consists only of a composite group of values concerning the nodes' identity, a covertness index, and their Relationship Sets' cardinalities and Shared Relationship Sets along with the Neighborhood Relationship Set. There is no single value assigned to the Edge-Vertex as such, and its purpose is to store an assorted set of values and information about a pair of nodes, as illustrated in Figure 3.5:

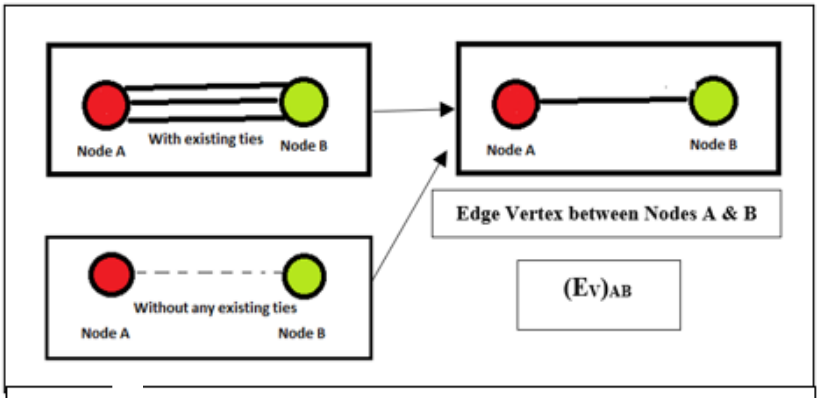


Figure 3.5 Illustration of how an Edge-Vector function is computed.

Figure 3.5 presents two cases. In the first case, there are some communication exchanges between nodes *A* and *B*, and in the second, there are no exchanges between the nodes. But, in both cases, there will be a resulting edge $\boldsymbol{e_{AB}}$. But, the Edge-Vertex values for both cases will vary significantly because the Edge-Vertex function is designed to store values emerging from each mail transaction between a given pair of nodes, and the list-set that is obtained by applying this function gets changed incrementally. The reason why the Edge-Vertex function is proposed in this study is made clear once we come to the part that measures collusion between covert edges. The information supplied by the function becomes very crucial then.

A more detailed analysis of how the set constituting an Edge-Vertex is constructed is illustrated in Figures 3.6 (a) to 3.6 (d), representing a small e-mail based network.

Figure 3.6 (a) shows the exchange of emails between two nodes, ***a*** and ***b,*** and copies of some of these emails sent out to other network nodes. Each mail exchanged between nodes ***a*** and ***b*** is marked with a solid line of a particular color if the same instance of mail exchange has been marked out as copies to one or more nodes in the same network (outside of the pair of nodes in question, i.e., ***a*** and ***b***) the copy is denoted by a dotted line with the same color as the instance of the mail exchange.

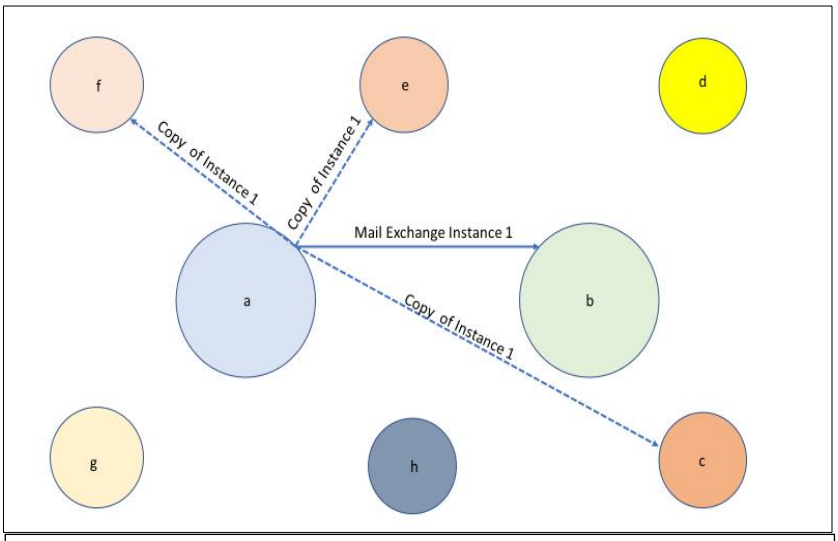


Figure 3.6 (a)

In the figure, node ***a*** sends a mail to ***b***, labeled as Instance 1 and colored blue. Three copies of the mail instance are marked as copies by ***a*** to nodes ***c***, ***e,*** and ***f***.

We are constructing the set on which Edge-Vertex $(E_v)_{ab}$ is based.

After the first instance of mail exchange between nodes ***a*** and ***b***,

$(E_v)_{ab} = \{Va, Vb, |R_{ab}|, |R_{ab}'|, C_{ab}, , \{(V_c, V_e, V_f)\}\}$

where**,**

V_a is node ***a,*** V_b is node ***b,*** R_{ab} is the **Relationship Set** between nodes ***a*** and ***b*** as defined earlier**,** R_{ab}' is the **Shared Relationship Set** between nodes ***a*** and ***b,*** C_{ab} is the **Covertness Index** of the tie between nodes ***a*** and ***b*** after this instance of an exchange of mail and the last entity in the list is the set of sets of nodes to which the particular instance of mail was marked out.

After the first instance of an exchange of mail between nodes ***a*** and ***b***, the **Relationship Set cardinality** between ***a*** and ***b*** is **1**. Since there is only a single mail, the **Shared Relationship Set cardinality** is also one as copies have been marked. It needs to be noted here that though three copies of the first instance are seen marked out from the mail exchanged between the dyad(***a,b***) (i.e., to nodes ***c***, ***e,*** and ***f***'), the **cardinality** will remain **1**. (However, the fact that three nodes have received copies of the instance of mail exchange is reflected in the last entity of the Edge-Vertex list). The entity that occurs after the cardinality entries is the **Covertness Index,** which is obtained by computing the ratio of the cardinalities of the Shared Relationship Set to that of the Relationship Set formed between nodes ***a*** and ***b*** deducted from 1.

After the first instance of mail transfer, the **Covertness Index**of edge (***a,b***) is **0,** which is found through the following calculation:

$|R_{ab}| = 1,\ |R_{ab}'| = 1,$

Overtness Index $= |R_{ab}'| \div |R_{ab}| = 1$

C_{ab} **= Covertness Index = (1 – Overtness Index) = (1 – 1) = 0;**

The last entity in the list set comprising the Edge-Vertex for edge (***a,b***) is the set of sets containing details of the nodes to which the copies of a particular instance of mail exchange are marked. Since we can see here that node ***a,*** which has sent the mail to ***b*** in this instance, it has marked out copies to nodes c, e, and f, this entity reads as $\{V_c, V_e, V_f\}$**.**

We can summarizetheoutputs after the first instance of mail exchange as follows:

1. Relationship Set**(**R_{ab}**)**:
   a) $R_{ab(1)} = 1$;
   b) R_{bc}**={1}**;
   c) $|R_{ab}| = 1$;
2. Shared Relation Set (R_{ab}'):

a) $R_{ab(1)}' = 1$;

b) $R_{ab}' = \{1\}$;

c) $|R_{ab}'| = 1$;

3. Covertness Index (C_{ab}) = 0:

4. Set of Copied Nodes = $\{\{V_c, V_e, V_f\}\}$;

5. $(E_v)_{ab} = \{V_a, V_b, |R_{ab}|, |R_{ab}'|, C_{ab}, , \{(V_c, V_e, V_f)\}$;

i.e. $(E_v)_{ab} = \{V_a, V_b, 1, 1, 0 , \{(V_c, V_e, V_f)\}\}$;

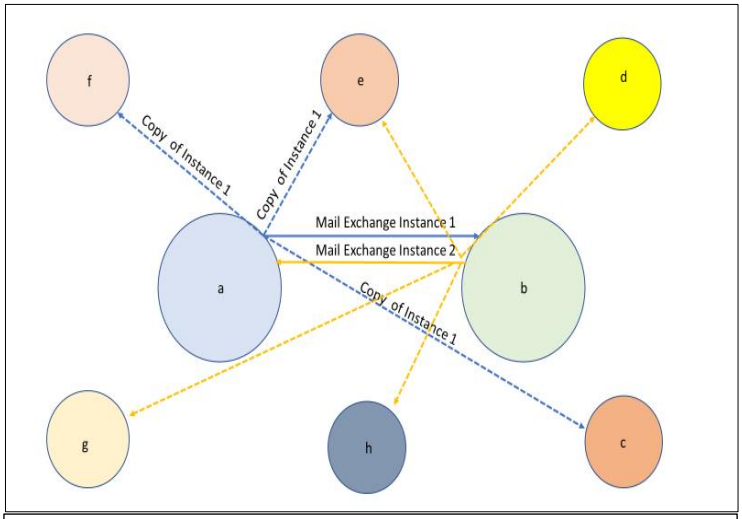


Figure 3.6(b)

Figure3.6 (b) above shows the second instance of mail being exchanged between nodes ***a*** and ***b***; it's node b that has sent a mail to node ***a***. The arrows representing this second instance are yellow. The mail exchange between ***a*** and ***b*** is shown in a solid yellow line, whereas the mails marked by node ***b*** to other nodes, namely, nodes ***d***, ***e***, ***g,*** and ***h,*** are shown in dotted format.

As in the first instance of mail exchange, we go on with constructing the set on which edge vertex $(E_v)_{ab}$ is based.

After the second instance of mail exchange between nodes ***a*** and ***b***, we have:

$(E_v)_{ab} = \{V_a, V_b, |R_{ab}|, |R_{ab}'|, C_{ab}, , \{(V_c, V_e, V_f), (V_e, V_d, V_g, V_h\}\}$

After the second instance of an exchange of the mail between nodes ***a*** and ***b***, the **cardinality** of the **Relationship Set** formed between ***a*** and ***b*** is **2**; since there are two mails between ***a*** and ***b*** at this stage, the **cardinality** of the **Shared Relationship Set** also increases to**2** as copies of emails have been marked out the second time around as well. The **CovertnessIndex**, as in the earlier case, gets computed by calculating the ratio of the cardinalities of the Shared Relationship Set and that of the Relationship Set, respectively, and then deducting from 1.

At this stage, the Covertness Index of the edge is 0, as is revealed through the following calculation:

$|R_{ab}| = 2$, $|R_{ab}'| = 2$,

Overtness Index = $|R_{ab}| \div |R_{ab}'| = 1$

C_{ab} = Covertness Index = (1 – Overtness Index) = (1 – 1) = 0;

The last entity in the set comprising the Edge-Vertex between nodes ***a*** and ***b*** is the set of sets containing details of the nodes to which the copies of a particular mail exchange are marked. Since we can see here that node ***b,*** which has sent the mail to node ***a*** in this instance, has marked out copies to nodes ***d***, ***e*** and ***g*** and ***h,*** this entity reads as **{{(V_c, V_e, V_f),(V_d, V_e, V_g, V_h)}.**

Summarized results after the second mail exchange:-

1. Relationship Set**(R_{ab})** :

d) **$R_{ab(2)} = 2$**;

e) **R_{bc}={1,1}**;

f) **$|R_{ab}| = 2$;**

2. Shared Relation Set (R_{ab}'):

d) **$R_{ab(2)}' = 1$**;

e) **R_{ab}'= {1,1}**; Increases by one element since there is at least one copy.

f) **$|R_{ab}'| = 2$;**

3. Covertness Index (C_{ab}) = **0**:

4. Set of Copied Nodes = **{(V_c, V_e,V_f), (V_d, V_e,V_g,V_h)};**

5. **$(E_v)_{ab}$ = {V_a, V_b, $\lvert R_{ab}\rvert$, $\lvert R_{ab}'\rvert$, C_{ab}, , {(V_c,V_e, V_f),,(V_d, V_e,V_g,V_h)}};**

6. **i.e. $(E_v)_{ab}$ = {V_a, V_b,2, 2, 0, {(V_c,V_e, V_f),(V_d, V_e,V_g,V_h)}};**

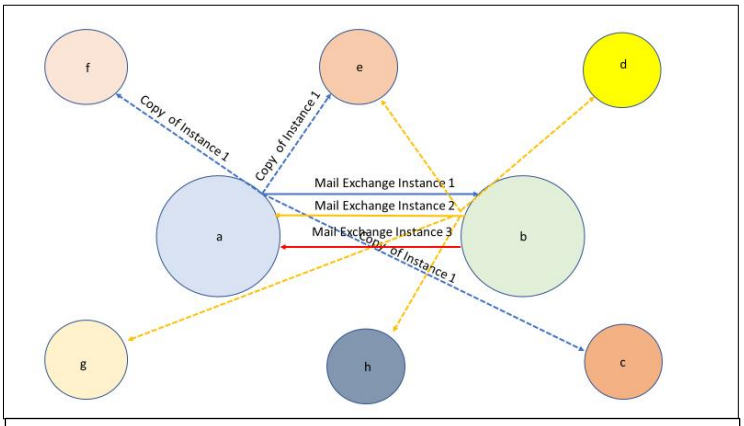


Figure 3.6(c)

The third instance of mail being exchanged between nodes ***a*** and ***b*** is shown in Figure 3.6 (c) above. It is seen that node ***b*** has sent a mail to node ***a***. The arrow representing this instance is red. The mail exchange between ***a*** and ***b*** is shown as a solid red line, and no copies of this instance are seen to be marked out.

As in the first two mail exchange instances, we continue constructing the set on which edge vertex $(E_v)_{ab}$ is based.

After the third instance of mail exchange between nodes ***a*** and ***b***,

$(E_v)_{ab}$ = {V_a, V_b, $\lvert R_{ab}\rvert$, $\lvert R_{ab}'\rvert$, C_{ab}, {(V_c,V_e, V_f), (V_e, V_d, V_g,V_h), (φ)}

As in the earlier instances**,**

V_a is node ***a,*** and V_b is node ***b,*** R_{ab} is the **Relationship Set** between nodes ***a*** and ***b*** as defined earlier**,** R_{ab}' is the Shared Relationship, **Set** between nodes ***a*** and ***b,*** C_{ab} is the **Covertness Index** of the tie between nodes ***a*** and ***b*** after this instance of an exchange of mail and the

last entity in the list is the set of sets of nodes to which the particular instance of mail was marked out.

After the third instance of an email exchange between nodes ***a*** and ***b***, the **Relationship Set cardinality** between ***a*** and ***b*** is **three** since three emails are exchanged between ***a*** and ***b***. Still, the **Shared Relationship Set cardinality** remains at **two** as no copies have been marked out this time. The **Covertness Index** is obtained as earlier by computing the Shared Relationship Set's cardinality ratio to that of the Relationship Set, respectively, and then deducting from 1.

The Covertness Index of the edge is computed to be after the third instance of mail exchange and calculated as follows:

$|R_{ab}| = 3$, $|R_{ab}'| = 2$,
Overtness Index = $|R_{ab}'| \div R_{ab}| = 2/3 = 0.67$
C_{ab} = Covertness Index = (1 – 0.67) = 0.33;

Again, the last entity in the list-set comprising the Edge-Vertex is the set of sets containing details of the nodes to which the copies of a particular mail exchange are marked. Since we can see here that node ***b,*** which has sent the mail to node ***a*** in this instance, has not marked out copies to any other nodes and this entity reads as **{{(V_c, V_e, V_f),(V_d, V_e, V_g, V_h), (φ)}.**

Summarized results after the third mail exchange:-

1. Relationship Set**(R_{ab})** :

g) $R_{ab(3)} = 3$;

h) R_{bc}={1,1,1};

i) $|R_{ab}| = 3$;

2. Shared Relation Set (R_{ab} '):

g) $R_{ab(-)}' = \phi$; Since there are no copies, there will be no entry in this set.

h) $R_{ab}' = \{1,1,\}$; (No change since there are no copies this time).

i) $|R_{ab}'| = 2$;(No change).

3. Covertness Index (C_{ab}) = **0.33**:

4. Set of Copied Nodes = $\{(V_c, V_e, V_f), (V_d, V_e, V_g, V_h), (\varphi)\}$;

5. $(E_v)_{ab} = \{V_a, V_b, |R_{ab}|, |R_{ab}'|, C_{ab}, , \{(V_c, V_e, V_f),,(V_d, V_e, V_g, V_h), (\varphi)\}\}$;

6. **i.e.** $(E_v)_{ab} = \{V_a, V_b, 3, 2, 0.33, \{(V_c, V_e, V_f),(V_d, V_e, V_g, V_h), (\varphi)\}\}$;

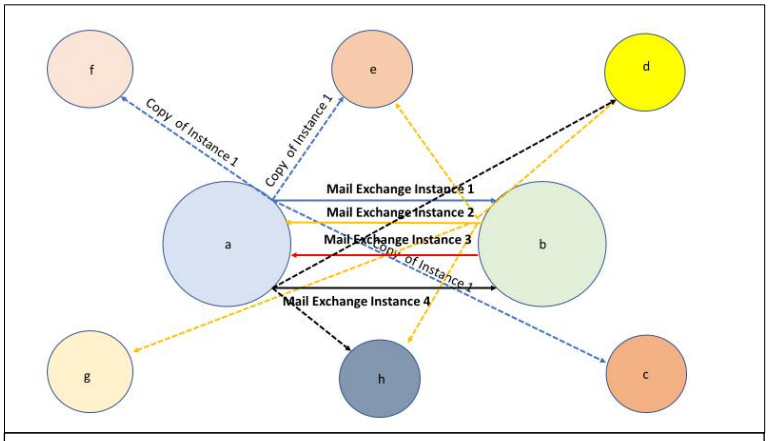


Figure 3.6(d)

The fourth and last instance of mail being exchanged between nodes ***a*** and ***b*** is shown in Figure 3.6 (d) above. Node ***a*** has sent a mail to node ***b***. The arrow representing this instance is black. As in the earlier instances, the mail exchange between ***a*** and ***b*** is shown as a solid black line, and two copies of this instance are seen to be marked out by node ***a***, i.e., to nodes ***d*** and ***h***, respectively (dotted black arrows).

After this instance,

$E_{ab} = \{V_a, V_b, |R_{ab}|, |R_{ab}'|, C_{ab}, \{(V_c, V_e, V_f), (V_e, V_d, V_g, V_h\}, (\varphi), (V_d, V_h)\}$

Four emails have been exchanged between nodes ***a*** and ***b*** at this stage; the **Relationship Set cardinality** between ***a*** and ***b*** is **four** since four emails are exchanged. The **cardinality** of the **SharedRelationshipSet**increases to **3** as copies have been marked out this time.

The Covertness Index is calculated, as shown below:

$|R_{ab}| = 4, \ |R_{ab}'| = 3,$

Overtness Index = $|R_{ab}'| \div |R_{ab}| = 3/4 = 0.75$

C_{ab} = Covertness Index = (1 – 0.75) = 0.25;

The last entity in the set comprising the Edge-Vertex reads after the last exchange of mail as

{{(V_c,V_e, V_f),(V_d, V_e,V_g,V_h), (φ), (V_d ,V_h) }.

Summarized results after the fourth mail exchange:-

1. Relationship Set**(R_{ab})** :

a) $R_{ab(4)}$ **= 4**;

b) R_{bc}**={1,1,1,1}**;

c) **$|R_{ab}|$ = 4;**

2. Shared Relation Set (R_{ab} ‘):

a) $R_{ab(3)}'$ **= 1**; This is the third entry in this set as there was no entry in the last (third) instance of mail transfer.

b) R_{ab} **‘= {1,1,1}**; Increases in size since there are copies this time.

c) **$|R_{ab}'|$ = 3;**

3. Covertness Index (C_{ab}) = **0.25**:

4. Set of Copied Nodes = **{(V_c, V_e,V_f), (V_d, V_e,V_g,V_h),(φ), (V_d,V_h)};**

5. E_{ab} **= {V_a, Vb, $|R_{ab}|$, $|R_{ab}'|$, C_{ab}, , {(V_c,V_e, V_f),,(V_d, V_e,V_g,V_h),(φ), (V_d,V_h) }} ;**

6. **i.e. E_{ab} = {V_a, V_b,4, 3, 0.25, {(V_c,V_e, V_f),(V_d, V_e,V_g,V_h)(φ), (V_d,V_h)}}**

The table3.1below illustrates how the list-set generated by the application of the Edge-Vertex function on dyad (*a*,*b*) evolves over the increasing instances of mail transfer:

Instance #	Dyad Pair		Relationship Set Cardinality	Shared Relationship Set Cardinality	Covertness Index	Set of Copied Nodes	Remarks
	Node#1	Node#2					
1.	Va	Vb	1	1	0	$\{(V_c, V_e, V_f)\}$	Mail copied
2.	Va	Vb	2	2	0	$\{(V_c, V_e, V_f), (V_d, V_e, V_g, V_h)\}$	Mail copied
3.	Va	Vb	3	2	0.33	$\{(V_c, V_e, V_f), (V_d, V_e, V_g, V_h), (\phi)\}$	Mail not copied
4.	Va	Vb	4	3	0.25	$\{(V_c, V_e, V_f), (V_d, V_e, V_g, V_h), (\phi), (V_d, V_h)\}$	Mail copied

Table 3.1 Table showing the Edge-Vertex list-set getting progressively updated after successive exchanges of mails between nodes *a* and *b*. The last column in the table above shows whether any copies were marked out during the instance of mail transfer between the nodes.

3.9 Overtness and Covertness of Ties Explained

Let us now define a function $\sigma_{ij}$where $\sigma_{ij} = (r_{ij}' / r_{ij})$

In all cases, $0 \leq \sigma_{ij} \leq$ one since the number of communications copied to other nodes can't exceed the total number of emails exchanged.

The closer σ_{ij} is to 1, the more accessible the communications from the dyad to other nodes; in other words, the edge of the dyad is more transparent and less covert in the communication interchanges between its constituent nodes.

Thus, going by the above discussions, we may now define the Overtness Index of the tie between the nodes i & j as σij or the proportion of the total number of communication instances shared with other nodes.

Definition ***(Overtness Index)****Consider the finite set of all communications exchanged between nodes* $\boldsymbol{V_i}$ *and* $\boldsymbol{V_j}$, $\boldsymbol{R_{ij} = \{R_{ij(1)}, R_{ij(2)}, \ldots, R_{ij(m)}\}}$ *comprising m entities and consider* $\boldsymbol{R_{ij}' = \{R_{ij}'_{(1)}, R_{ij}'_{(2)}, \ldots R_{ij}'_{(k)}\}}$ *comprising k entities where k <= m, constituting the set of*

all mail communications between nodes V_i *and* V_j *which have been copied to nodes other than* V_i*and* V_j*. Then the ratio of their respective cardinalities* σ_{ij} where $\sigma_{ij} = (r_{ij}'/ r_{ij})$ *defines the* ***Overtness Index*** *of the nodes* V_i*and* V_j*.*

Likewise, the **Covertness Index** of the Edge between the nodes ***i***&***j*** will be defined by the formula:

$$C_{ij} = (1 - \sigma_{ij})$$

Definition ***(Covertness Index)****Consider the finite set of all communications exchanged between nodes* V_i *and* V_j $R_{ij} = \{R_{ij(1)}, R_{ij(2)}, \ldots, R_{ij(m)}\}$ *comprising m entities and consider* $R_{ij}' = \{R_{ij}'_{(1)}, R_{ij}'_{(2)}, \ldots R_{ij}'_{(k)}\}$ *comprising k entities where k <= m, constituting the set of all mail communications between nodes* V_i *and* V_j *which have been copied to nodes other than* V_i*and* V_j*. Then the ratio of their respective cardinalities when deducted from 1, i.e.,* $C_{ij} = (1- \sigma_{ij})$, where $\sigma_{ij} = (r_{ij}'/ r_{ij})$ *defines the* ***Covertness Index*** *of the nodes* V_i*and* V_j*.*

The concepts of Overtness and Covertness are explained illustratively by using the same example of mail exchanges between two nodes in a simple four-node network as in the earlier
Case. The figures below in sequence Figures 3.7(a) to (d) illustrate how the Covertness Index is constructed:

In the figures below, we have four nodes in a Network: ***a***, ***b***, ***c*** & ***d***.

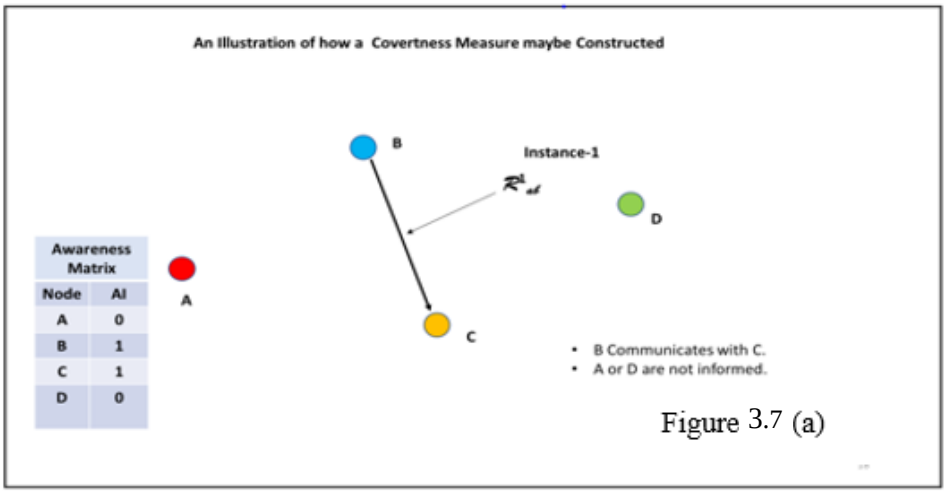
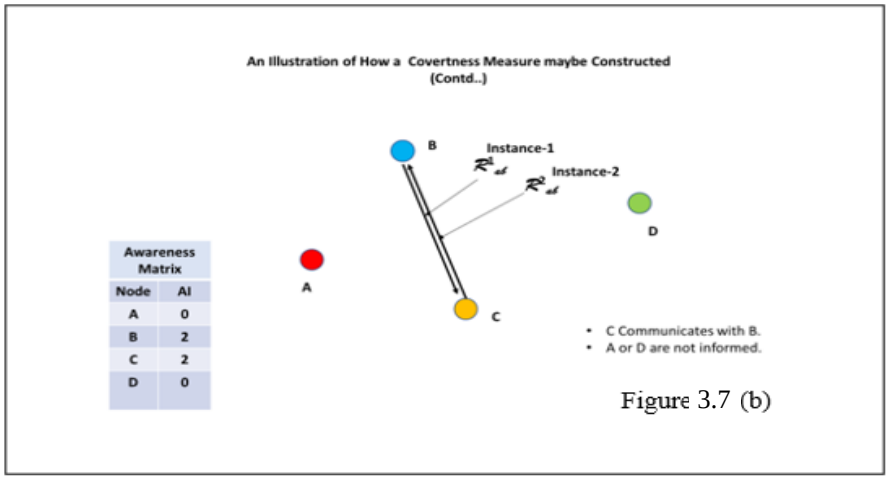


Figure 3.7 (a)

To start with, node ***b*** sends a mail to node ***c*** without marking any copies to either node ***a*** or node ***d***. In the **Awareness Matrix**[38] shown within the figure, the cells corresponding to node ***b*** and node ***c***are populated with **1**, reflecting the communications status. It may be seen that since no copies are marked to either node ***a*** or node ***d***, the cells corresponding to them remain unpopulated.

[38]An Awareness Matrix is described as a Matrix that reflects how many copies of mails exchanged between two nodes have been marked (copied) to other nodes and also indicates the identities of the nodes privy to such copied mails. An Awareness Matrix is specific to a node pair in a Social Network. This concept has been enlarged upon in the succeeding sections.



In Fig 3.7(b) above, we see that a mail has been directed back from node ***c*** to node ***b*** again, with no copies. In the Awareness Matrix shown to the bottom left of Figure 3.7 (b), in the cells corresponding to nodes***b***&***c***, the second column in the Awareness Index (or AI)[39], which counts the number of emails exchanged, increases to 2. Again, the cells corresponding to nodes ***a*** and ***d*** remain at zero, reflecting that there are no copies received by them of the emails exchanged between nodes ***b*** and ***c***.

Figure 3.7 (c) below shows that the third instance of mail is seen exchanged between nodes ***b*** and ***c***. This time around, however, a copy is marked to node ***a***. Thus, the count in the second column of the Awareness Matrix increases to **3** for both nodes ***b*** and ***c***. More importantly, the column corresponding to node ***a*** is now populated with a **1**, reflecting the node's status as the recipient of a copy of one of the instances of mail exchanged between nodes ***b*** and ***c***.

39 An Awareness Index is an entry in an Awareness Matrix that reflects the number of copied mails received by a node other than the nodes in the Dyad between whom the interchange of mails is going on. The concept is explained in a succeeding section.

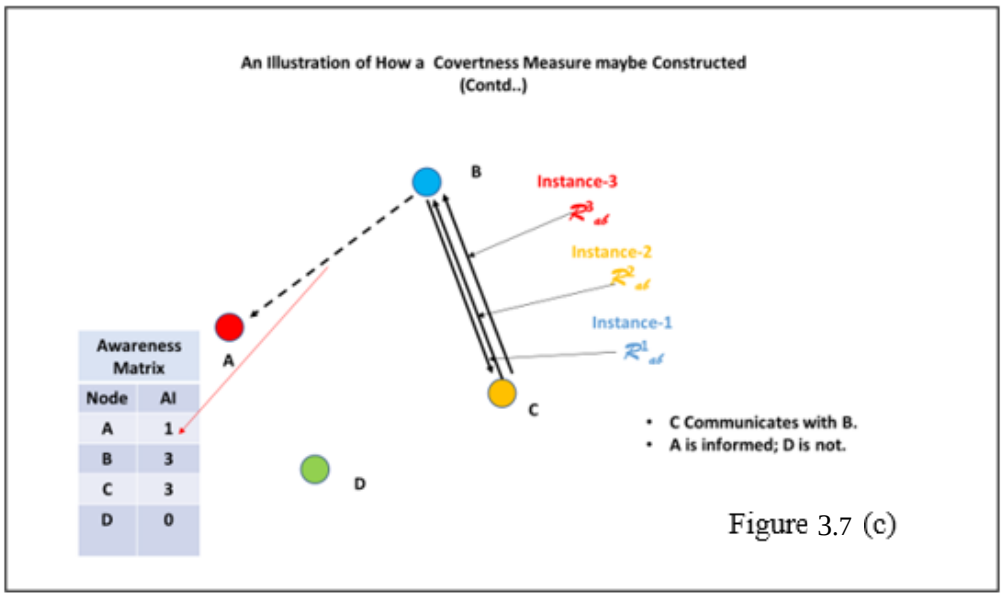
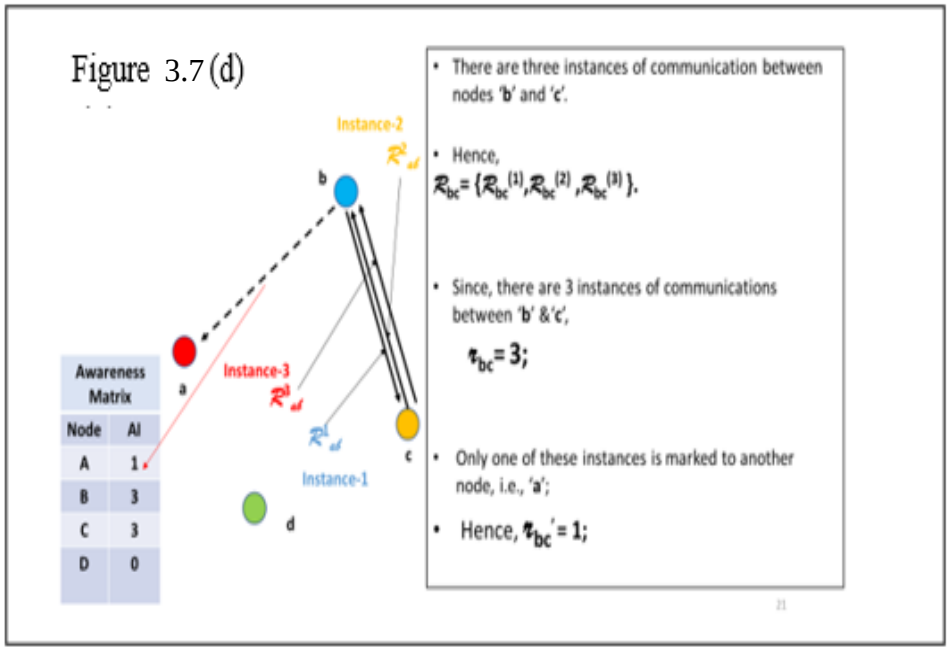


Figure 3.7 (c)

The mechanism by which the Covertness Index (CI) of the Information Tie between nodes ***b*** and ***c*** is calculated is shown in the text box to the right in Figure 3.7 (d). The total number of emails exchanged between nodes ***b*** and ***c*** is **3**. Of these three emails exchanged, only one mail is seen to be copied to a third node, namely node ***a***. Thus, two of the three mails exchanged between nodes ***b***and ***c***are confined between these nodes, i.e., the information in these emails has not left the dyad formed between nodes ***b*** and ***c***. As per the construct proposed in this study, the extent to which the information exchanges between ***b*** and ***c*** is not shared comes to **2/3** or **0.67 (67%),** which is also the **Covertness Index** of the edge or tie between nodes ***b*** and ***c***.



3.10 Developing a Balanced Covertness Index

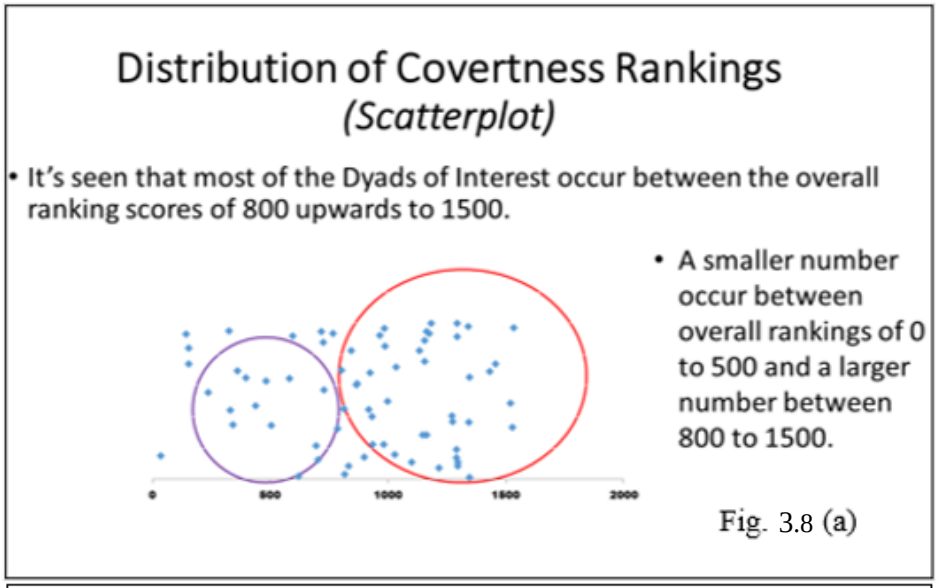
However, the above formulation is agnostic of the number of communication interchanges between the constituent nodes in the Edge. The Covertness Index should depend upon Rij,' i.e., the number of instances known to other nodes, and the cardinality of $\boldsymbol{R_{ij}}$, i.e., the number of instances of communications between ***i*** and ***j***. One of the ways this can be done is by re-defining the Covertness Index (or CI) as

$\boldsymbol{C_{ij} = C_{ij} \log (R_{ij})}$ **when** $\boldsymbol{R_{ij} = 0}$

$\boldsymbol{C_{ij} = 0}$ **when** $\boldsymbol{R_{ij} = 0}$

The reasons for selecting this redesign becomes obvious as we proceed to the actual calculations of the Covertness Index of ties between nodes. A sample calculation has been done in the subsequent sections to demonstrate the efficacy of the revised formulation. The figures below (not based on actual values, basically drawn to show the patterns) illustrate the improvement in results when the Revised formulation is applied to the ENRON dataset.

The scatter plots' horizontal axis represents the rankings of the edges of interest (EoIs) based on the Covertness Index scores associated with them. Figure 3.8 (a) shows the distribution of the EoIs when the unrevised Covertness Index metric is applied to the edges and their rankings calculated. Figure 3.8 (b) shows the shift leftwards along the horizontal axis of the rankings of the EoIs as the revised Covertness Index metric is applied to the edges in the ENRON mail corpus. The highest rankings occur at the horizontal axis's origin and decrease as we move to the right. The higher the Covertness Index value of an edge, the higher its rank, and the more it is positioned towards the left on the horizontal axis. The revised Covertness Index makes the rankings of the EoIs better and allows them to be detected more easily, as we will see later.



Scatterplot distribution of Edges of Interest's based on CI scores before applying the revised formulation.

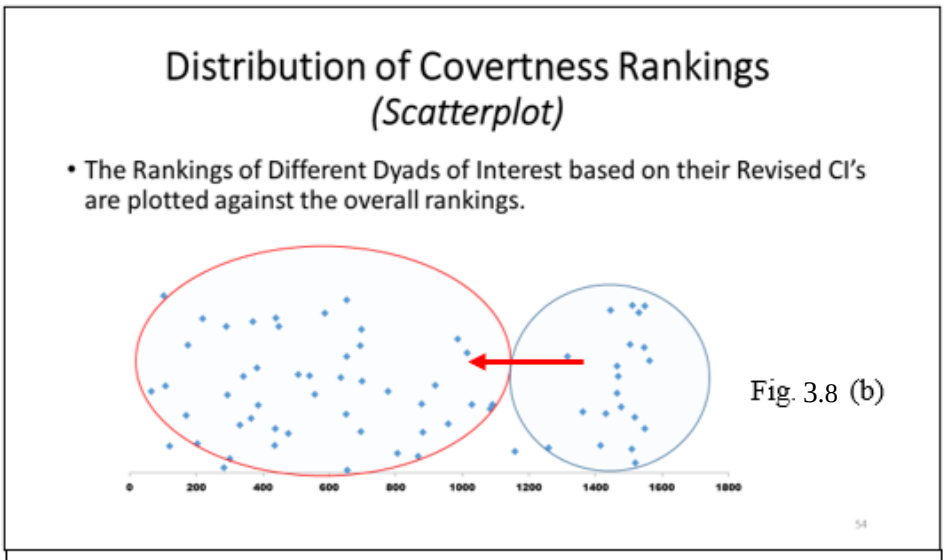


Fig. 3.8 (b)

Scatterplot distribution of Edges of Interest's shifts leftwards based on CI scores after applying the revised formulation.

Thus, it's trivial to observe from a comparative view of the scatterplots of the rankings of the edges of interest before and after applying the revised formulation that the Edges of Interest rankings have shifted to the left. That is, more of the Edges of Interest's now figure in the higher ranks than earlier, which is the improvement that has resulted from the revision of the CI formula.

A comparative graphical representation given in Figure 3.9below shows the improvement in covertness rankings of the edges of interest (EoIs) after applying the revised metric. Figure 3.10 shows the lift chart, which gives a clear picture of the comparative rankings of the EoIs before and after the application of the revised metric. The improvement arises from the utilization of the volume of transactions between the nodes in a dyad. Earlier, many of the edges had been assigned high rankings because they had not shared any copies of their exchanges. Still, this model failed to consider the possibility of very few mails having been exchanged and resulting in high index values based on very few copies having been marked out.

Thus, pairs of nodes with a high volume of mail exchanges between them may have marked out a marginally less percentage of their transactions as copies to outside mails are liable to be left out of the higher rankings. The revised metric's application solves this problem largely, as shown in Figure 3.9 and Figure 3.10. This is somewhat of a paradox, which is explained in the sections following in this chapter.

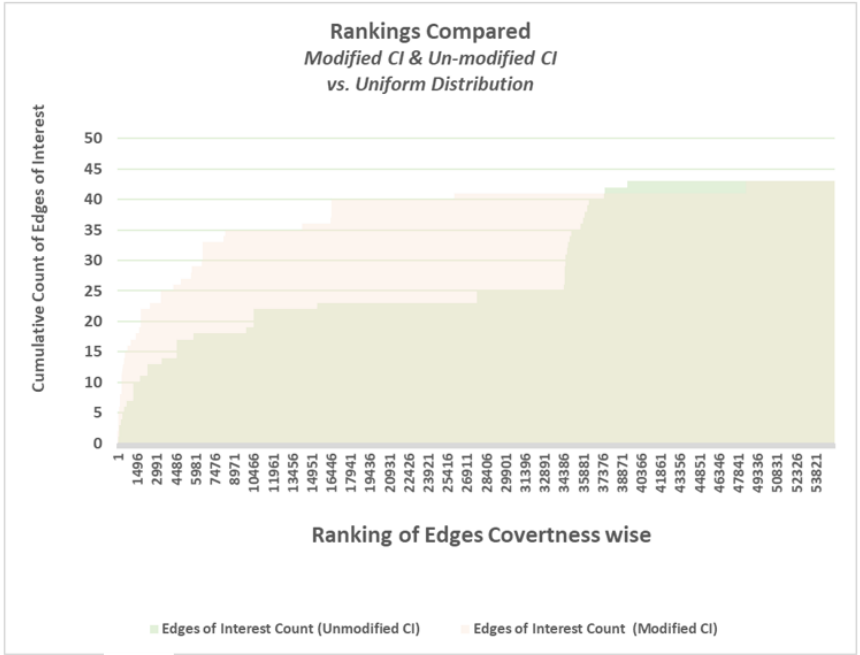


Figure 3.9 Graphical comparison of the Edges of Interest or EoIs before the application of the modified formula for calculating the Covertness Index (area in light green) and after applying the modified formula (area in light orange). The modified variant improves the chances of detection of covert edges measurably.

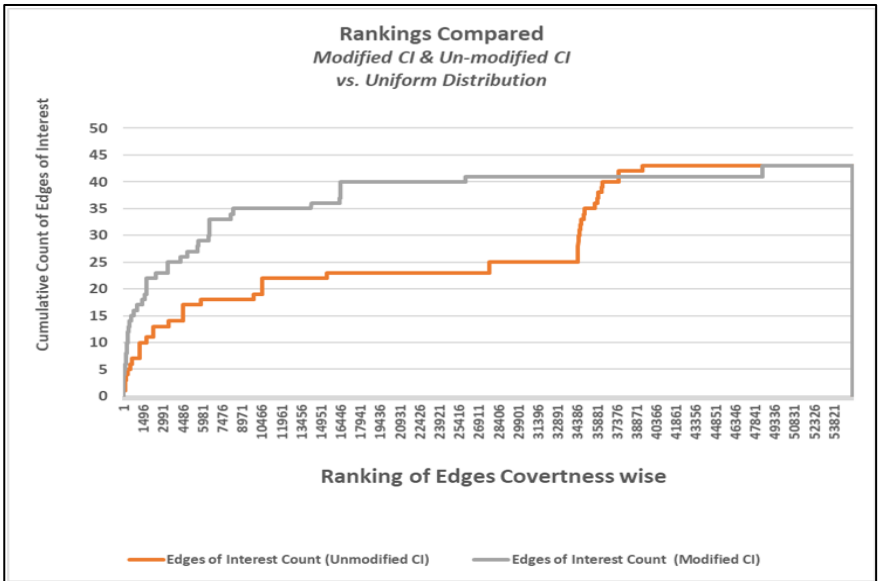


Figure 3.10 A lift chart showing the improvement in rankings of the Edges of Interest after the application of the revised formulation. In the plot, the line in grey shows the position after applying the modification and the line in orange shows the position without effecting the modification.

3.11 ENRON email Dataset: A Brief History

Enron was founded in July 1985 when Texas-based Houston Natural Gas merged with InterNorth, a Nebraska-based natural gas company. In its first few years, the new company was simply a natural gas provider. Still, by 1989 it had begun trading natural gas commodities, and in 1994 it began trading electricity, and by 2001 it was executing on-line trades worth about $2.5 billion a day. By the late 1990s, Enron had begun shuffling much of its debt obligations into offshore partnerships—many created by Chief Financial Officer Andrew Fastow. At the same time, the company was reporting inaccurate trading revenues. Enron was also using its partnerships to sell contracts back and forth to itself and booking revenue each time.

In February 2001, Jeffrey Skilling, the president, and the chief operating officer, took over as Enron's chief executive officer, while former CEO Kenneth Lay stayed on as chairman. In August, however, Skilling abruptly resigned, and Lay resumed the CEO role. By this point, Lay had received an anonymous memo about the Fastow partnerships, which warned

of possible accounting scandals. As rumors about Enron's troubles abounded, the firm shocked investors on October 16 when it announced that it would post a $638 million loss for the third quarter and take a $1.2 billion reduction in shareholder equity due in part to Fastow's partnerships. Simultaneously, some officials at Arthur Andersen LLP, Enron's accountant, began destroying Enron audits documents.

By October 22, the Securities and Exchange Commission had begun an inquiry into Enron and the partnerships; a week later, the inquiry had become a full investigation. Its stock value began to crater—it fell below $1 per share by the end of November and was delisted on Jan. 16, 2002. Many executives at Enron were indicted for various charges, and some were later sentenced to prison.

3.12 Nature of the ENRON Dataset

The email inboxes of 151 executives were uploaded onto the Internet by the Federal Energy Regulatory Commission (FERC) on March 26, 2003, in the public interest. The federal agency decided to post hundreds of thousands of e-mails that Enron executives sent and received from 1998 through 2002. The FERC eventually culled the trove to remove the most sensitive and personal data after receiving complaints. Even so, the "Enron e-mail corpus," as the cleaned-up version is now known, remains one of the largest public domain databases of real e-mails in the world. Please see FERC (2013) for the Federal Energy Regulatory Commission's website on the Enron investigation, FERC (2003) for the final order releasing the data to the public, and the informative book on the ENRON scandal by McLean and Elkind (2013) for a popular account of the Enron scandal.

This corpus is valuable to computer scientists and social-network theorists in ways that the e-mails' authors and recipients could never have intended. The mails are a rich brew of how real people in a real organization use e-mail—full of mundane lunch plans, boring meeting notes, embarrassing flirtations that revealed at least one extramarital affair, the damning missives that spelled out corruption. Research into the corpus is prolific and wide-

ranging, so much so that it has become the foundation of hundreds of research studies in fields as diverse as machine learning and workplace gender studies. A selection from the large range of ENRON studies is briefly presented here to highlight some of the research-related methodologies that the corpus has spurred and the approaches adopted. One of the earliest research studies on the subject was by Shetty and Adibi (2004) whodeveloped an RDBMS (MySQL) database of the corpus, Wang et al. (2014) have utilized the corpus for conducting anomaly detection studies in a dynamic network, Diesner et al. (2005) have provided a social network analysis procedure that focuses on changes in behavior during the scandal period, Deitrick et al. (2012) have developed a neural network model predicting the gender of an email-based on the email exchanges that happened in ENRON, Peterson et al. (2011) have proposed measures of formality in the email correspondence, Chapanond et al. (2005) have adopted a graph-theoretic and spectral classification analysis, Martin et al. (2005) have suggested techniques for detection of abnormal email activity in sent messages, Zhou et al. (2006) proposed a probabilistic (Bayesian) approach to community detection for identifying employees who were involved in the scam, and Zhou et al. (2007) have also developed methodologies for effecting data cleaning with focus on email aliases, Hardin and Sarkis (2015) discuss six measures of the Enron corpus based on the adjacency matrices of the email network, and suggest how these measures can be used in undergraduate education and research. They also provide a brief analysis of the group membership of the most connected cliques, found through the method of hierarchical clustering and hierarchical decomposition. Alkhereyf & Rambow (2017) have developed a classifier that trains on the Enron email corpus and tests the Enron email corpora. They show that information from the email exchange networks improves the performance of classification

This wide-ranging and roving research has had widespread applications: computer scientists have used the corpus to train systems that automatically prioritize certain messages in an in-box and alert users that they may have forgotten about an important message. Other researchers have used the Enron corpus to develop systems that automatically organize or summarize messages. The data set has somehow touched much

of today's software for fraud detection, counterterrorism operations, and mining workplace behavioral patterns over e-mail.

The reason why the Enron email corpus is used in this study is access to a huge volume of existing research, which enables the comparison of different techniques and evolution of better network analytics with demonstrable improvements in outcomes.

3.13 Previous Research on ENRON Dataset

This dissertation focuses on the ENRON email corpus, which, strictly speaking, is not a covert network. The financial scam that engulfed the corporation was the handiwork of a select group of actors who planned their concert strategies. There have numerous analyzes of the ENRON corpus. Most approaches have focused on semantic-based algorithms where key words connected to the scam have been used to classify and cluster the employees of interest to the investigation. There also have been some structural approaches, namely, variants of the geodesic path algorithm of Djikstra (1959) or the ShortestPaths Network Search Algorithm, SPNSA(Magalingam, Davis and Rao 2015), wherein the authors have presented an algorithm designed to extract a smaller, more manageable, network of possible relationships from a large dataset of interactions and have further developed this algorithm to show that it performs well in a variety of scenarios, and can extract meaningful sub-networks for a criminal investigator to start an investigation.

Most of these researches on the ENRON corpus are based on a post hoc scenario, i.e., the analysis is aware of the employees involved, and then the research evolves methodologies to identify them optimally. Though many of the techniques do yield very impressive results in terms of bringing out the conspiracy structure from within the larger organizational network of ENRON, it needs to be borne in mind that in most cases, analysts may not aware that a conspiracy or an offense is underway or may just have an inkling without knowing about the principal actors and the way they are collaborating. This research orients itself to

such a blind scenario. It presumes not to know either the employees of interest or the acts of collaboration to commit fraudulent activities.

3.14 Metrics to Identify Covert Structures

The idea here is first to apply the attribute of covertness (described earlier) of ties between pairs of nodes and then rank them in a descending order of the index values, which is the preliminary step to building the first layer of communities in the ENRON network. But the higher-ranked edges in terms of covertness in their ties are not necessarily related to each other. That is, each edge may be confining information related to the exchange of emails between its constituent nodes for different reasons.

In this study, we seek to ascertain which of the edges (and, as by extension, the constituent pair of nodes) were involved in activities about insider trading of stocks before the scam coming loose, or just being privy to the events leading to it. As we have discussed earlier, the top-ranked covert Edges may be confining information due to reasons unconnected to our study. They may be ranked higher as they might have been secretive in a more fastidious manner.

The challenge then is to evolve a subsequent series of steps that will not only take into account the scores generated from the application of covertness index but also cluster the ‘covert’ Edges into groups that will reflect the commonness of the covert intentions, i.e., the effort will be to cluster those top-ranked Edges which have specifically been part of the efforts to conceal information about the scam.

The flow chart below shows the series of steps schematically to identify top-ranked covert edges and then build bigger sub-nets based on common intent.

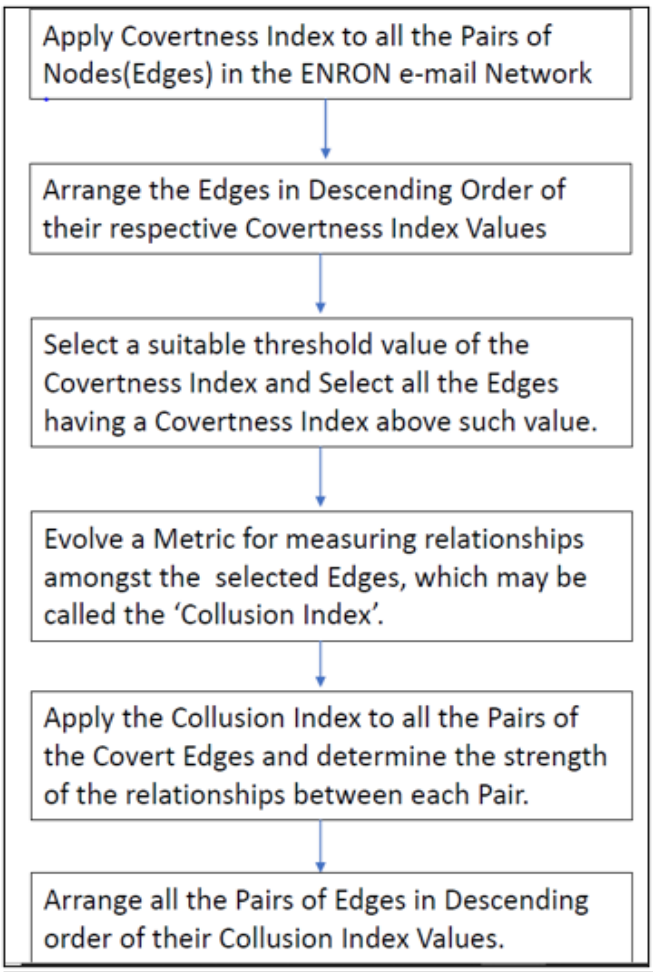


Figure 3.11 Flow Chart showing the steps taken in this study to first tag all available dyadic pairs or edges in the corpus with a Covertness Index value and then implement a ranking system. The edges with the highest index values will occupy the topmost ranks. A selection of these top ranked edges is made based on a heuristic threshold value of the Covertness Index. Following this, the second part of the study links pairs of these edges by a similarity index called the Collusion Index and covert community structures are the result.

3.15 ENRON Email Dataset in Numbers

Research into the ENRON email corpus reveals 34980 unique nodes, including many belonging to non-ENRON domains. Many of these nodes are not employees or even persons; rather, they are inanimate, for example, outlook-migration-team@enron.com.

Of these, the number of unique nodes belonging to the ENRON domain (i.e., possessing the suffix @enron.com) is 4643. The number of available Edges between these nodes comes to 122987 in the sense that there are these the number of pairs of nodes that have exchanged at least one mail amongst themselves. However, for this study's purposes, the non-ENRON domain mails have been removed from the above number, and the total number of unique Edges comes down to 55288. The template below summarizes some of the numbers connected to the ENRON e-mail corpus which have been used in this report:

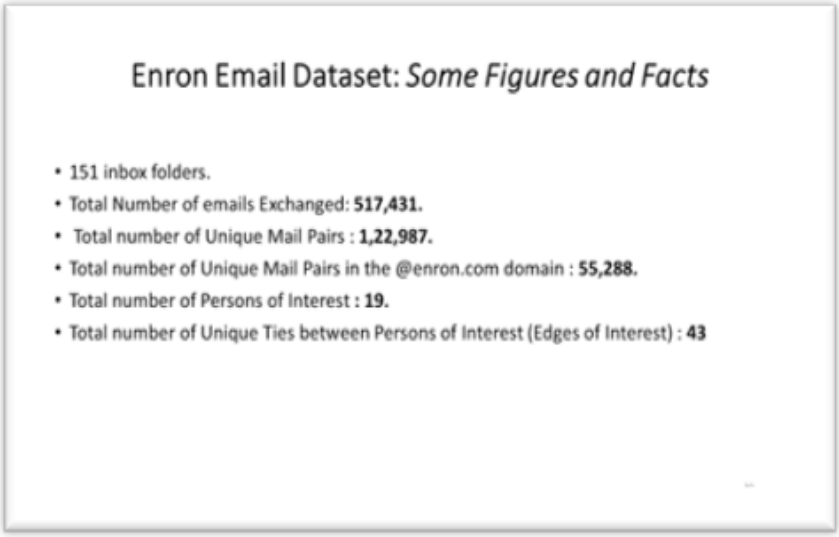


Figure 3.12 Salient facts and figures connected to the ENRON email corpus.

And as mentioned earlier, the email inboxes of only 151 employees are fully available in the public domain. The details of the remaining persons employed in ENRON can only be inferred from the information available in the different sections (like the 'From,' 'To,' 'cc' and 'bcc') within the existing in-boxes. Thus, the details of the network are substantially incomplete, which adds to the challenge.

For research, this study has considered 19 employees of ENRON who were either indicted during the subsequent trial proceedings or were in some way connected to the matter (We may call these players 'Complimentary Participants' described by Krebs (2001) In his seminal study regarding the 9/11 conspiracy, He defined *ComplimentaryParticipants* as "essentially co-conspirators who did not board the planes and who served as "conduits for money and also provided needed skills and knowledge."For this study's purposes, however, Participatory Participants' definition is broadened to mean all the players or nodes who might have been aware of the intentions or actions of the core participants who were later indicted). The broadening of the covert sub-net that participated in the actions leading to the crisis in ENRON allows what Krebs describes as "The addition of complimentary participants brought "shortcuts" to the network and improved the flow of communication. We assume that it also placed the overall network at a greater risk of detection." (Krebs, 2001).

The same strand of thought pervades the study by Morselli et al. (2007), who describe the utility of adding such Complimentary Participants thus, "The transition from an action network to a complete network that incorporates complimentary participants results in an inverse pattern when studying criminal enterprise networks" (p.147). The Caviar network[40]studied by Morselli (2009) included a greater number of participants than did Krebs' terrorist network (2001,2002). Furthermore, the Caviar network's action segment, the traffickers, represented most of the network's participants. Although larger than the terrorist network, the Caviar network was more clustered with shorter distances separating participants. Morselli describes the impact of including complimentary participation on the Caviar network model structure; thus - "The geodesic range for Caviar's trafficking segment was smaller (between 1 and 4). Adding non-traffickers to the network increased this range, but only slightly (between 1 and 5). The overall average path length for the trafficking segment of the network was 2.15. The addition of 28 complimentary participants, such as accountants, lawyers, legitimate importers, border agents, and other

[40]The Caviar network represents the profit-driven criminal enterprise network. It was reconstructed using electronic surveillance data gathered during a 2-year (1994–1996) tandem investigation (Project Caviar) by the Montreal Police, the Royal Canadian Mounted Police, and other national and regional law-enforcement agencies from various countries (i.e., England, Spain, Italy, Brazil, Paraguay, and Colombia). The investigation was aimed at dismantling a series of hashish and cocaine distribution chains that spanned across several countries and resulted in the seizures of four hashish and eight cocaine consignments.

non-traffickers, increased this metric by 37%, to 2.95. In contrast to the terrorist network, in which complementary actors made the action segment of the network more efficient by reducing the distance between participants, the addition of complementary participants to this drug trafficking network increased distance and therefore assured greater security for all involved."(p.147) It needs to be noted that the use of the term *complementary participants* is synonymous with the use of *facilitators* used by Klerks (2001).

The list of ENRON employees who are of some interest to this study's purposes are termed as nodes of interest or NoIs, and their list is furnished in the table3.2below. The table itself is partitioned into two columns, the first one comprising employee mail-ids whose inboxes are available in the public domain for analysis. The second column contains employee mail-ids whose inboxes are not provided initially. This partition is made to test the impact of incomplete data on the results obtained through the experiments conducted during this research.

NoI without Inboxes	**NoI with Inboxes**
andrew.fastow@enron.com	david.delainey@enron.com
lea.fastow@enron.com	jeff.skilling@enron.com
richard.causey@enron.com	kenneth.lay@enron.com
tim.belden@enron.com	jeff.dasovich@enron.com
jeff_dasovich@enron.com	jeffrey.shankman@enron.com
james.steffes@enron.com	karen.denne@enron.com
Rob.bradley@enron.com	steven.kean@enron.com
	maureen.mcvicker@enron.com
	sally.beck@enron.com
	louise.kitchen@enron.com
	vince.kaminskyenron.com
	angela.schwarz@enron.com

Table 3.2 Bi-columnar table showing the employee mail-ids of interest. With and without mailboxes.

Figures3.13and 3.14below show the existing e-mail links amongst the Nodes of Interest (NoIs) in the ENRON mail corpus. (The mapping of the Node-Ids to the Mail-Ids of the individual employees of interest is shown in table 3.3). The size of the labels against each node is proportionate to that node's degree, i.e., the number of links the employee

represented by the node has got through the exchange of emails with other employees (nodes). The links or edges are undirected. Each existing tie amongst the noise is denoted as an Edge of Interest (EoI). An EoI is defined to exist as a link between two nodes (employees) comprising a dyad and is unweighted. That is, the edge between two nodes isn't influenced by the number of incoming or outgoing information exchanges (e-mails). An edge is defined to have a value of 1 if any mail has been exchanged between the nodes it connects and 0 if there are no mails exchanged along with it. Each edge in the ENRON mail corpus is acted upon by a function termed as the Edge-Vertex function or $(E_V)_{ij}$, where i and j are the nodes, the tie between which forms the substrate upon which the Edge-Vertex function acts. The concept of the Edge-Vertex function has been defined earlier in this study. For this study's purposes, the tie or the edge between the constituent nodes is an atomic unit that serves as the building block of covert communities within the ENRON mail network. In effect, each edge plays the same role as a node or a vertex in conventional social network analysis techniques.

In the ENRON email dataset, 19 employees are within the focus of this study as players who have some knowledge about the insider trading scam that happened in the period between 1998 to 2002. Many of these employees are not having visible ties in the form of mail exchanges between themselves. Scrutiny reveals that there are only 43 edges amongst these employee mail-ids, whereas, in theory, there should have been 81 undirected edges ((19 x 18)/2, or, n(n-1)/2), which is a fairly dense matrix. The subnetwork density of mail exchanges that may be defined amongst these employees is 53% (43/81). We may compare this with the ENRON mail network's overall density, where there are 4600 or so unique mail-ids that result in a potential 10,000,000 (10 million-plus) ties. In contrast, only 123,000 edges are available (i.e., at least one mail has been exchanged between the dyads' nodes that contain these edges). The ENRON mail network's overall density of linkage comes to approximately 0.01 (123,000 / 10,000,000) or 1%.

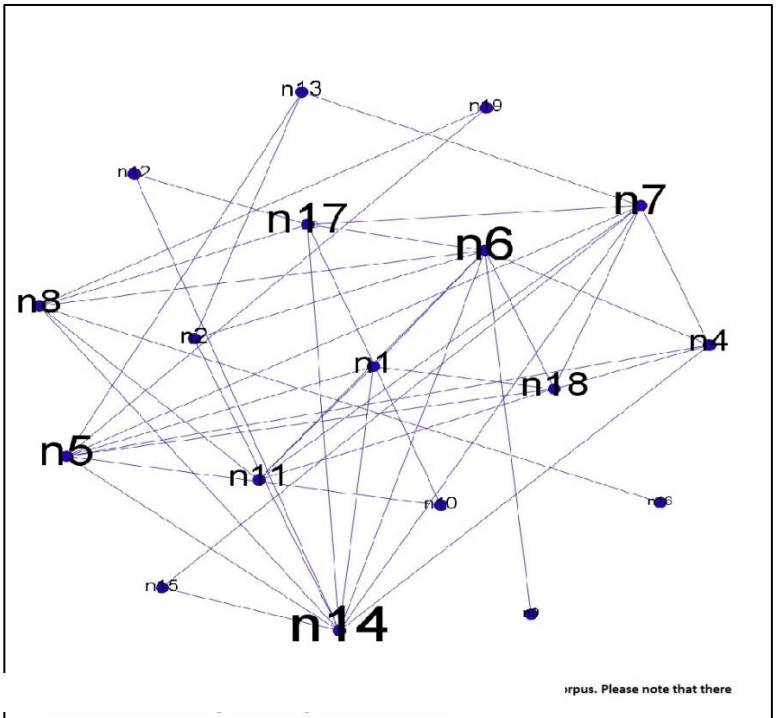


Figure 3.13

Graph showing the links amongst the nodes of interest (NoIs) in the ENRON e-mail network.

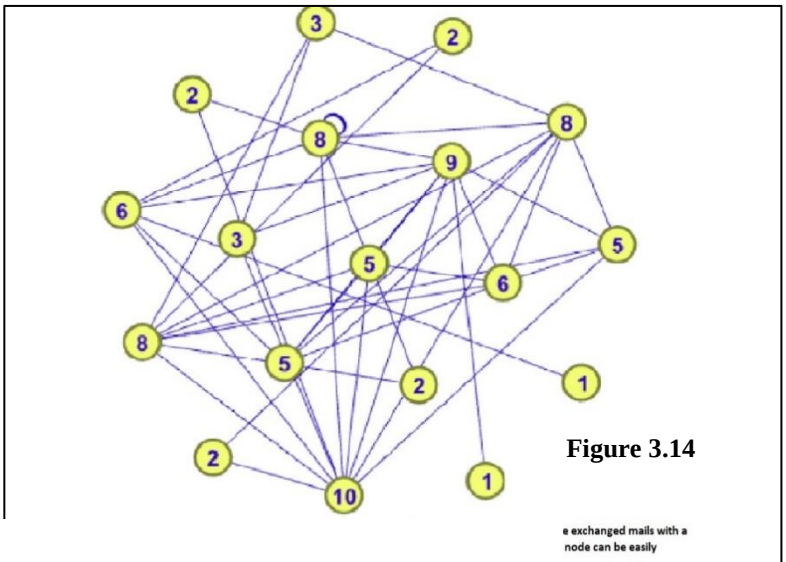


Figure 3.14

The same graph with the nodes now displaying their respective degrees.

Node-Id	Mail-Id
1	andrew.fastow@enron.com
2	richard.causey@enron.com
3	lea.fastow@enron.com
4	tim.belden@enron.com
5	david.delainey@enron.com
6	jeff.skilling@enron.com
7	kenneth.lay@enron.com
8	jeff.dasovich@enron.com
9	jeffrey.shankman@enron.com
10	sally.beck@enron.com
11	louise.kitchen@enron.com
12	james.steffes@enron.com
13	karen.denne@enron.com
14	maureen.mcvicker@enron.com
15	rob.bradley@enron.com
16	jeff_dasovich@enron.com
17	steven.kean@enron.com
18	vince.kaminski@enron.com
19	angela.schwarz@enron.com

Table 3.3 Table showing Node Id to employee Mail-Id mapping for the network diagrams at Figures 4.15 and 4.16 above.

3.16 Applying Covertness Metric to Dataset

The challenge is to ascertain if a viable sub-net (community) comprising these nodes of interest (NoI) or at least a substantial portion of them can be constructed based around the Index of Covertness evolved earlier. The Covertness Index, whose procedure is outlined in an earlier section, is applied to each of the 55,288 Edges, and each of the above Edges has then been assigned an index.

An illustration of this process has been explained below:

We consider the instance of the edge formed between two of ENRON employees, namely, Kevin Hannon and Kenneth Lay.

It's seen from the data that the number of emails exchanged between these two nodes is 24.
Of these exchanges, no e-mails are seen to be copied to others.

Thus,

Let, ρ_{ij} = Total no. of mails exchanged = 24;

Let, ρ_{ij}' = Number of Mails copied to others = 0;

And let, σ_{ij} = Covertness Index for the tie (or edge) between nodes *i* and *j*.

Node *i* represents the mail-id of Kevin Hannon, and the node *j* represents the mail-id of Kenneth Lay.

Hence, the Covertness Index for ties between Kevin Hannon & Kenneth Lay is calculated as below:

σ_{ij} = (24 – 0)/ 24= 1.00. (A value of 1 signifies that Hannon and Lay's tie is perfectly opaque or covert).

Based on calculations like the one illustrated above, Covertness Index values have been calculated for all possible edges used to exchange messages within the ENRON e-mail network. Further, a ranking in descending order of the values of the Covertness Index has been worked out. That is, the ones with the highest Covertness Index values have been ranked higher, and those with lower index values occur down the ladder of rankings. Table 3.4shows the top 100 edges ranked according to their Covertness Index values, i.e., the edges having the highest values of Covertness Index occupying the top 100. The meaning of this exercise becomes clear once we look for the edges of interest to the study (EoIs),i.e., the edges connecting pairs of ENRON employees who are in some way connected with the scandal. The same table has been presented in table 3.5 with the EoIs highlighted in yellow. The rankings reflect the status of the opacity of the information exchanges that

have been transacted along the corresponding edges by the constituent nodes of a dyad that are joined by them. The fundamental philosophy of this study's approach to the concept of covertness is to devise a measure for measuring the confinement of information exchanges. Confinement of information is possible through many mechanisms such as deception (where the actors are exchanging data, but, in a manner invisible to an external observer), pre-existing ties (that is, the actors were in close touch with each other at some point before the surveillance and these ties are no longer reflected in the current time) or through plain silence, i.e., the existing communications between the actors are visible and complete. The actors avoid sharing it with the others in the network to keep their dealings a secret. Irrespective of the mechanism to keep the information exchanges under a lid, the covertness metric proposed in this study can draw out the essence of these efforts by the concerned actors (employees of interest in the ENRON network) information from seeping out.

It bears repeating here that there are only 43 possible edges of interest-based on the communications between the nodes of interest or NoIs comprising ENRON employees who had some part in the matter in question. To locate them in a 55,288 edges dataset is, to say the least, searching for the proverbial needle in a haystack. The real objective of developing a Covertness Index metric is to enhance the visibility of covert edges within this ocean of information.

Table 3.4 Edges ranked as per their Covertness Index values

From	To	Covertness Index	Ranking
jill.chatterton@enron.com	robert.badeer@enron.com	1	1
jill.chatterton@enron.com	tim.belden@enron.com	1	2
marie.heard@enron.com	kimberly.allen@enron.com	1	3
mollie.gustafson@enron.com	center.dl-portland@enron.com	1	4
sylvia.hu@enron.com	felecia.acevedo@enron.com	1	5
tim.belden@enron.com	portland.desk@enron.com	1	6
veronica.gonzalez@enron.com	bob.schorr@enron.com	1	7
amanda.rybarski@enron.com	edith.cross@enron.com	1	8
amanda.rybarski@enron.com	leonardo.pacheco@enron.com	1	9
amanda.rybarski@enron.com	mike.curry@enron.com	1	10
arsystem@mailman.enron.com	elizabeth.sager@enron.com	1	11
becky.spencer@enron.com	sandi.braband@enron.com	1	12
cathy.phillips@enron.com	mark.frevert@enron.com	1	13
chris.germany@enron.com	james.barker@enron.com	1	14
deb.korkmas@enron.com	ann.white@enron.com	1	15
ina.rangel@enron.com	jason.williams@enron.com	1	16
ina.rangel@enron.com	kelli.stevens@enron.com	1	17
ina.rangel@enron.com	tom.donohoe@enron.com	1	18
kay.chapman@enron.com	david.w.delainey@enron.com	1	19
kay.chapman@enron.com	rick.buy@enron.com	1	20
larry.campbell@enron.com	spendegr@enron.com	1	21
leticia.botello@enron.com	dennis.benevides@enron.com	1	22
leticia.botello@enron.com	james.lewis@enron.com	1	23
lorna.brennan@enron.com	stephen.dowd@enron.com	1	24
maritta.mullet@enron.com	david.bush@enron.com	1	25
shari.wicks@enron.com	john.ambler@enron.com	1	26
shari.wicks@enron.com	michael.grimes@enron.com	1	27
simone.la@enron.com	sally.beck@enron.com	1	28
stacy.walker@enron.com	dan.bruce@enron.com	1	29
stacy.walker@enron.com	dave.schafer@enron.com	1	30
david.delainey@enron.com	sally.beck@enron.com	1	31
andrea.ring@enron.com	sandra.brawner@enron.com	1	32
cheryl.johnson@enron.com	kysa.alport@enron.com	1	33
christi.nicolay@enron.com	christopher.calger@enron.com	1	34
david.delainey@enron.com	scott.josey@enron.com	1	35
david.steiner@enron.com	center.dl-portland@enron.com	1	36
mika.watanabe@enron.com	steven.kean@enron.com	1	37
arsystem@mailman.enron.com	louise.kitchen@enron.com	1	38
bernadette.hawkins@enron.com	andre.cangucu@enron.com	1	39
bernadette.hawkins@enron.com	andrew.fastow@enron.com	1	40
bob.ambrocik@enron.com	aaron.adams@enron.com	1	41
bob.ambrocik@enron.com	ana.agudelo@enron.com	1	42
bob.ambrocik@enron.com	billie.akhave@enron.com	1	43
chris.germany@enron.com	theresa.branney@enron.com	1	44
d..hogan@enron.com	lindsay.culotta@enron.com	1	45
eddie.zhang@enron.com	tara.piazze@enron.com	1	46
ina.rangel@enron.com	mog.heu@enron.com	1	47
janette.elbertson@enron.com	robert.bruce@enron.com	1	48
justin.rostant@enron.com	a..bibi@enron.com	1	49
kay.chapman@enron.com	james.ajello@enron.com	1	50

Table 3.5 Top-ranked edges as per their Covertness Index values. The presence of a single Edge of Interest may be seen amongst the entries.

From	To	Covertness Index	Ranking
jill.chatterton@enron.com	robert.badeer@enron.com	1	1
jill.chatterton@enron.com	tim.belden@enron.com	1	2
marie.heard@enron.com	kimberly.allen@enron.com	1	3
mollie.gustafson@enron.com	center.dl-portland@enron.com	1	4
sylvia.hu@enron.com	felecia.acevedo@enron.com	1	5
tim.belden@enron.com	portland.desk@enron.com	1	6
veronica.gonzalez@enron.com	bob.schorr@enron.com	1	7
amanda.rybarski@enron.com	edith.cross@enron.com	1	8
amanda.rybarski@enron.com	leonardo.pacheco@enron.com	1	9
amanda.rybarski@enron.com	mike.curry@enron.com	1	10
arsystem@mailman.enron.com	elizabeth.sager@enron.com	1	11
becky.spencer@enron.com	sandi.braband@enron.com	1	12
cathy.phillips@enron.com	mark.frevert@enron.com	1	13
chris.germany@enron.com	james.barker@enron.com	1	14
deb.korkmas@enron.com	ann.white@enron.com	1	15
ina.rangel@enron.com	jason.williams@enron.com	1	16
ina.rangel@enron.com	kelli.stevens@enron.com	1	17
ina.rangel@enron.com	tom.donohoe@enron.com	1	18
kay.chapman@enron.com	david.w.delainey@enron.com	1	19
kay.chapman@enron.com	rick.buy@enron.com	1	20
larry.campbell@enron.com	spendegr@enron.com	1	21
leticia.botello@enron.com	dennis.benevides@enron.com	1	22
leticia.botello@enron.com	james.lewis@enron.com	1	23
loma.brennan@enron.com	stephen.dowd@enron.com	1	24
maritta.mullet@enron.com	david.bush@enron.com	1	25
shari.wicks@enron.com	john.ambler@enron.com	1	26
shari.wicks@enron.com	michael.grimes@enron.com	1	27
simone.la@enron.com	sally.beck@enron.com	1	28
stacy.walker@enron.com	dan.bruce@enron.com	1	29
stacy.walker@enron.com	dave.schafer@enron.com	1	30
david.delainey@enron.com	sally.beck@enron.com	1	31
andrea.ring@enron.com	sandra.brawner@enron.com	1	32
cheryl.johnson@enron.com	kysa.alport@enron.com	1	33
christi.nicolay@enron.com	christopher.calger@enron.com	1	34
david.delainey@enron.com	scott.josey@enron.com	1	35
david.steiner@enron.com	center.dl-portland@enron.com	1	36
mika.watanabe@enron.com	steven.kean@enron.com	1	37
arsystem@mailman.enron.com	louise.kitchen@enron.com	1	38
bernadette.hawkins@enron.com	andre.cangucu@enron.com	1	39
bernadette.hawkins@enron.com	andrew.fastow@enron.com	1	40
bob.ambrocik@enron.com	aaron.adams@enron.com	1	41
bob.ambrocik@enron.com	ana.agudelo@enron.com	1	42
bob.ambrocik@enron.com	billie.akhave@enron.com	1	43
chris.germany@enron.com	theresa.branney@enron.com	1	44
d..hogan@enron.com	lindsay.culotta@enron.com	1	45
eddie.zhang@enron.com	tara.piazze@enron.com	1	46
ina.rangel@enron.com	mog.heu@enron.com	1	47
janette.elbertson@enron.com	robert.bruce@enron.com	1	48
justin.rostant@enron.com	a..bibi@enron.com	1	49
kay.chapman@enron.com	james.ajello@enron.com	1	50

A closer look at the second of the two tables reveals a single EoI (David Delainey – Sally Beck), which makes its appearance within the top 50 ranked Edges going by their Covertness Index values. This pair has been highlighted in yellow in the second table.

The complete picture of the 43 Edges of Interest rankings is shown in table 3.6 below.

From	To	Covertness Index	Ranking
david.delainey@enron.com	sally.beck@enron.com	1	32
maureen.mcvicker@enron.com	david.delainey@enron.com	1	122
maureen.mcvicker@enron.com	kenneth.lay@enron.com	1	123
jeff.dasovich@enron.com	louise.kitchen@enron.com	1	204
maureen.mcvicker@enron.com	andrew.fastow@enron.com	1	331
maureen.mcvicker@enron.com	richard.causey@enron.com	1	505
karen.denne@enron.com	richard.causey@enron.com	1	681
karen.denne@enron.com	david.delainey@enron.com	1	1214
maureen.mcvicker@enron.com	rob.bradley@enron.com	1	1215
vince.kaminski@enron.com	andrew.fastow@enron.com	1	1216
maureen.mcvicker@enron.com	tim.belden@enron.com	1	1729
maureen.mcvicker@enron.com	jeff.skilling@enron.com	1	2292
tim.belden@enron.com	jeff.skilling@enron.com	1	2293
vince.kaminski@enron.com	tim.belden@enron.com	1	3421
jeff.dasovich@enron.com	jeff_dasovich@enron.com	1	4540
sally.beck@enron.com	steven.kean@enron.com	1	4541
vince.kaminski@enron.com	kenneth.lay@enron.com	1	4542
james.steffes@enron.com	steven.kean@enron.com	1	5849
louise.kitchen@enron.com	kenneth.lay@enron.com	1	9876
andrew.fastow@enron.com	jeff.skilling@enron.com	1	10496
kenneth.lay@enron.com	tim.belden@enron.com	1	10497
louise.kitchen@enron.com	jeff.skilling@enron.com	1	10498
jeff.dasovich@enron.com	jeff.skilling@enron.com	1	15406
david.delainey@enron.com	angela.schwarz@enron.com	1	27723
richard.causey@enron.com	jeff.skilling@enron.com	1	27724
steven.kean@enron.com	kenneth.lay@enron.com	0.956043956	34423
jeff.dasovich@enron.com	angela.schwarz@enron.com	0.933729822	34455
jeff.dasovich@enron.com	steven.kean@enron.com	0.929648241	34465
karen.denne@enron.com	kenneth.lay@enron.com	0.927419355	34470
rob.bradley@enron.com	kenneth.lay@enron.com	0.9	34533
david.delainey@enron.com	kenneth.lay@enron.com	0.890909091	34547
vince.kaminski@enron.com	jeff.skilling@enron.com	0.863636364	34661
steven.kean@enron.com	jeff.skilling@enron.com	0.851239669	34723
david.delainey@enron.com	tim.belden@enron.com	0.821011673	34888
jeffrey.shankman@enron.com	jeff.skilling@enron.com	0.804878049	34973
louise.kitchen@enron.com	vince.kaminski@enron.com	0.727272727	35702
andrew.fastow@enron.com	louise.kitchen@enron.com	0.714285714	35894
vince.kaminski@enron.com	david.delainey@enron.com	0.710526316	35987
jeff.dasovich@enron.com	maureen.mcvicker@enron.com	0.666666667	36221
david.delainey@enron.com	andrew.fastow@enron.com	0.666666667	36317
steven.kean@enron.com	steven.kean@enron.com	0.615384615	37527
james.steffes@enron.com	maureen.mcvicker@enron.com	0.607142857	37559
maureen.mcvicker@enron.com	steven.kean@enron.com	0.5	39330

Table 3.6 The rankings of all the 43 Edges of Interest.

Based on our results, we can summarize the prevalence of the Edges of Interest in the total distribution as per the table below:

Top-ranked Edges (Overall)	Prevalence of Edges of Interest	Cumulative Occurrence of EoIs
1 - 100	1	1
101 - 500	4	5
500 - 1000	2	7
1000 - 1500	3	10
1500 - 2000	1	11
2000 - 3000	2	13
3000 - 4000	1	14
4000 - 5000	3	17
> 4000	26	43
Total	**43**	**43**

Table3.7 Table showing rank distribution of EoIs within the overall rankings of all edges

The above distribution is presented graphically in the figure below.

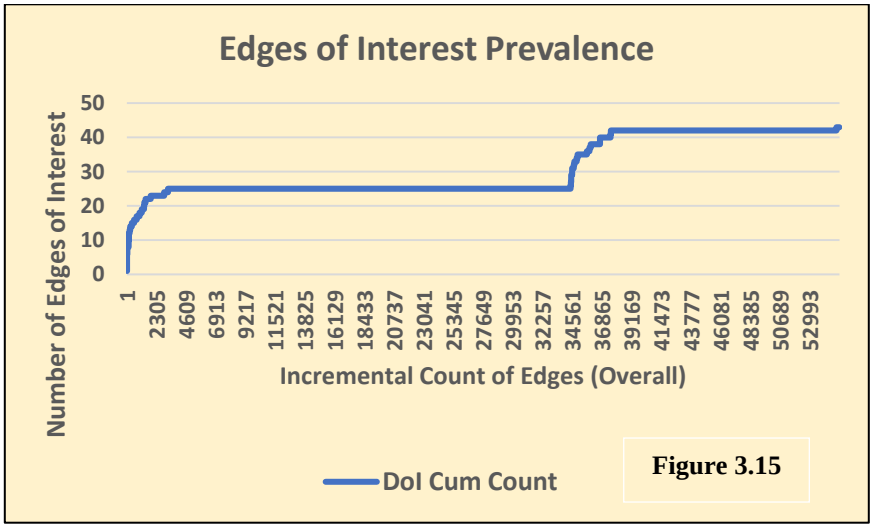


Figure 3.15

The ratio of the number of Edges of Interest (EoI) to the set of all edges comes to 0.000778 (43/55288) if we deem that the EoIs are distributed uniformly throughout the set of edges in the dataset. The plot above shows that the 13 EoIs are occurring in the top-ranked 2500 Edges. The ratio of the number of EoIs to the number of top-ranked covert edges improves to (13/2500) = 0.0052 if the covertness metric is applied. In the graph in figure 3.23 below, the prevalence of EoIs given a Uniform Distribution is shown for comparative purposes.

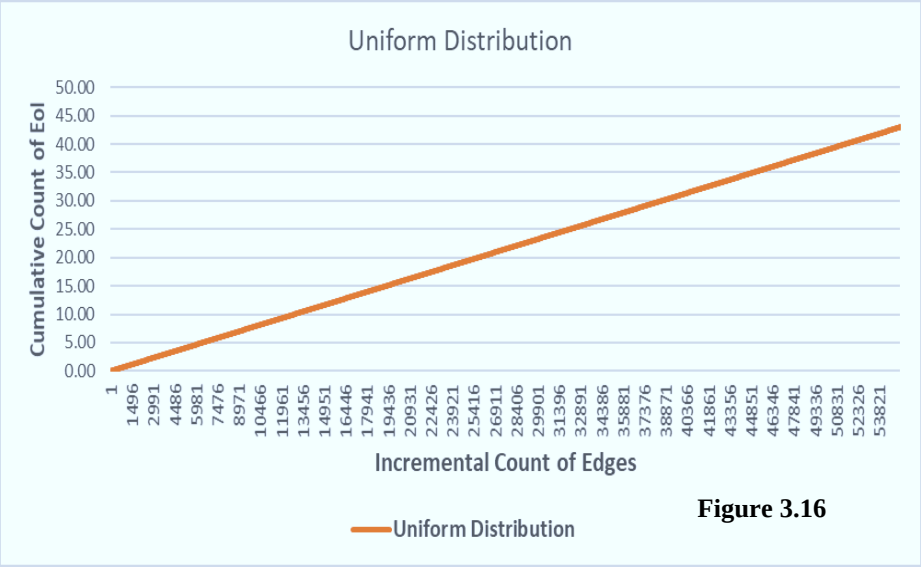


Figure 3.16

The above plot shows the linear pattern of growth of the prevalence of Edges of Interest provided that the EoIs are distributed uniformly throughout the distribution. The cumulative number of EoIs reaches 43 (i.e., the total number of EoIs in the dataset), eventually at the distribution's culmination.

Table 3.8 shows how the use of the Covertness Index metric enhances the chances of identifying the EoIs.The table compares the prevalence of the EoIs in the given distribution after employing the Covertness Index mechanism vis-à-vis the prevalence if we presume a Uniform Distribution:-

Top-ranked Edges (Overall)	Cumulative Occurrence of EoIs (Using CI)	Cumulative Occurrence of EoIs (Uniform Dist.)	Probability of Identifying an EoI using Covertness Index	The ratio of Number of EoI to Top-ranked edges using UniformDist
1 - 100	1	0.08	0.01	0.0008
101 - 500	5	0.39	0.01	0.0008
500 - 1000	7	0.78	0.007	0.0008
1000 - 1500	10	1.17	0.007	0.0008
1500 - 2000	11	1.56	0.006	0.0008
2000 - 3000	13	2.33	0.004	0.0008
3000 - 4000	14	3.11	0.0035	0.0008
4000 - 5000	17	3.88	0.0034	0.0008
>5000	43	43		

Table 3.8 Table showing the comparison between the performances of the Covertness Index Model and a Uniform Distribution Model.

The chances of detection of covert edges within the distribution improve markedly once the covertness metric is applied. In a real-world situation where there is a requirement of mounting surveillance over millions of nodes in a social network to detect only a handful of the covert ones, applying the proposed Covertness Index metric to the ties (or edges) substantially improves the chances of identification, which is particularly true in an environment where enforcement agencies may not have access to all the mails or contents of information exchanges. Besides, even in situations where such intrusive access to information is available, it'll still be arduous computationally to parse tens or hundreds of millions of messages or information exchanges to filter out contents interesting to the inquiry. Having only a few select edges (and the nodes associated with them) is an easy solution to bypass computationally expensive interventions and duck any legal issues arising from privacy and encryption.

The advantages of yoking the Covertness Index metric to enhance chances of identification of the Edges of Interest as compared to a Uniform Distribution is apparent from the lift chart provided below: -

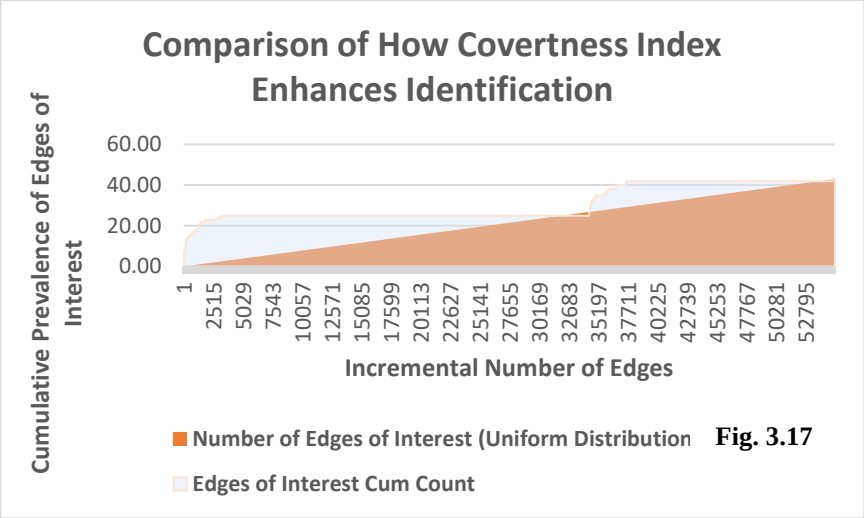


Fig. 3.17

3.17 Modified Covertness Index

Though the use of a mechanism involving the Covertness Index did boost the chances of identifying covert edges significantly, the rankings reveal a typical feature. It's easily observed that most of the top-ranked Edges are those whose constituent node pairs have not marked out any emails to other nodes at all, i.e., their Covertness Index is a perfect 1.

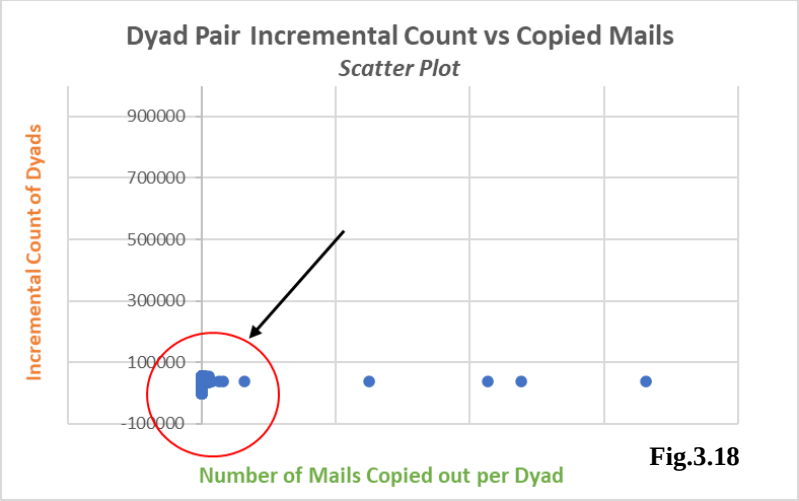


Fig.3.18

Figure 3.26 above is a scatter plot showing the correlation between the number of mails copied out by the dyads (shown on the horizontal axis) and the dyads' ranking based on their covertness indexes on the vertical axis. It's seen from the graph above that the data points are all concentrated around the 0 marks on the horizontal axis, i.e., the overwhelming majority of the top-ranked pairs haven't marked out any copies at all, which is an indicator of the fact that if we use the covertness index in its current context, many pairs which have no mails copied out score high ranks irrespective of the volume of information exchanged, i.e., there is little correlation with the data kept confined.

A similar inference is reached when we correlate the dyads' rankings with the total number of emails exchanged between the constituent nodes, which may be seen from the scatter plot below, reflecting the correlation between the covertness ranking dyads and the number of emails exchanged. Most of the top ranks are occupied by dyads, which have exchanged fewer emails comparatively. In contrast, dyads whose constituent nodes have exchanged more mails score lower in the rankings.

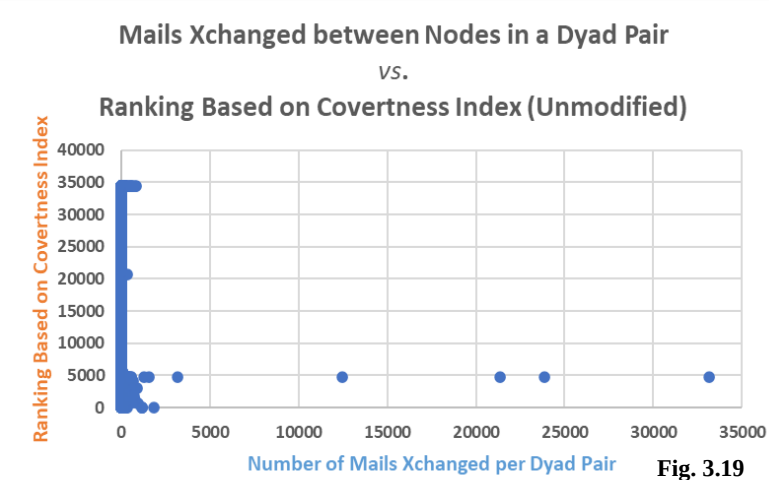


Fig. 3.19

This downside to the present state of estimation of covertness of an edge vertex (of a dyad) is illustrated by the example shown below.

Let's consider the Edge Vertex case formed by David Delainey and Sally Beck, two ENRON employees who are persons of interest in the case in question. The Covertness

Index of the tie between these two has been worked out. The Covertness Index of the Edge Vertex of another dyad of interest, i.e., Jeff Dasovich and Angela Schwarz, is worked out subsequently. The pairs have been selected specially to show the skewness caused by using the current methodology. The first pair have exchanged very few mails (47) but has marked out no copies to other nodes, whereas the second pair have exchanged a large number of mails (1177) but has marked out some as copies to outside nodes. But the covertness ranking of the first pair is far higher than that of the second pair. Thus, despite the large volume of information exchange, the second pair loses out in ranking.

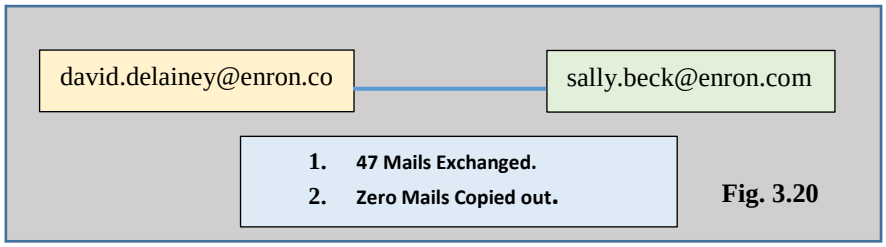


Fig. 3.20

The Covertness Index of the above pair is calculated as follows:

1. Number of mails exchanged = 47.
2. Number of mails copied out = 0.
3. The ratio of mails copied out to mails exchanged = 0.
4. Overtness Index = Ratio above = 0.
5. Covertness Index = 1 – (Overtness Index) = 1 – 0 =1.

We may compare this with the Covertness Index calculated in respect of the tie between two other employees of our interest, namely, Jeff Dasovich and Angela Schwarz.

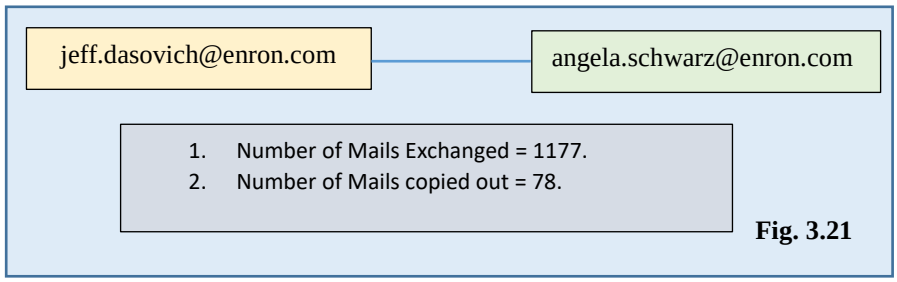


Fig. 3.21

The Covertness Index of the above pair is calculated as follows:

6. Number of mails exchanged = 1177.
7. Number of mails copied out = 78.
8. The ratio of mails copied out to mails exchanged = 0.0663.
9. Overtness Index = Ratio above = 0.0663.
10. Covertness Index = 1 – (Overtness Index) = 1 – 0.0663 =0.9337.

Thus, though the Edge between the said dyad is quite high at 0.9337, its rank in the Covertness List comes to 34454, extremely low. Besides, as stated earlier, the exchange of mails is fairly intense at 1177 mails. We may compare this with the number of emails exchanged between the previous pair of employees, i.e., David Delainey and Sally Beck, who have exchanged only 47 mails. That is to say, this pair has exchanged hardly about 4% of what the employees in the second pair (Jeff Dasovich and Angela Schwarz) have exchanged, yet their covertness ranking comes to within the top 100 overall. It needs to be noted here that we've considered only two dyads of interest here. If we look at the top ranks, all the reckoning pairs are sailing in the same boat practically. All of these dyads have not marked out copies to outside nodes. If we recall the covertness attribute definition, such pairs are 'Perfectly Covert' or opaque.

However, it may be observed that most of these 'Perfectly Covert' dyads have exchanged very few mails, which may have happened due to one or more of three reasons:

a) The nodes in the dyad are genuinely secretive and have communicated through means other than through mail exchange and would have led to fewer traces on the e-mail-based network, but the information would have been transferred all the same.

b) A more likely possibility is that the information about these dyads is incomplete. It may be recalled that the ENRON e-mail corpus comprises only 151 inboxes, basically of such employees who were thought to be of interest by the investigators at the point in time. This narrow selection has likely led to a substantial loss of information, and many mail exchanges that might have happened are simply no longer available.

c) There is also the likelihood that the high ranked covert dyads' constituent nodes have not exchanged many emails. The fact that nothing is marked out as a copy is more or less an indicator of the scant nature of the relationship between them.

Be that as it may, it leads us to the question of whether the number of emails exchanged between the constituent nodes of a dyad (which constitutes the Edge Vertex basically) has any impact upon the calculation of a putative Covertness Index. In other words, by simply considering the ratio between the mails copied out and the total emails exchanged, the complete picture of covertness is not revealed, which leads us to the issue of what other factors may influence the covertness in tie beside the ratio that has been just described in the paper. May we use the Covertness Index in combination with other more traditional network attributes to enhance its effectiveness, or should we just use the Index *in sui generis*?

As described elsewhere in this study, graph theory interventions in social network analysis identify the prominent nodes or actors and relationships (ties)at both the individual and group levels of analysis. The weapons of choice in graph-theoretic approaches are the various centrality measures that a node's prominence within a network encapsulates the relationships among all nodes. An individual node's ranks based around its centrality reflect its greater visibility to the other network nodes, and similar arguments may be made for ties (Edge Vectors) defining a relationship. As discussed earlier, the most widely used centrality measures in social network analysis are degree, closeness, and betweenness (Freeman, 1977, 1979). It's worth looking at the value of each of these measures in refining the Covertness Index proposed in this study.

The simplest of the centrality measure and perhaps the most easily applicable of all centrality measures is the Degree of a node or vertex. The degree is also called the Degree Centrality of a node in a network. Although degree centrality is a simple centrality measure, it can be very illuminating. In a social network, individual actors or nodes with

relationships with many others might have more influence, more access to information, or more prestige than those with fewer relationships (Newman, 2004, pp 168 -172).

Another candidate measure combined with the newly evolved Covertness Index is a natural extension is *eigenvector centrality*. Whereas degree centrality invests a centrality point for every neighbor a node is connected to in a network, eigenvector centrality looks at the neighboring node's quality. Not all neighbors are equivalent in a social network context. If any of the neighboring nodes is important in some meaningful manner, the node whose centrality is in question also gets more weightage for its links or ties. The concept of how eigenvector centrality has been applied in social networks, particularly in dark or covert networks, has been analyzed in an earlier section. Here, we're just considering its value as a combination or a companion attribute.

The other centrality measures that are popular in social network analytics are Closeness Centrality and Betweenness Centrality. Closeness centrality measures the mean distance from a node to other nodes. Closeness metric between a pair of nodes is based upon the geodesic path(s) between them. A geodesic path concept has been explained earlier and generally means the number of paths or edges connecting a pair of interest nodes. The mean geodesic distance from any node *i* to *j* is obtained by dividing the network nodes' geodesic distance. As has been discussed, this centrality measure is low for nodes separated by shorter geodesic distances. We may describe the importance of this measure in a social network in terms of a node with a low closeness measure, i.e., a person with a lower mean distance to others, which would imply that this node (or person) can reach others in the social network more expeditiously than others who possess higher mean distance scores. The distance measure is somewhat counter-intuitive in the sense that nodes with higher influence have lower scores, and nodes with lesser influence have higher scores. Hence, Closeness Centrality is generally considered as the inverse of the Mean Geodesic Distance of a node.

Another popular measure of centrality is the *betweenness centrality*, which measures the extent to which a node lies on the path between other vertices. This measure was proposed

by Freeman (1977,1979) and also by Anthonisse (1971). Betweenness is simply the number of geodesic paths passing through a particular node. Thus, nodes with high betweenness centrality tend to significantly influence their control over the information passing between other nodes in the network. As we have seen earlier, researchers and investigators of covert networks tend to favor this centrality measure since they can simply remove the nodes with the best values of betweenness centrality to disrupt communications in the network. This centrality measure differs from the other centrality measures in as much as it doesn't reflect how well connected a node is. Instead, it shows how much a node falls between other nodes in the network. Hence, if we take betweenness centrality scores as an index to measure a node's status, it may return high scores even if it has a low degree, has ties to other nodes that have a low degree, and might even have a long distance from others on an average.

The figure below (reproduced from Newman (2010, p.188)) shows how this is possible. Node A lies on a bridge between two groups on a network. Since any shortest path or any other path between a node in any component group and a node in another component group in a social network must pass along the bridge, Node A acquires a high betweenness score. The node may well be on the periphery of the network and is likely to have low eigenvector and closeness centrality, and its degree centrality is only two. Still, it will have a high influence on negotiating between two disparate groups within the network. As mentioned here, such nodes with a high between centrality are called *brokers* in a sociological context.

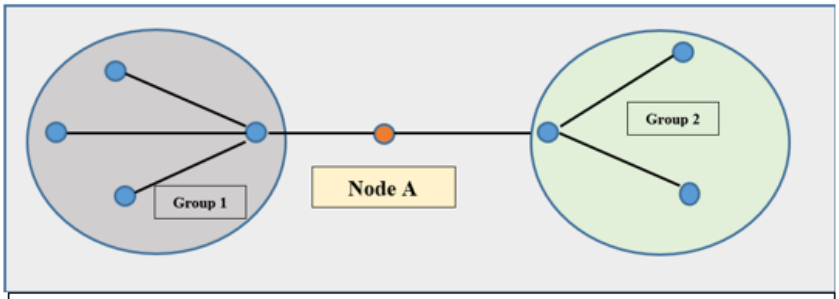


Figure 3.22:A low-degree node with high betweenness. In this sketch of a network, node A lies on a bridge joining two sub-nets of other nodes within the Network. All paths within the two groups must pass through Node A, so it has a high betweenness centrality even though its degree centrality is low at a paltry 2.

The sections above describe the comparative benefits and disadvantages of the principal centrality measures, which this study seeks to deploy as companions to the newly evolved Covertness Index. While considering their strengths and weaknesses, it is worthwhile noting that the Covertness Index is defined along with a tie or an edge connecting a dyad (of nodes). Since this study focuses on mail exchanges between the constituent nodes of a dyad, the utility of all the centrality measures we discussed just prior is extremely narrow. We may use only that centrality attribute that best relates to a relationship between two nodes to exclude other properties within a network. This isolated need allows us to use the degree centrality as a companion attribute, and only that part of the degree centrality relates to the tie in the dyad. The concept of an Edge Vertex has been discussed at some length in the previous sections. Since the Edge Vector is limited only to the pair of nodes within the dyad in question, we can extract only the fraction of the degree centrality values that accrue to both the nodes from the interchange of mails (or information).

This analysis is best illustrated from the series of diagrams depicting the same fictional email network which had been described in an earlier section and which has been reproduced below again for the sake of clarity:

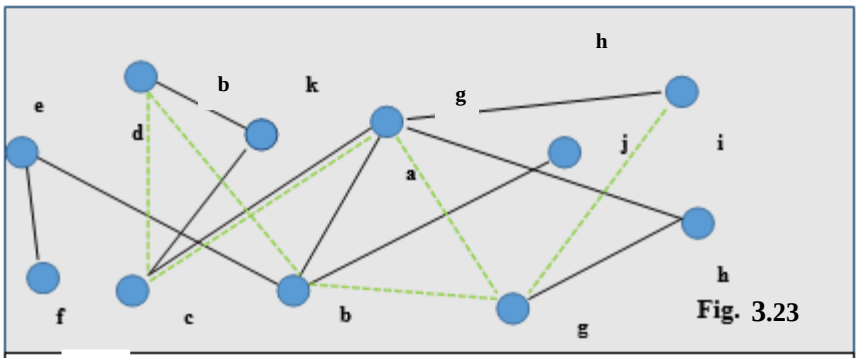


Fig. 3.23

Fig. 3.23 The above Social Network is an example of an e-mail exchange community. Each mail-id which is the equivalent of a node in a network is named as an alphabet. The solid lines that link the nodes represent actual exchanges of mail. The dotted lines in green indicate mails copied from one node to the other. The links or edges are not shown as directionally oriented since the study looks at undirected links. That is, it doesn't really matter who has mailed whom.

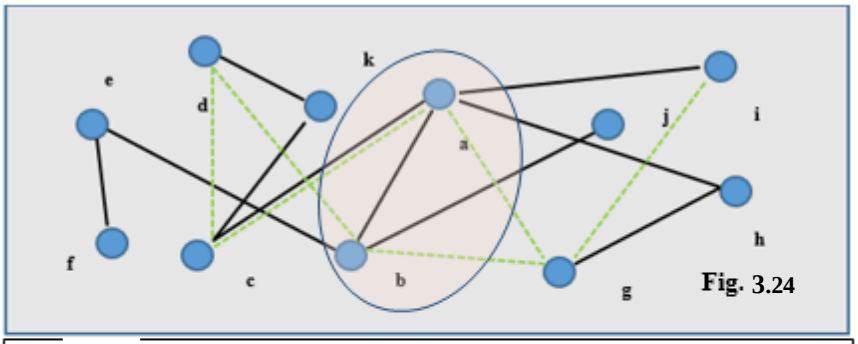


Fig. 3.24

Fig 3.24 The dyadic pair of interest i.e. the Nodes 'a' and 'b' have been highlighted in the figure above. There is an Edge Eab between the nodes defining the mail exchange relationship or tie between them. The dotted lines emanating from the pair of nodes are symbolic of the mails which have been copied out from the overall mail exchanges between nodes 'a' and 'b'.

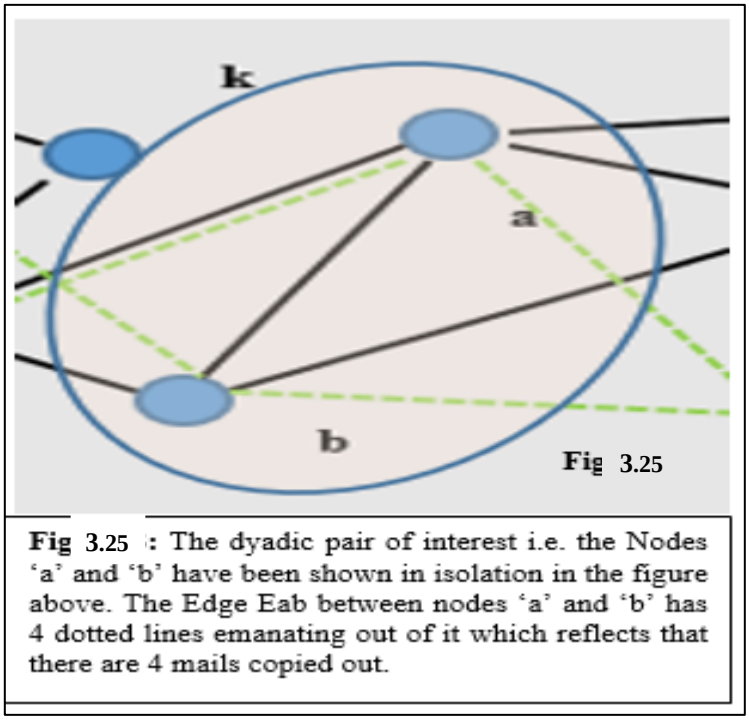


Fig 3.25

Fig 3.25 : The dyadic pair of interest i.e. the Nodes 'a' and 'b' have been shown in isolation in the figure above. The Edge Eab between nodes 'a' and 'b' has 4 dotted lines emanating out of it which reflects that there are 4 mails copied out.

From the figures above, it's clear that if we are to ascribe an index or a value based on the tie between the pair of nodes comprising the dyad, the use of a fraction of the degree centrality that is the number of edges connected to each of node *a* and *b* is the most accessible instrument amongst the centrality measures enumerated earlier, albeit, in a

modified mechanism. The way degree centrality is applied in these curated circumstances is explained in the passages below.

In the instant case, it's seen that node *a* is seen to have exchanged emails with nodes *c, b,* and *h*. Hence its degree centrality is 3. Similarly, *b* is seen to have exchanged emails with nodes *e* and *a,* and thus, its degree centrality comes to 2. It needs to be pointed out here that the degree centrality measure does not indicate the total number of emails exchanged by these nodes. For instance, let's suppose that nodes *a* and *c* have exchanged ten mails and nodes *a* and *b* have exchanged 20 mails, and nodes *a* and *h* have exchanged 15 mails. Then, node *a* is a part of 45 mail exchanges, which is not reflected by its degree centrality of 3. What is proposed here is to use the volumes of mail exchanged between two nodes as a single attribute of the tie or the edge between the nodes.

It's been analyzed in the passages previous that the Covertness Index matric envisaged has a few drawbacks. There is a need to decide which centrality measure to incorporate as a *companion attribute* to the existing covertness index to bring about a more balanced outcome. From the nature of the list-set contents that define an Edge-Vertex function, it's seen that the impact of degree centrality is the maximum. The number of emails exchanged between the constituent nodes of the pair is a fraction of the total mails received by either of the nodes from the nodes they have ties with, which is a form of degree centrality, and the emails exchanged between only the two of them is a fraction of the degree. No other centrality measure, be it eigenvalue or closeness or betweenness centrality, has such a proximal impact upon the performance of edges, and this doesn't necessarily mean that we can't use any other centrality; conversely, if any dyad is situated strategically between two discrete components of the network, we may use betweenness. Closeness or Eigenvalue centrality can be employed after some knowledge is acquired about the influence of the nodes in a network, about which this study presumes agnosticism. The only attribute the study has prior awareness about is the number of mails and copies exchanged, which is nothing but a facet of degree centrality alone. After identifying nodes as covert or otherwise within a network, the other structural aspects of the nodes (or edges) such as

their placement, proximity to influential nodes or subnets, etc. can be looked at from a more balanced perspective.

Suppose we look at the edge between nodes a and b as an entity akin to a single vertex. In that case, its degree centrality can be estimated as the total number of mails existing in the tie between nodes *a* and *b*, which will be a fraction of the total mails sent or received by nodes *a* and *b* in their capacity. It needs to be noted that for this study, the mails copied out from a dyad (or edge) do not constitute a part of the degree centrality of the edge.

In light of these arguments, in respect of an Edge Vertex, where the defining attribute is the tie or relationship between a dyad's constituent nodes, two candidate attributes are similar to the degree centrality. One is the total number of emails exchanged between the dyad's constituent nodes (irrespective of the mail exchange direction). The second is the total number of emails copied to nodes outside of the pair. As discussed earlier, the number of mails opens a window into the volume of information interchanged between the social network actors. There are good arguments that run counter to this in as much as two covert actors may try and hide their interactions through deception (such as the use of covert channels) and make active efforts to hide the ties between themselves by communicating as little and as cryptically as possible. In such circumstances, the question that arises is whether to ascribe any importance to the number of mails alongside the ties' covertness index. The answer is an emphatic yes, as the analysis below shows.

It needs to be recalled that a substantial part of the network-related information is absent, missing, or deliberately kept hidden (Sparrow,1991). The addition of the number of mail exchanges between a node as an extra attribute in the covertness analysis bolsters the information we have about the relationship's nature. Though this study is limited to a non-longitudinal analysis of a network (i.e., looking only at the ties of a snapshot of the network or just a cross-section of it), the framework we presume to be possessing for our study is the accumulated information about the network, e.g., the number of emails that have been exchanged from the inception of the network (or at least since we began looking at its evolution). Suppose two actors have been partners in a covert enterprise. In that case, it

stands to reason that we look at their interaction in its entirety, given that the study aims to identify the confinement of information. Even if we presume for the sake of argument that these two actors were enjoying a benign relationship till a point in time and then commenced their attempts to transform their ties into a covert one, the pre-existing information will be useful.

Pre-existing ties' crucial importance has focused on a recent study in corruption-related social networks (Diviák, Dijkstra & Snijders, 2019). Pre-existing ties are a measure of trust between actors in a social network. In the research on covert networks, there has been a great emphasis on the importance of ties based on trust[41]. (Erickson, 1981; Krebs, 2002; Milward and Raab, 2006; Oliver et al. ,2014; Robins, 2009; van der Hulst, 2011). Pre-existing ties, meaning ties established before the criminal act itself, may be crucial sources of trust, making their analysis important (Morselli and Roy 2008). In the literature on social networks research, pre-existing ties include marriage, sharing the same classes in a university or even a karate class, or membership in a board of directors. The existence of such a tie does not by itself create trust between the two actors. It may only potentially facilitate the future build-up of trust between the same set of actors. Still, by ignoring such ties, there is an unacceptable risk of omitting vitally relevant information.

In an email network, like the one which is the subject of scrutiny in this study, the closest attribute that we have to pre-existing ties is the corpus of emails exchanged between two actors (or nodes) and, like pre-existing ties, the existing mail correspondence between the actors is a form of kinship relations, friendships, or relations based on shared ideology or shared affiliation to the same organizations or institutions. The question that now arises is, to what extent was the connivance dependent on the mail correspondence? This may well be answered by the outcomes(e.g., the conspiracy to commit insider trading on Enron shares, for instance). Ties in this dimension capture the notion of Krebs' (2002) and Everton's (2012) dimension of trust, Papachristos and Smith (2014; Smith and

[41]Trust may be defined as the expectation of reciprocation and of not breaking the confidentiality in a covert environment (Campana and Varese 2013; von Lampe and Ole Johansen 2004).

Papachristos 2016) personal ties and partly legal ties, and Gerdes' (2015a) links of training, ideology, family and friendship (Diviák et al., 2019).

In the earlier arguments' backdrop, the Covertness attribute is modified to include the number of emails exchanged as a multiplicand. However, instead of using the number itself, this study suggests the use of its logarithmic value. In large and busy networks, the number of mail exchanges can be huge, to the extent of millions, making the calculation unwieldy and unbounded. Using the logarithm reduces the index value to manageable proportions.

This modification in calculating the Covertness Index is tried on the example cited earlier and may revisit.

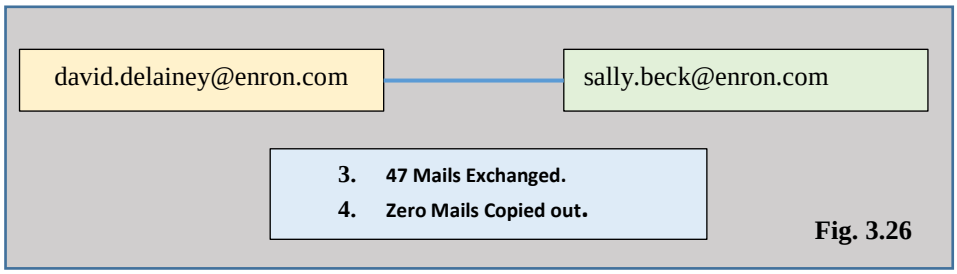


Fig. 3.26

The Covertness Index of the above pair is calculated as follows:

1. Number of mails exchanged = 47.
2. Number of mails copied out = 0.
3. The ratio of mails copied out to mails exchanged = 0.
4. Overtness Index = Ratio above = 0.
5. Covertness Index = 1 – (Overtness Index) = 1 – 0 =1.
6. Modified Covertness Index = 1 x Log (47) = 1.67.

We may compare this with the Covertness Index calculated upon the tie between two other employees of our interest: Jeff Dasovich and Angela Schwarz. A pictorial depiction of the mail exchange status between them and the copies mark out by the pair to other nodes is shown in Figure 3.27 below.

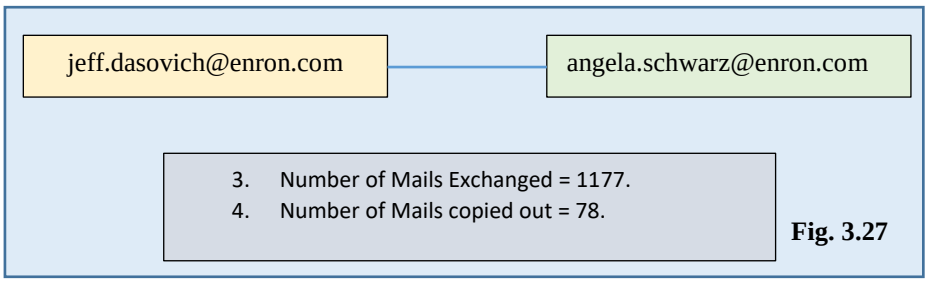


Fig. 3.27

The modified Covertness Index of the above pair is calculated as follows:

1. Number of mails exchanged = 1177.
2. Number of mails copied out = 78.
3. The ratio of mails copied out to mails exchanged = 0.0663.
4. Overtness Index = Ratio above = 0.0663.
5. Covertness Index = 1 – (Overtness Index) = 1 – 0.0663 =0.9337.
6. Modified Covertness Index = 0.9337 x Log (1177) = 2.87.

Thus, the Covertness Index of both pairs is reversed now in terms of value. Earlier, the edge joining the pair of David Delainey and Sally Beck had a Covertness Index of 1.0, which was higher than the Covertness Index of the second pair, which we had considered, namely, Jeff Dasovich and Angela Schwarz, which came to 0.93 approximately. After modifying the formula, the edge connecting David Delainey and Sally Beck change to 1.67, which is less than the modified value of the Covertness Index of the edge tying the second pair, namely, Jeff Dasovich and Angela Schwarz (which is now 2.87). This way, we successfully factored in the number of emails exchanged between a pair of actors in the network and rationalized the covertness rankings.

The graphs in the figures below (Figures 3.28 and 3.29) represent an improvement in ranking based on Covertness Index after applying the modification suggested earlier. The lift achieved is considerable. Though the cumulative count of Edges of Interest (EoIs) match up towards the end, the prevalence of the EoIs till the first 10,000 ranks is

measurably denser, which improves the chances of detecting such edges. The horizontal axis in the first of the figures shows the incremental ranking of all dyads numerically. The second figure has the incremental values arranged logarithmically for more clarity.

The same graphs are repeated in the figures on the page following (Figures 3.30 and 3.31) with the addition of the prevalence of the Edges of Interest (EoIs) based on a presumed uniform distribution throughout the set of all edges in the network. The enhancements brought about by applying the Covertness Index of the ties, both modified and unmodified, stand out in stark contrast. The uniform distribution network model, which has been extensively used in this dissertation, can be described as being akin to the concept of a *random graph*[42] in which the covert edges are presumed to be equally distributed over the entire set of edges. From a mathematical perspective, random graphs are used to answer actual graphs or networks' properties. Its practical applications are found in all areas in which complex networks need to be modeled – many random graph models are thus known, mirroring the diverse types of complex networks encountered in different areas. Here, the uniform distribution model has been used for comparing the performance of the models proposed in the study.

[42] The concept of a random graph was introduced through a seminal article published in 1959 by Erdős and Rényi.The construction of a random graph is defined by Newman, Watts and Strogatz (2002, p.2567) thus – "One takes some number *N* of nodes or "vertices" and places connections or "edges" between them, such that each pair of vertices *i, j* has a connecting edge with independent probability *p.* "…. This example is one of the simplest models of a network there is, and is certainly the best studied; the random graph has become a cornerstone of the discipline known as discrete mathematics."

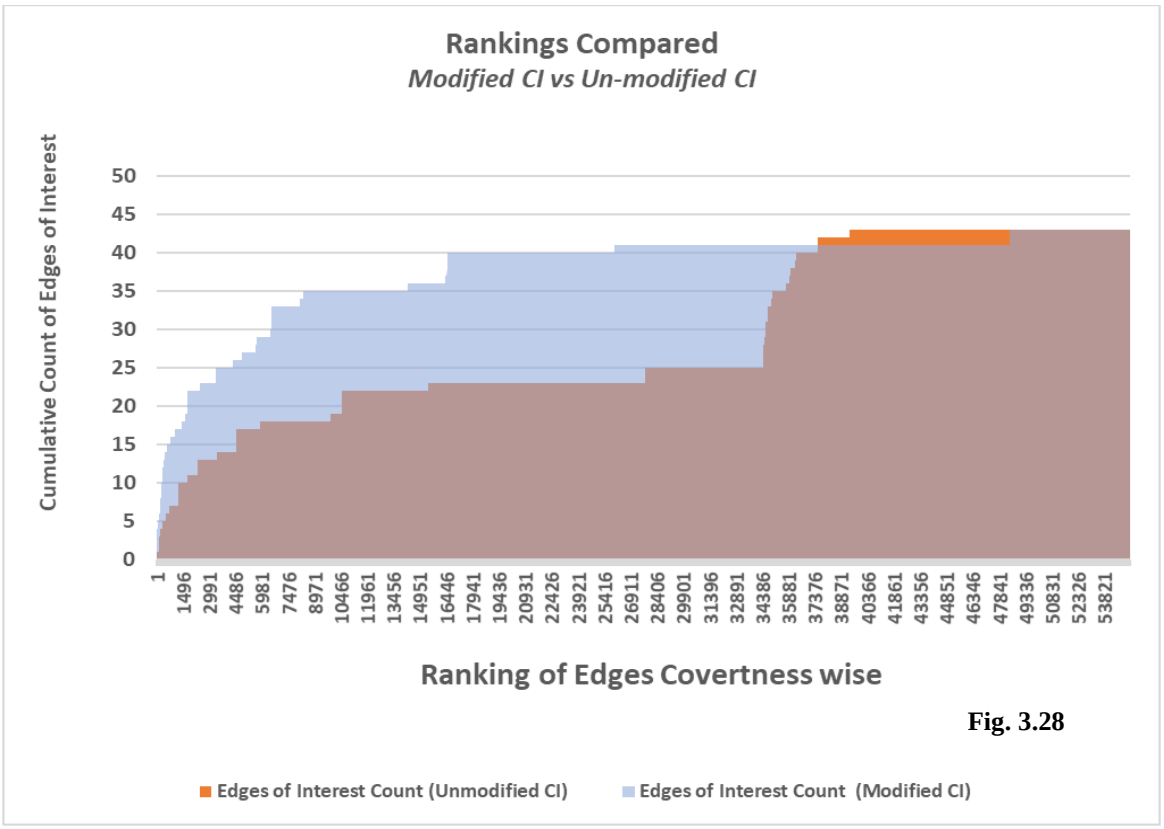


Fig. 3.28

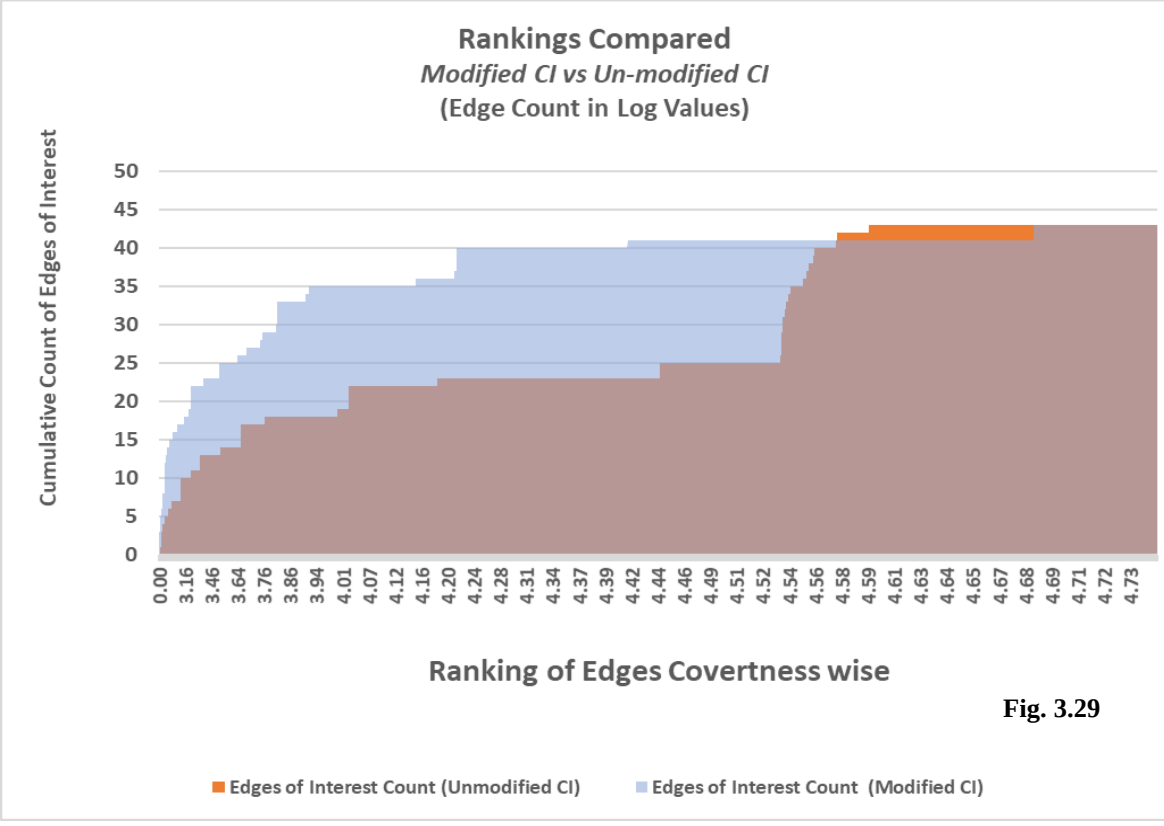


Fig. 3.29

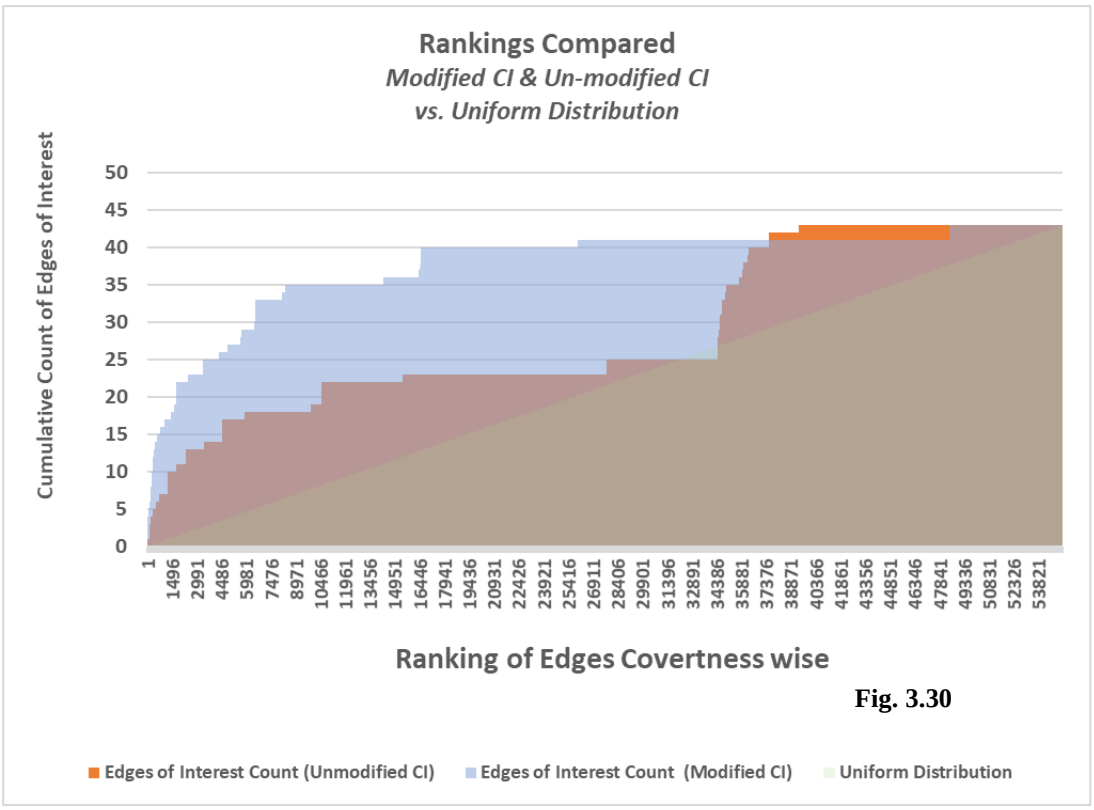


Fig. 3.30

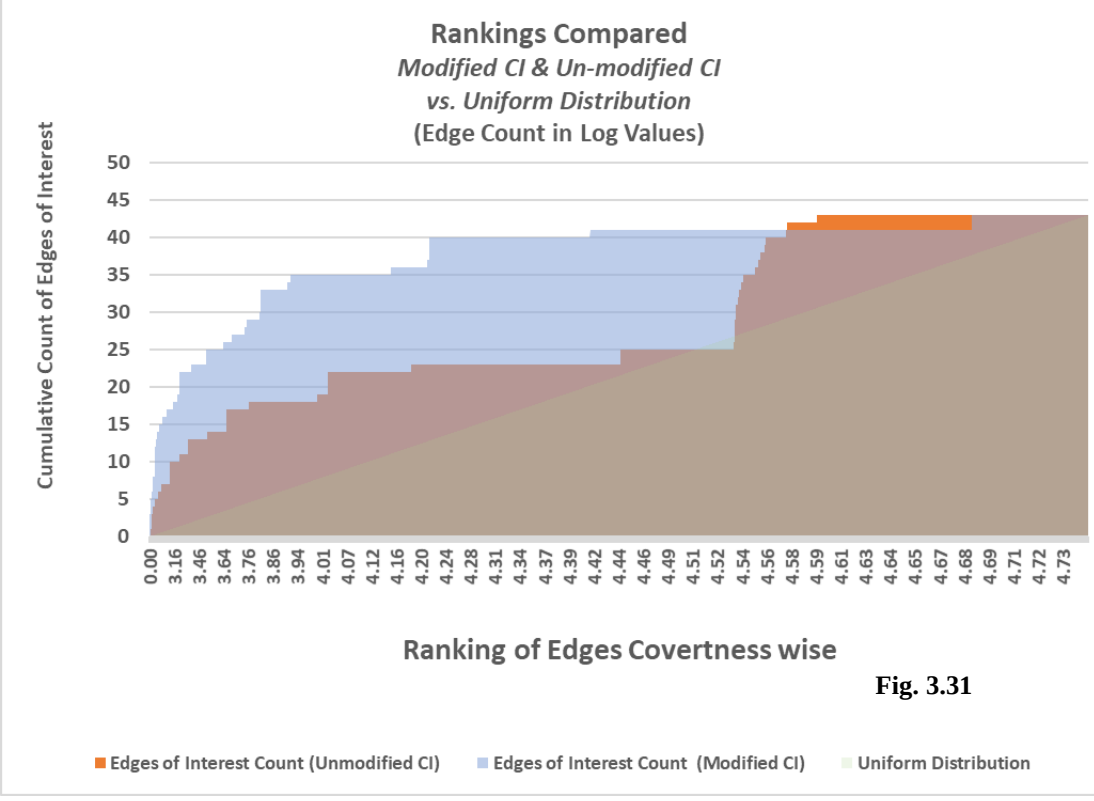


Fig. 3.31

The revised rankings of the Dyads based on the modified scores of their Covertness Indexes are shown in the tables (Table 3.9 and Table 3.10) below. The first table reveals a snapshot of the overall picture of the ranking (top 50ranks), whereas the next one highlights the rankings of the top-ranked Edges of Interest within the top ranks of all edges.

From	To	Modified CI	Ranking
kay.mann@enron.com	suzanne.adams@enron.com	3.07	1
evelyn.metoyer@enron.com	kate.symes@enron.com	2.91	2
jeff.dasovich@enron.com	angela.schwarz@enron.com	2.87	3
jeff.dasovich@enron.com	steven.kean@enron.com	2.86	4
jeff.dasovich@enron.com	beverly.aden@enron.com	2.85	5
kerri.thompson@enron.com	kate.symes@enron.com	2.77	6
veronica.espinoza@enron.com	william.bradford@enron.com	2.76	7
veronica.espinoza@enron.com	debbie.brackett@enron.com	2.75	8
bill.iii@enron.com	portland.shift@enron.com	2.73	9
kristin.walsh@enron.com	louise.kitchen@enron.com	2.67	10
kate.symes@enron.com	sharen.cason@enron.com	2.59	11
kate.symes@enron.com	stephanie.piwetz@enron.com	2.57	12
kate.symes@enron.com	kimberly.hundl@enron.com	2.39	13
kay.mann@enron.com	heather.kroll@enron.com	2.35	14
chris.germany@enron.com	alvin.thompson@enron.com	2.35	15
kenneth.thibodeaux@enron.com	john.allison@enron.com	2.31	16
bill.williams@enron.com	todd.bland@enron.com	2.29	17
susan.scott@enron.com	ted.noble@enron.com	2.26	18
pete.davis@enron.com	craig.dean@enron.com	2.26	19
john.arnold@enron.com	ina.rangel@enron.com	2.26	20
mary.hain@enron.com	phillip.allen@enron.com	2.25	21
brant.reves@enron.com	susan.bailey@enron.com	2.24	22
robin.rodrigue@enron.com	becky.pitre@enron.com	2.24	23
karen.denne@enron.com	kenneth.lay@enron.com	2.22	24
kysa.alport@enron.com	tom.alonso@enron.com	2.20	25
pete.davis@enron.com	bert.meyers@enron.com	2.19	26
victor.lamadrid@enron.com	f..brawner@enron.com	2.17	27
pete.davis@enron.com	bill.williams.iii@enron.com	2.16	28
steven.kean@enron.com	kenneth.lay@enron.com	2.16	29
victor.lamadrid@enron.com	chuck.ames@enron.com	2.16	30
chris.germany@enron.com	edward.terry@enron.com	2.16	31
daren.farmer@enron.com	megan.parker@enron.com	2.15	32
david.delainey@enron.com	rob.milnthorp@enron.com	2.15	33
kenny.soignet@enron.com	berney.aucoin@enron.com	2.15	34
kysa.alport@enron.com	robert.badeer@enron.com	2.15	35
kenny.soignet@enron.com	john.arnold@enron.com	2.15	36
kenny.soignet@enron.com	phillip.allen@enron.com	2.15	37
kay.chapman@enron.com	raymond.bowen@enron.com	2.13	38
carla.hoffman@enron.com	jeff.richter@enron.com	2.11	39
debra.davidson@enron.com	portland.desk@enron.com	2.11	40
robin.rodrigue@enron.com	kathy.reeves@enron.com	2.11	41
cara.semperger@enron.com	portland.shift@enron.com	2.08	42
john.forney@enron.com	portland.shift@enron.com	2.07	43
jeff.dasovich@enron.com	mpalmer@enron.com	2.07	44
rosalee.fleming@enron.com	cliff.baxter@enron.com	2.06	45
pete.davis@enron.com	albert.meyers@enron.com	2.05	46
jennifer.mcquade@enron.com	andy.zipper@enron.com	2.02	47
kenneth.thibodeaux@enron.com	beth.apollo@enron.com	2.01	48
jeffrey.gossett@enron.com	kam.keiser@enron.com	2.00	49
ray.alvarez@enron.com	tim.belden@enron.com	2.00	50

Table3.9 Rankings of edges after applying the Modified Covertness Index

From	To	Modified CI	Ranking
kay.mann@enron.com	suzanne.adams@enron.com	3.07	1
evelyn.metoyer@enron.com	kate.symes@enron.com	2.91	2
jeff.dasovich@enron.com	angela.schwarz@enron.com	2.87	3
jeff.dasovich@enron.com	steven.kean@enron.com	2.86	4
jeff.dasovich@enron.com	beverly.aden@enron.com	2.85	5
kerri.thompson@enron.com	kate.symes@enron.com	2.77	6
veronica.espinoza@enron.com	william.bradford@enron.com	2.76	7
veronica.espinoza@enron.com	debbie.brackett@enron.com	2.75	8
bill.iii@enron.com	portland.shift@enron.com	2.73	9
kristin.walsh@enron.com	louise.kitchen@enron.com	2.67	10
kate.symes@enron.com	sharen.cason@enron.com	2.59	11
kate.symes@enron.com	stephanie.piwetz@enron.com	2.57	12
kate.symes@enron.com	kimberly.hundl@enron.com	2.39	13
kay.mann@enron.com	heather.kroll@enron.com	2.35	14
chris.germany@enron.com	alvin.thompson@enron.com	2.35	15
kenneth.thibodeaux@enron.com	john.allison@enron.com	2.31	16
bill.williams@enron.com	todd.bland@enron.com	2.29	17
susan.scott@enron.com	ted.noble@enron.com	2.26	18
pete.davis@enron.com	craig.dean@enron.com	2.26	19
john.arnold@enron.com	ina.rangel@enron.com	2.26	20
mary.hain@enron.com	phillip.allen@enron.com	2.25	21
brant.reves@enron.com	susan.bailey@enron.com	2.24	22
robin.rodrigue@enron.com	becky.pitre@enron.com	2.24	23
karen.denne@enron.com	kenneth.lay@enron.com	2.22	24
kysa.alport@enron.com	tom.alonso@enron.com	2.20	25
pete.davis@enron.com	bert.meyers@enron.com	2.19	26
victor.lamadrid@enron.com	f..brawner@enron.com	2.17	27
pete.davis@enron.com	bill.williams.iii@enron.com	2.16	28
steven.kean@enron.com	kenneth.lay@enron.com	2.16	29
victor.lamadrid@enron.com	chuck.ames@enron.com	2.16	30
chris.germany@enron.com	edward.terry@enron.com	2.16	31
daren.farmer@enron.com	megan.parker@enron.com	2.15	32
david.delainey@enron.com	rob.milnthorp@enron.com	2.15	33
kenny.soignet@enron.com	berney.aucoin@enron.com	2.15	34
kysa.alport@enron.com	robert.badeer@enron.com	2.15	35
kenny.soignet@enron.com	john.arnold@enron.com	2.15	36
kenny.soignet@enron.com	phillip.allen@enron.com	2.15	37
kay.chapman@enron.com	raymond.bowen@enron.com	2.13	38
carla.hoffman@enron.com	jeff.richter@enron.com	2.11	39
debra.davidson@enron.com	portland.desk@enron.com	2.11	40
robin.rodrigue@enron.com	kathy.reeves@enron.com	2.11	41
cara.semperger@enron.com	portland.shift@enron.com	2.08	42
john.forney@enron.com	portland.shift@enron.com	2.07	43
jeff.dasovich@enron.com	mpalmer@enron.com	2.07	44
rosalee.fleming@enron.com	cliff.baxter@enron.com	2.06	45
pete.davis@enron.com	albert.meyers@enron.com	2.05	46
jennifer.mcquade@enron.com	andy.zipper@enron.com	2.02	47
kenneth.thibodeaux@enron.com	beth.apollo@enron.com	2.01	48
jeffrey.gossett@enron.com	kam.keiser@enron.com	2.00	49
ray.alvarez@enron.com	tim.belden@enron.com	2.00	50

Table 3.10 Prevalence of EoIs within the top 50 rankings of edges after applying the Modified Covertness Index.

From	To	CI Value	Ranking
kate.symes@enron.com	lester.rawson@enron.com	1.99	1
christi.nicolay@enron.com	jeff.brown@enron.com	1.99	2
scott.neal@enron.com	kimberly.brown@enron.com	1.99	3
chris.germany@enron.com	crystal.hyde@enron.com	1.98	4
ray.alvarez@enron.com	mike.swerzbin@enron.com	1.98	5
david.delainey@enron.com	tim.belden@enron.com	1.98	6
kysa.alport@enron.com	shift.dl-portland@enron.com	1.98	7
m..forney@enron.com	joe.errigo@enron.com	1.97	8
kay.mann@enron.com	david.fairley@enron.com	1.96	9
m..forney@enron.com	joe.capasso@enron.com	1.95	10
dan.hyvl@enron.com	kim.ward@enron.com	1.95	11
jeffrey.keeler@enron.com	stanley.horton@enron.com	1.95	12
kay.chapman@enron.com	janet.dietrich@enron.com	1.93	13
shona.wilson@enron.com	leslie.reeves@enron.com	1.93	14
cindy.derecskey@enron.com	richard.shapiro@enron.com	1.92	15
mary.hain@enron.com	steve.c.hall@enron.com	1.92	16
chris.germany@enron.com	dick.jenkins@enron.com	1.91	17
becky.spencer@enron.com	samantha.boyd@enron.com	1.90	18
victor.lamadrid@enron.com	kevin.alvarado@enron.com	1.90	19
lexi.elliott@enron.com	mark.lindsey@enron.com	1.90	20
kevin.hyatt@enron.com	market.team@enron.com	1.89	21
kay.mann@enron.com	matthew.berry@enron.com	1.89	22
drew.fossum@enron.com	lorraine.lindberg@enron.com	1.89	23
hector.mcloughlin@enron.com	bob.hall@enron.com	1.89	24
mark.mccoy@enron.com	stacey.neuweiler@enron.com	1.88	25
vincent.strohmeyer@enron.com	jerry.peters@enron.com	1.87	26
alan.comnes@enron.com	robert.badeer@enron.com	1.86	27
kate.symes@enron.com	mark.confer@enron.com	1.86	28
ray.alvarez@enron.com	michael.driscoll@enron.com	1.86	29
scott.neal@enron.com	dick.jenkins@enron.com	1.85	30
brant.reves@enron.com	stephanie.panus@enron.com	1.85	31
susan.mara@enron.com	mark.palmer@enron.com	1.85	32
debra.perlingiere@enron.com	sylvia.pollan@enron.com	1.84	33
chris.germany@enron.com	meredith.mitchell@enron.com	1.84	34
chris.germany@enron.com	jesse.villarreal@enron.com	1.84	35
jeff.dasovich@enron.com	jdasovic@enron.com	1.83	36
jinsung.myung@enron.com	benjamin.rogers@enron.com	1.83	37
susan.scott@enron.com	kevin.hyatt@enron.com	1.83	38
rebecca.cantrell@enron.com	barry.tycholiz@enron.com	1.82	39
janet.butler@enron.com	daniel.allegretti@enron.com	1.82	40
caroline.abramo@enron.com	sara.shackleton@enron.com	1.82	41
chris.germany@enron.com	judy.townsend@enron.com	1.82	42
kay.young@enron.com	gerald.nemec@enron.com	1.82	43
david.delainey@enron.com	w.duran@enron.com	1.82	44
allison.navin@enron.com	steven.kean@enron.com	1.80	45
ginger.dernehl@enron.com	mark.palmer@enron.com	1.79	46
taffy.milligan@enron.com	alan.aronowitz@enron.com	1.78	47
eric.benson@enron.com	richard.shapiro@enron.com	1.78	48
stanley.horton@enron.com	peggy.fowler@enron.com	1.78	49
lorna.brennan@enron.com	w..mcgowan@enron.com	1.78	50

Table 3.11 Prevalence of EoIs within the next 50 rankings of edges after applying the Modified Covertness Index.

The results obtained after applying the Modified Covertness Index to the ties between all the network dyads yield better results. Tables 3.12 and 3.13 below reflect how the edges of interest (EoIs) ' rankings transform after the rectification. Earlier, when the Covertness Index was applied, a skewed distribution favoring dyads hadn't copied out any of the mails they had exchanged. The ranking of such 'perfectly opaque' edges dominated the front half of the rankings. The pairs that had sizeable mails' exchanges but had some of these copied out to nodes outside the pairs scored far lesser on the rankings.

The second of the tables below shows the comparative rankings of the edges of interest (EoIs) before and after applying the modified formula for the ties between the actors who focus on study in the ENRON mail network. The pairs which had sizeable exchanges of mails now move up in the rankings. To the extent that many of the pairs which were languishing at the bottom of the rankings table following the application of the initial (unmodified version) of the Covertness Index appear at the very top after modifying the formula on the above lines. The upending of the rankings brings to the fore the importance of the number of emails exchanged (incidence) in calculating the covertness of ties

Sl. No.	From	To	Modified CI	Ranking after Modifying the CI
1	jeff.dasovich@enron.com	angela.schwarz@enron.com	2.87	4
2	jeff.dasovich@enron.com	steven.kean@enron.com	2.86	5
3	karen.denne@enron.com	kenneth.lay@enron.com	2.22	25
4	steven.kean@enron.com	kenneth.lay@enron.com	2.16	30
5	david.delainey@enron.com	tim.belden@enron.com	1.98	57
6	steven.kean@enron.com	jeff.skilling@enron.com	1.77	105
7	vince.kaminski@enron.com	jeff.skilling@enron.com	1.68	168
8	david.delainey@enron.com	sally.beck@enron.com	1.67	173
9	maureen.mcvicker@enron.com	david.delainey@enron.com	1.61	285
10	maureen.mcvicker@enron.com	kenneth.lay@enron.com	1.61	286
11	jeff.dasovich@enron.com	maureen.mcvicker@enron.com	1.60	327
12	rob.bradley@enron.com	kenneth.lay@enron.com	1.60	328
13	jeff.dasovich@enron.com	louise.kitchen@enron.com	1.57	393
14	david.delainey@enron.com	kenneth.lay@enron.com	1.55	464
15	maureen.mcvicker@enron.com	andrew.fastow@enron.com	1.51	566
16	maureen.mcvicker@enron.com	richard.causey@enron.com	1.45	775
17	karen.denne@enron.com	richard.causey@enron.com	1.38	1029
18	vince.kaminski@enron.com	david.delainey@enron.com	1.34	1395
19	jeffrey.shankman@enron.com	jeff.skilling@enron.com	1.30	1629
20	karen.denne@enron.com	david.delainey@enron.com	1.26	1766
21	maureen.mcvicker@enron.com	rob.bradley@enron.com	1.26	1767
22	vince.kaminski@enron.com	andrew.fastow@enron.com	1.26	1768
23	maureen.mcvicker@enron.com	tim.belden@enron.com	1.18	2465
24	maureen.mcvicker@enron.com	jeff.skilling@enron.com	1.08	3336
25	tim.belden@enron.com	jeff.skilling@enron.com	1.08	3337
26	james.steffes@enron.com	maureen.mcvicker@enron.com	1.06	4317
27	vince.kaminski@enron.com	tim.belden@enron.com	1.00	4849
28	maureen.mcvicker@enron.com	steven.kean@enron.com	0.98	5620
29	louise.kitchen@enron.com	vince.kaminski@enron.com	0.98	5695
30	david.delainey@enron.com	andrew.fastow@enron.com	0.92	6464
31	jeff.dasovich@enron.com	jeff_dasovich@enron.com	0.90	6529
32	sally.beck@enron.com	steven.kean@enron.com	0.90	6530
33	vince.kaminski@enron.com	kenneth.lay@enron.com	0.90	6531
34	steven.kean@enron.com	steven.kean@enron.com	0.87	8144
35	james.steffes@enron.com	steven.kean@enron.com	0.85	8312
36	louise.kitchen@enron.com	kenneth.lay@enron.com	0.70	14223
37	andrew.fastow@enron.com	louise.kitchen@enron.com	0.60	16385
38	andrew.fastow@enron.com	jeff.skilling@enron.com	0.60	16476
39	kenneth.lay@enron.com	tim.belden@enron.com	0.60	16477
40	louise.kitchen@enron.com	jeff.skilling@enron.com	0.60	16478
41	jeff.dasovich@enron.com	jeff.skilling@enron.com	0.48	25958
42	david.delainey@enron.com	angela.schwarz@enron.com	0.00	48444

Table 3.12 Changed Rankings of the edges of interest (EoIs) after the Covertness Index is modified.

Sl. No.	From	To	Modified CI	Ranking after Modifying the CI	Ranking before Modification
1	jeff.dasovich@enron.com	angela.schwarz@enron.com	2.87	4	34455
2	jeff.dasovich@enron.com	steven.kean@enron.com	2.86	5	34465
3	karen.denne@enron.com	kenneth.lay@enron.com	2.22	25	34470
4	steven.kean@enron.com	kenneth.lay@enron.com	2.16	30	34423
5	david.delainey@enron.com	tim.belden@enron.com	1.98	57	34888
6	steven.kean@enron.com	jeff.skilling@enron.com	1.77	105	34723
7	vince.kaminski@enron.com	jeff.skilling@enron.com	1.68	168	34661
8	david.delainey@enron.com	sally.beck@enron.com	1.67	173	32
9	maureen.mcvicker@enron.com	david.delainey@enron.com	1.61	285	122
10	maureen.mcvicker@enron.com	kenneth.lay@enron.com	1.61	286	123
11	jeff.dasovich@enron.com	maureen.mcvicker@enron.con	1.60	327	36221
12	rob.bradley@enron.com	kenneth.lay@enron.com	1.60	328	34533
13	jeff.dasovich@enron.com	louise.kitchen@enron.com	1.57	393	204
14	david.delainey@enron.com	kenneth.lay@enron.com	1.55	464	34547
15	maureen.mcvicker@enron.com	andrew.fastow@enron.com	1.51	566	331
16	maureen.mcvicker@enron.com	richard.causey@enron.com	1.45	775	505
17	karen.denne@enron.com	richard.causey@enron.com	1.38	1029	681
18	vince.kaminski@enron.com	david.delainey@enron.com	1.34	1395	35987
19	jeffrey.shankman@enron.com	jeff.skilling@enron.com	1.30	1629	34973
20	karen.denne@enron.com	david.delainey@enron.com	1.26	1766	1214
21	maureen.mcvicker@enron.com	rob.bradley@enron.com	1.26	1767	1215
22	vince.kaminski@enron.com	andrew.fastow@enron.com	1.26	1768	1216
23	maureen.mcvicker@enron.com	tim.belden@enron.com	1.18	2465	1729
24	maureen.mcvicker@enron.com	jeff.skilling@enron.com	1.08	3336	2292
25	tim.belden@enron.com	jeff.skilling@enron.com	1.08	3337	2293
26	james.steffes@enron.com	maureen.mcvicker@enron.con	1.06	4317	37559
27	vince.kaminski@enron.com	tim.belden@enron.com	1.00	4849	3421
28	maureen.mcvicker@enron.com	steven.kean@enron.com	0.98	5620	39330
29	louise.kitchen@enron.com	vince.kaminski@enron.com	0.98	5695	35702
30	david.delainey@enron.com	andrew.fastow@enron.com	0.92	6464	36317
31	jeff.dasovich@enron.com	jeff_dasovich@enron.com	0.90	6529	4540
32	sally.beck@enron.com	steven.kean@enron.com	0.90	6530	4541
33	vince.kaminski@enron.com	kenneth.lay@enron.com	0.90	6531	4542
34	steven.kean@enron.com	steven.kean@enron.com	0.87	8144	37527
35	james.steffes@enron.com	steven.kean@enron.com	0.85	8312	5849
36	louise.kitchen@enron.com	kenneth.lay@enron.com	0.70	14223	9876
37	andrew.fastow@enron.com	louise.kitchen@enron.com	0.60	16385	35894
38	andrew.fastow@enron.com	jeff.skilling@enron.com	0.60	16476	10496
39	kenneth.lay@enron.com	tim.belden@enron.com	0.60	16477	10497
40	louise.kitchen@enron.com	jeff.skilling@enron.com	0.60	16478	10498
41	jeff.dasovich@enron.com	jeff.skilling@enron.com	0.48	25958	15406
42	david.delainey@enron.com	angela.schwarz@enron.com	0.00	48444	34423
43	richard.causey@enron.com	jeff.skilling@enron.com	0.00	48445	27724

Table 3.13 Rankings of the edges of interest (EoIs) compared with previous rankings after the Covertness Index is modified.

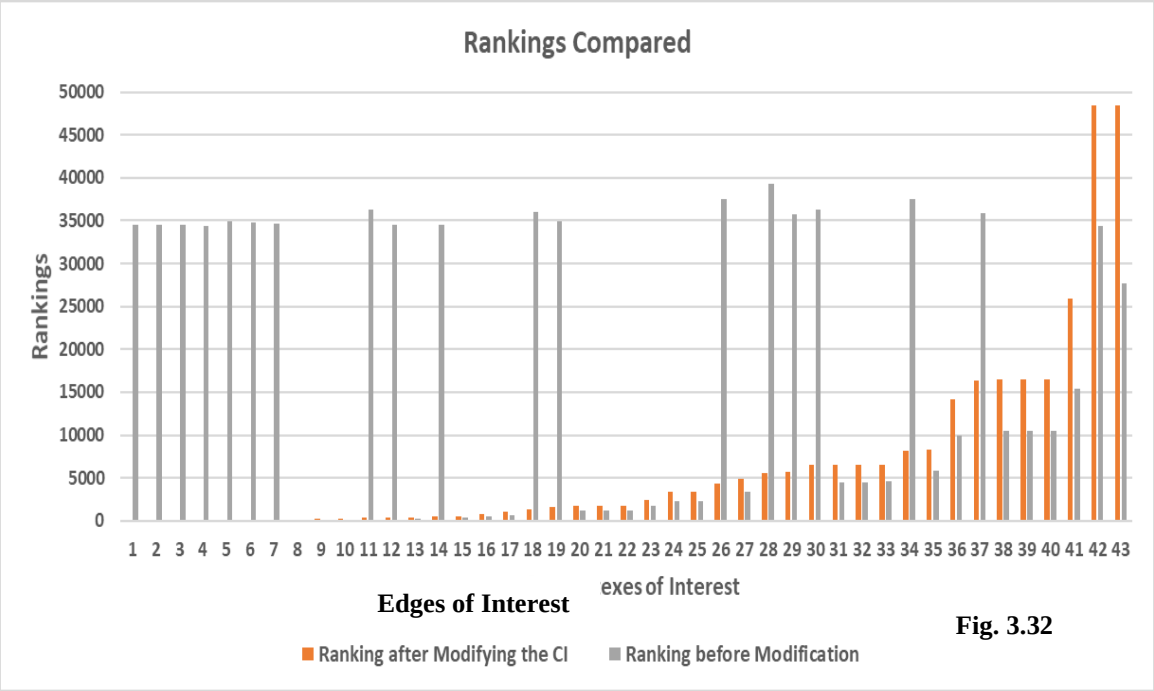


Fig. 3.32

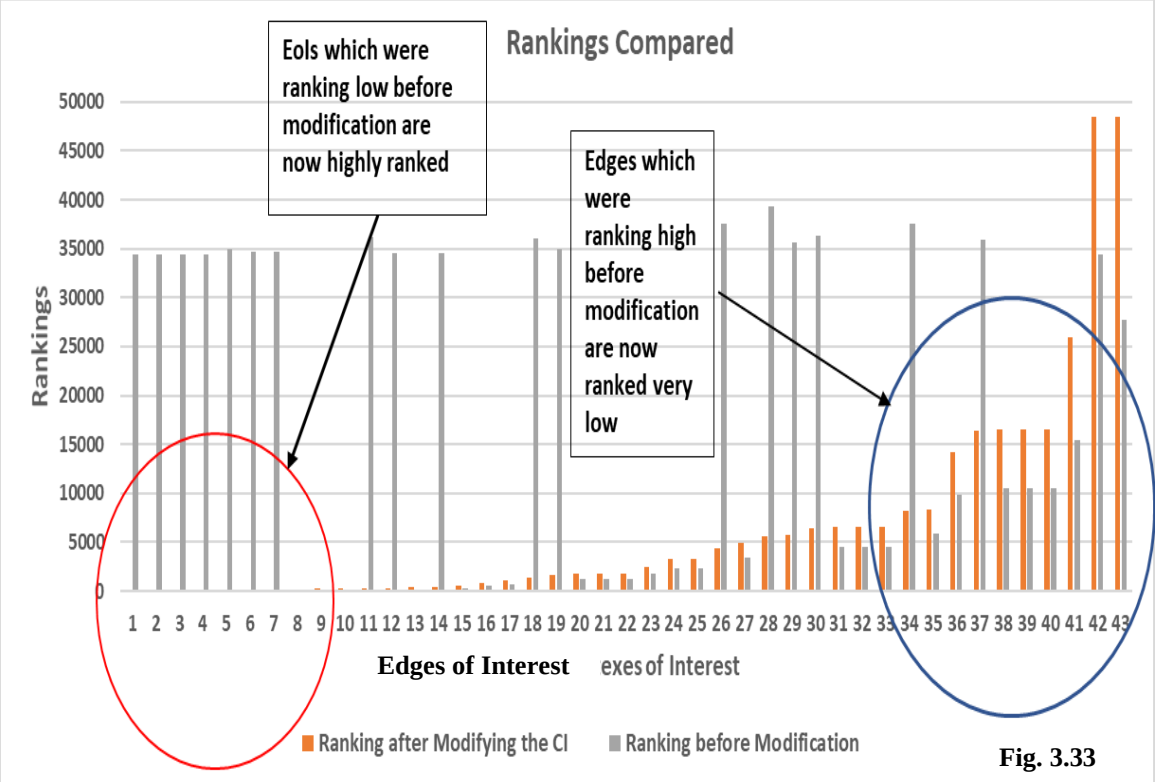


Fig. 3.33

The two graphs are shown earlier emphasize the enhancement in the covertness profile of the edges of interest (EoIs) upon applying the originally modified formula. In parallel, the figures also draw attention to the importance of factoring in the total amount of information exchanged between the constituent nodes in a dyad.

A scatterplot diagram showing the top ranks' skewness towards dyads having no mails copied out from amongst the mail exchanges between their constituent nodes is shown in Figure3.34 below. The plot in Figure 3.34 reflects the correlation between covertness-based rankings of the edges and the number of mails that the constituent nodes of the dyad the edges belong to have sent out to the other nodes outside the pair. A clear correlation is visible from the scatterplot. To further elucidate the trend, another version of the same plot is presented in Figure 3.35, highlighting that all the dyads that have not sent out any copies are concentrated in the top ranks (please see the encircled area in the same graph). This trend is somewhat of an obstacle since many node pairs may have exchanged very few mails due to genuine reasons and wouldn’t have found the opportunity or need to mark them out as copies. Quite simply, these node pairs are *false positives* blotting out the *true positive* covert pairs of nodes dyads. To an extent, false positives will remain since many innocuous pairs of nodes will share characteristics with genuinely covert node pairs and may mimic them. But the numbers of false positives need to be further pruned for effective analysis and better surveillance focus. At this stage where pruning is needed, the role of the modified covertness metric comes into play. The second scatterplot between the mails copied and the covertness ranking in the context of a modified Covertness Index is also shown in Figure 3.36.

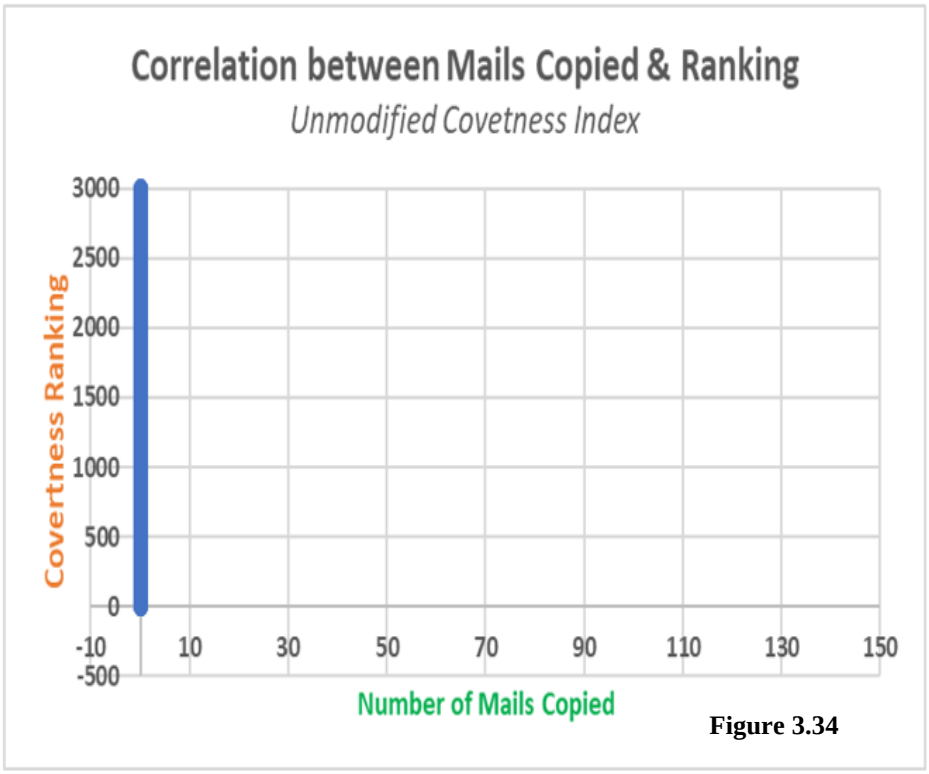


Figure 3.34

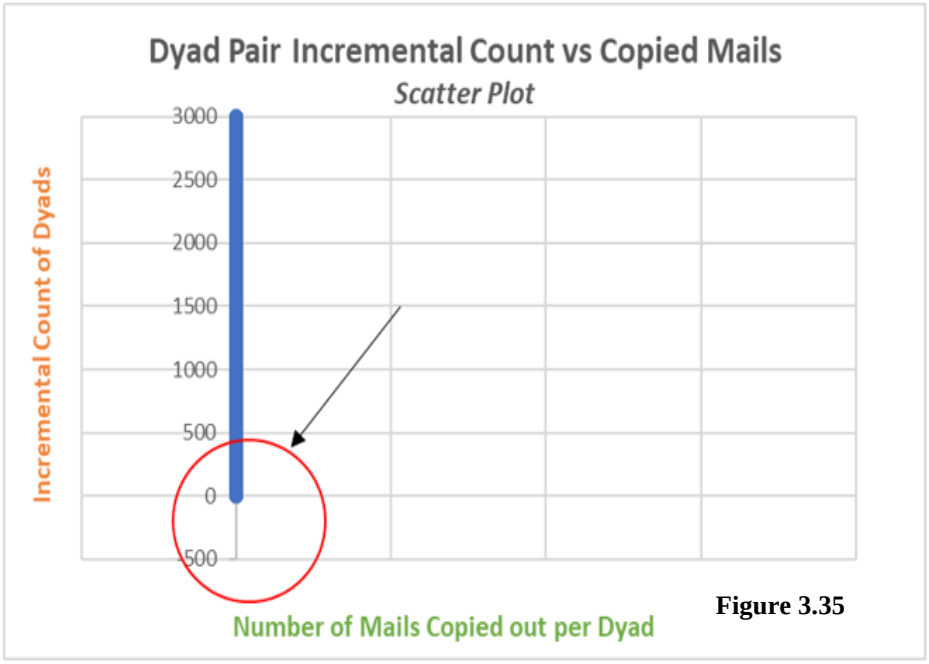


Figure 3.35

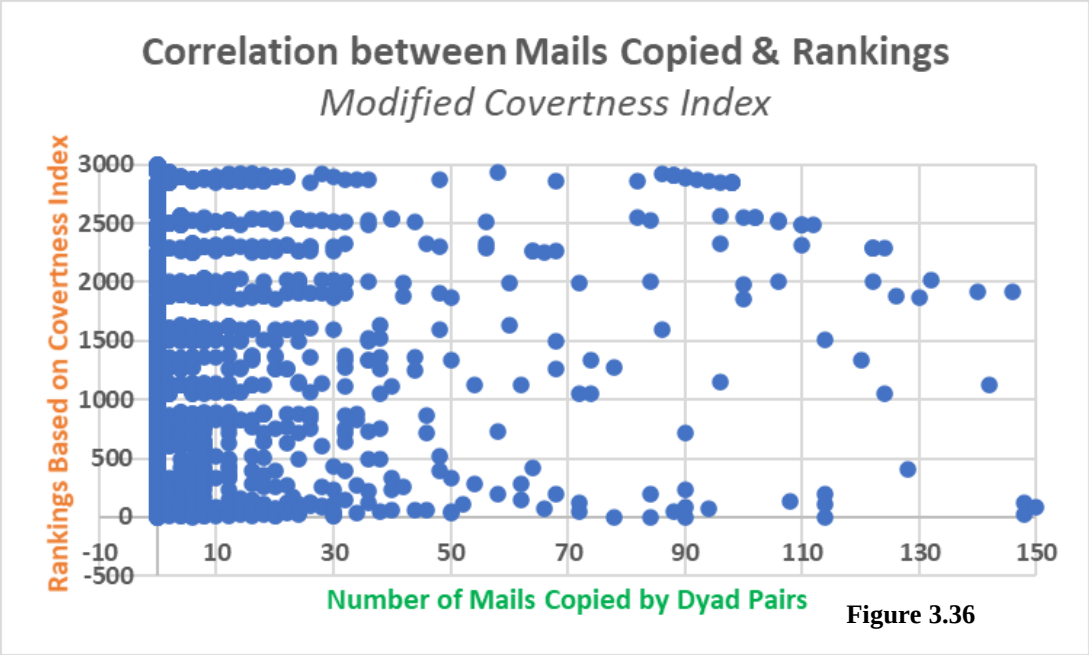


Figure 3.36

In the first of the two plots shown above, it's seen that all of the first 3000 top ranks by covertness are occupied by dyads, which have not marked out any of the emails exchanged as copies, i.e., the 'perfectly covert' or opaque edges. But, this is not seen in the second of the plots (in which the Covertness Index has been modified), where many of the top ranks by covertness index values are seen occupying a majority of the top ranks.

A similar conclusion can be drawn from the scatter plots shown below (Figures 3.37 and 3.38), which exhibit the correlation between the total number of emails exchanged between the constituent nodes of a dyad and the dyad's ranking based on the Covertness Index defined on its constituent edge. In the first of the two plots below, i.e., Figure 3.37, the scatter plot is in respect of the Covertness Index calculated as per the initial (unmodified) formula. In this plot, we may observe that dyads whose constituent nodes have exchanged a higher number of emails, say around 200 – 500, have very low rankings (many are at ranks 20,000 to 40,000). But dyads whose nodes have exchanged a lesser number of emails (less than 50) enjoy far better rankings on an average (5000 or less). Thus, in a way, dyads active in higher volumes of mail exchanges are getting disincentivized in rankings, which becomes a distinct disadvantage when analysts add pre-existing tie related data to the existing set-up as and how they become available. Extra volumes mean lower rankings,

and paradoxically when such dyads go out of reckoning, the information they contain also moves out from the scope of analysis. In this study, the edges are undirected and have binary orientation, i.e., a tie can have a value of 0 if the pair of nodes that it connects have exchanged no mails or a value of 1 if the associated nodes have exchanged one or more mails. The edge value isn't weighted in any way by the number of mail transactions that have occurred along with it. The Edge-Vertex function also doesn't keep a record of the number of mail exchanges directly. This important parameter can be factored into the calculation of covertness to make it a part of the Covertness Index, which is the basis for the modified formula.

This picture is more nuanced when we apply the modification to the Covertness Index formula, as is evident from the second of the scatterplots below Figure 3.38). The same dyads whose constituent node has exchanged a higher number of mails (200 – 500) enjoy better ranking (less than 500 or so), and the pairs whose nodes have exchanged less number of mails score lower on the rankings. In this way, the modified Covertness Index formula conserves the net information exchanged within a dyad, which is crucial if we are to drill further into the nature of their relationships later on.

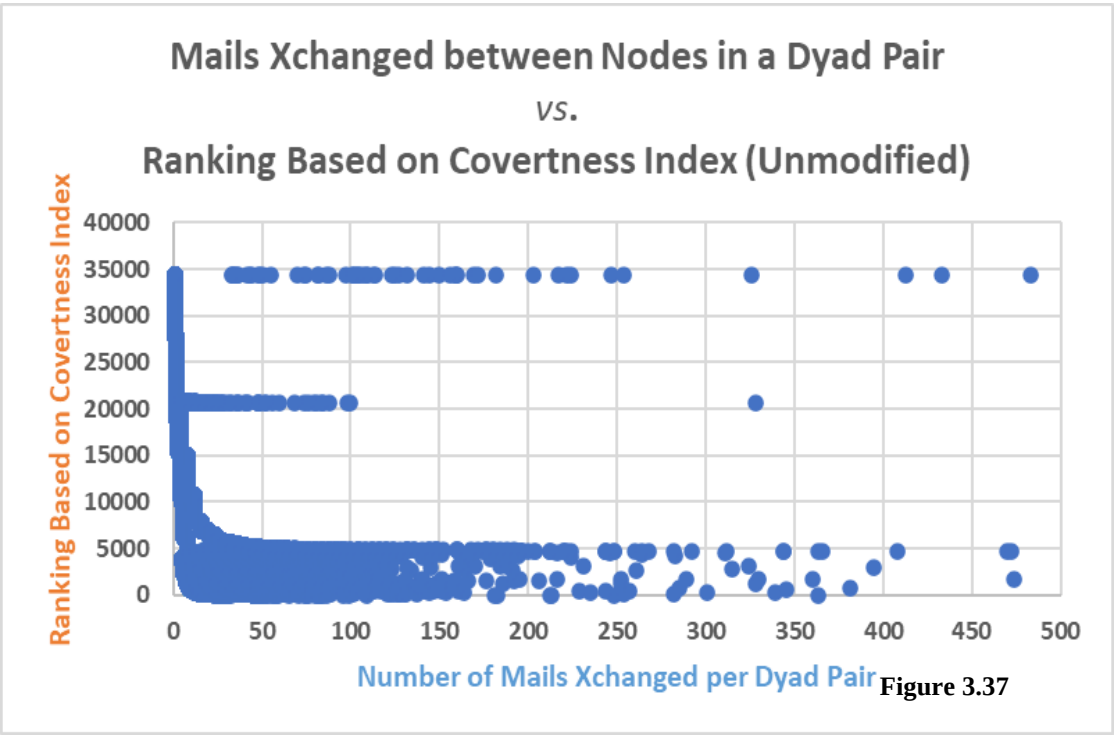


Figure 3.37

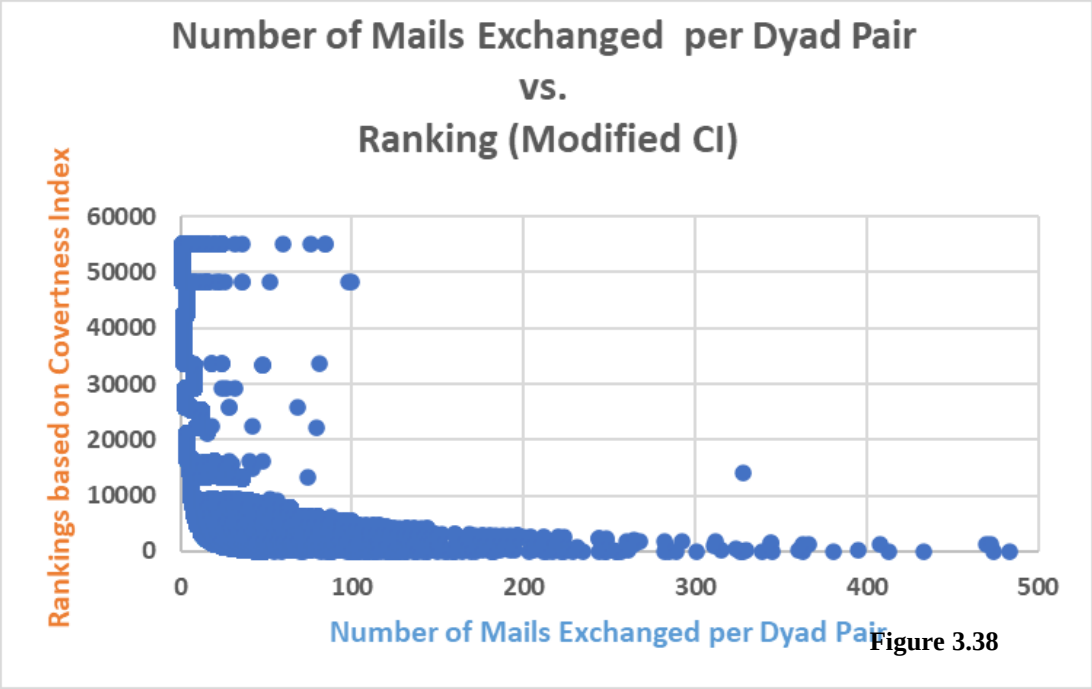


Figure 3.38

3.18 Selecting a Threshold of Covertness

The introduction of the Covertness Index[43] as a tie-based measure is seen to improve the chances of detecting Edge Vertexes of Interest to the study. In other words, the 'Needle in the haystack' problem of identifying 43 Edges of Interest in a set of 55000 plus edges is mitigated considerably with this approach.

If we have to consider all the 55288 edges in the reckoning, any surveillance operations involving multiples of this number will be exponentially more complex. Suppose a social network happens to bigger than the ENRON e-mail network, as most of them are likely to be. For instance, the worldwide web has typically billions of webpages, and the number of nodes on the Internet may still be higher.

[43]From this stage onwards the term 'Covertness Index' will mean the Modified Covertness Index.

Therefore, this study suggests a threshold of covertness that will serve as the floor of our analysis. That is, only such dyads having edges possessing a minimum value of the Covertness Index will be considered for analysis. This threshold will be a heuristic depending upon the circumstances in which the inquiry is being conducted.

Most analyses of covert networks are done with limited resources, both in terms of computational bandwidth and human resources. The time available is also fairly limited. It's often necessary to select what the analyst may consider as the 'best snapshot' of the solution in these situations. This solution may well be a subset of the whole, but one that sufficiently represents the problem.

In the present instance where the study focuses on detecting covertness of edges within the ENRON email corpus, the necessity is to identify the most covert specimens amongst the edges and then apply further analytics to unearth more of the related covert edges. This problem fits what Moore and Mertens (2011, p351) describes as a situation that happens if there is a whole stack of needles, but it's required to find the longest one, or the shortest, or the sharpest. The class of solutions is called Approximations. The need for an approximate path to solving any issue is justified by the famous quote often ascribed to Aristotle – "It is the mark of an educated mind to rest satisfied with the degree of precision which the nature of the subject admits and not to seek exactness where only an approximation is possible."

Moore and Mertens (2011, p.355) describe approximations thus – " If an Optimization problem is NP-hard, we can't expect to find an efficient algorithm that finds the optimal solution. But we might hope for an algorithm that finds a *good* solution – one that is guaranteed to be not much worse than the optimum." Using the notation of Moore and Mertens (2011), the "not much worse." An algorithm of this type that attempts to find a solution to an optimization problem using a reasonable rule of thumb is called a *heuristic*.[44]

[44]From the Greek εὑρίσκω,for "I discover". a *heuristic*is any approach to problem solving or self-discovery that employs a practical method that is not guaranteedtobe optimal, perfect or rational, but which is nevertheless sufficient for reaching an immediate, short-term goal.

Finding an optimal solution is impossible or impractical; heuristic methods can be used to speed up finding a satisfactory solution. Heuristics can be mental shortcuts that ease the cognitive load of making a decision.ionsApproximations are heuristic in nature. Wikipedia[45] defines a **heuristic** in "computer science, artificial intelligence, and mathematical optimization as a technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution, which is achieved by trading optimality, completeness, accuracy, or precision for speed. In a way, it can be considered a shortcut."

Another definition of a **heuristic function,** also called simply a **heuristic**, is a function that ranks alternatives in search algorithms at each branching step based on available information to decide which branch to follow. It may approximate the exact solution without producing the identical result.. (https://en.wikipedia.org/wiki/Heuristic_(computer_science)).

Going by this definition, Wikipedia further explains that a heuristic's objective is to solve a reasonable time frame that is good enough to solve the problem. **This solution may not be the best solution to this problem, or it may simply approximate the exact solution**. **But it is still valuable because finding it does not require a prohibitively long time**.

Most real-world applications have a complexity that matches the NP-hard postulates. Hence, the results of NP-hard computer science problems make heuristics the only viable option for various complex optimization problems that need to be routinely solved in real-world applications. Heuristics underlie the whole field of Artificial Intelligence and the computer simulation of thinking, as they may be used in situations where there are no known algorithms (Apter, 1970, p. 83).

The trade-off[46] criteria for deciding whether to use a heuristic for solving a given problem include the following:

- *Optimality:* When several solutions exist for a given problem, does the heuristic guarantee that the best solution will be found? Is it necessary to find the best solution?

[45]https://en.wikipedia.org/wiki/Heuristic_computer_science, (accessed June 6, 2020.)

- *Completeness:* When several solutions exist for a given problem, can the heuristic find them all? Do we need all solutions? Many heuristics are only meant to find one solution.
- *Accuracy and precision:* Can the heuristic provide a confidence interval for the purported solution? Is the error bar on the solution unreasonably large?
- *Execution time*: Is this the best-known heuristic for solving this type of problem? Some heuristics converge faster than others. Some heuristics are only marginally quicker than classic methods.

It may be difficult to decide whether the heuristic solution is good enough because the theory underlying heuristics is not very elaborate. The rationale for selecting a threshold of covertness is thus well established in the present case. Three different levels of the threshold are chosen just to have a comparative picture of the results. There is a progressive decrease in the cut-offs of the Covertness Index values and a corresponding increase in the numbers of dyads selected for further screening.

3.19 Metrics for Measuring Performance

3.19.1 Accuracy

The model we seek to build in the first part of the experiment is a type of engine that seeks to classify edges as covert (and of interest) and not covert (not of interest). Several types of metrics and scales measure performance in classifiers. The most naïve and popular is Accuracy. As a heuristic, or rule of thumb, accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally. However, it does not give detailed information regarding its application to the problem.

The problem with using accuracy as your main performance metric is that it does not do well when there is a severe class imbalance. The present problem at hand is a good example where Accuracy may not work well, explained through an example below.

[46] https://en.wikipedia.org/wiki/Heuristic_(computer_science) (accessed June 6, 2020)

Suppose we were to claim to create a model to identify covert Edge Vertices of Interest (EoIs) with greater than 99% accuracy. Would it be sufficient for detecting the Edges of Interest over the entire distribution? Well, here is the model: let's simply label every single edge as not a covert one. Given the 55300 edges in the overall ENRON mail corpus and the 43 (confirmed) Edges of Interest, which had something or the other to do with the scandal, this model achieves an astounding accuracy of 99.992%! That might sound impressive, but hides the crucial fact that we have not identified any covert edges at all though they exist! While this solution has nearly-perfect accuracy, this problem is one in which accuracy is not an adequate metric.

The covert edge detection task is an example of an imbalanced classification problem: we have two classes we need to identify — covert edges and not covert edges, i.e., edges which are not of any interest to us — with one category representing the overwhelming majority of the data points. Another imbalanced classification problem occurs in disease detection when the public's rate is very low. In both these cases, the positive class — disease or covert edge detection — is greatly outnumbered by the negative class. These problems are examples of the fairly common case in social networks when accuracy is not a good measure for assessing model performance.

3.19.2 Recall

Intuitively, we know that proclaiming all data points as negative in the covert edge detection problem is not helpful and, instead, we should focus on identifying the positive cases. The metric our intuition tells us we should maximize is known in statistics as **recall**, or the ability of a model to find all the relevant cases within a dataset. The precise definition of recall is the number of true positives divided by the number of true positives and false negatives. True positives are data points classified correctly as positive by the model (meaning they are correct), and false negatives are data points the model identifies wrongly as negatives (incorrect). In the instant case, true positives are correctly identified covert edges, and false negatives would be the edges the model labels as *not* covert edges that

were covert. Recall can be thought of as a model's ability to find all the data points of interest in a dataset.

Recall calculates how many of the Actual Positives our model captures through labeling it as Positive (True Positive). Applying the same understanding, it is trivial to note that Recall is an ideal metric used to select the best model when there is a high cost associated with False Negatives.

To quote another instance, in fraud detection or sick patient detection, if a fraudulent transaction(Actual Positive) is predicted as non-fraudulent(Predicted Negative), the consequence can be harmful for the financial institution.

Similarly, in sick patient detection, a sick patient (Actual Positive) goes through the test and is predicted as not sick (Predicted Negative). The cost associated with False Negative will be extremely high if the sickness is contagious.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Table 3.14 (a) Confusion matrix showing Actual Positive results.

True Positive + False Negative = Actual Positive

However, there is an additional point to note at this stage: if we label all edges as covert and of interest, then Recall goes to 1.0! We end up with a perfect classifier. Well, not exactly. As with most concepts in data science, there is a trade-off in the metrics chosen to maximize. In the case of a recall, when we increase the recall, we decrease the precision. Again, we intuitively know that a model that labels 100% of the edges as covert is probably not useful because we would have to keep surveillance on all ENRON employees, a costly proposition indeed, and one that the study aimed to mitigate in the first place! This new model suffers from low **Precision** or a classification model's ability to identify only the relevant data points.

3.19.3 Precision

Precision is defined as the number of true positives divided by the number of true positives and false positives. False positives are cases the model incorrectly labels as positive that are negative. In the case being studied here, the edges the model classifies as covert are not covert at all. While Recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant were relevant. Precision talks about how precise/accurate the model is; that is, out of those predicted positive, how many are positive.

Precision is a good measure to determine when the costs of False Positive is high. For instance, in email spam detection, a false positive means that a non-spam email (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

Thus, we may define Precision as per the formula below:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Table 3.14 (b) Confusion Matrix showing total predicted positive results.

True Positive + False Positive = Total Predicted Positive

Now, we can see that our first model, which labeled all edges as *not* covert, wasn't very useful. Although it had near-perfect accuracy, it had 0 Precision and 0 Recall because there were no True Positives! Say we modify the model slightly and identify a single edge correctly as a covert one; now, our precision will be 1.0 (no false positives). However, our Recall will be very low because we will still have many false negatives. Suppose we go to the other extreme and classify all edges as covert. In that case, we will have a recall of 1.0 — we'll detect every covert edge of interest — but our Precision will be very low, and we'll end up keeping costly surveillance on what would be many uninvolved employees. In other words, as we increase precision, we decrease recall and vice-versa. The see-saw nature of the correlation between Precision and Recall in a model is shown in Figure3.39 below:

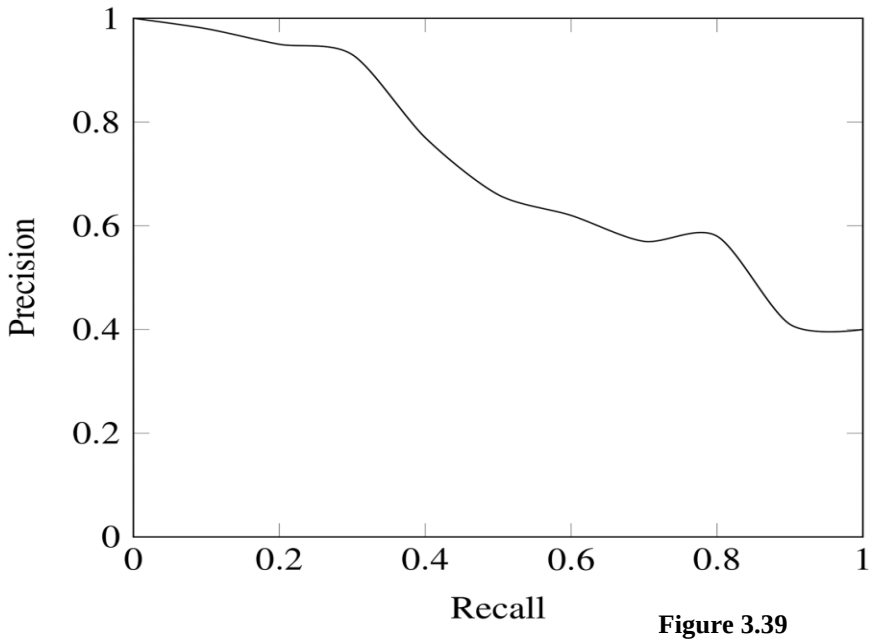


Figure 3.39

3.19.4 Combining Precision and Recall

We might know that we want to maximize either Recall or Precision at the other metric's expense in some situations. For example, in preliminary disease screening of patients for follow-up examinations, we would probably want a recall near 1.0 — we want to find all patients who have the disease — and we can accept a low Precision if the cost of the follow-up examination is not significant. However, in cases where we want to find an optimal blend of Precision and Recall, we can combine the two metrics using the F1 score.

The F1 score is the harmonic mean of precision and recall taking both metrics into account in the following equation:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The harmonic mean instead of a simple average is used because it dis-incentivizes extreme values. A classifier with a Precision of 1.0 and a Recall of 0.0 has a simple average of 0.5 but an F1 score of 0. The F1 score gives equal weight to both measures and is a specific example of the general Fβ metric, where β can be adjusted to give more weight to either Recall or Precision. (There are other metrics for combining precision and recall, such as the Geometric Mean of Precision and Recall, but the F1 score is the most commonly used.) If we need to create a balanced classification model with the optimal balance of Recall and Precision, we try to maximize the F1 score.

3.19.5 Visualizing Precision and Recall

Having discussed various metrics to evaluate our model's efficacy, we briefly discuss a few techniques to explain how the concepts described above may be applied.

By far, the most common technique is computing the *confusion matrix*, which is useful for quickly calculating Precision and Recall given the predicted labels from a model. A confusion matrix for binary classification shows the four different outcomes: True Positive, False Positive, True Negative, and False Negative. The actual values form the columns and the predicted values (labels) form the rows. The intersection of the rows and columns shows one of the four outcomes. For example, if we predict a data point to be positive, it turns out to be negative, then it’s termed as a false positive.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Table 3.15 An example of a Confusion Matrix

Going from the Confusion Matrix to the Recall and Precision requires finding the respective values in the matrix and applying the equations:

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \qquad precision = \frac{true\ positives}{true\ positives + false\ positives}$$

We can recast the above matrix into one that suits the purposes of our study. In the instant case, the True Positives are those dyads whose edges are covert and are the ones that are sought to be detected. The False Positives are the dyads whose edges have been wrongly classified as covert and of interest to us. False Negatives are those dyads whose constituent nodes have ties that are covert and of interest to the investigator but wrongly classified as benign edges and liable to be left out of the inquiry's scope. Finally, True Negatives in the Confusion Matrix above correspond to the dyads whose edges are not covert or interested in the investigation and correctly identified.

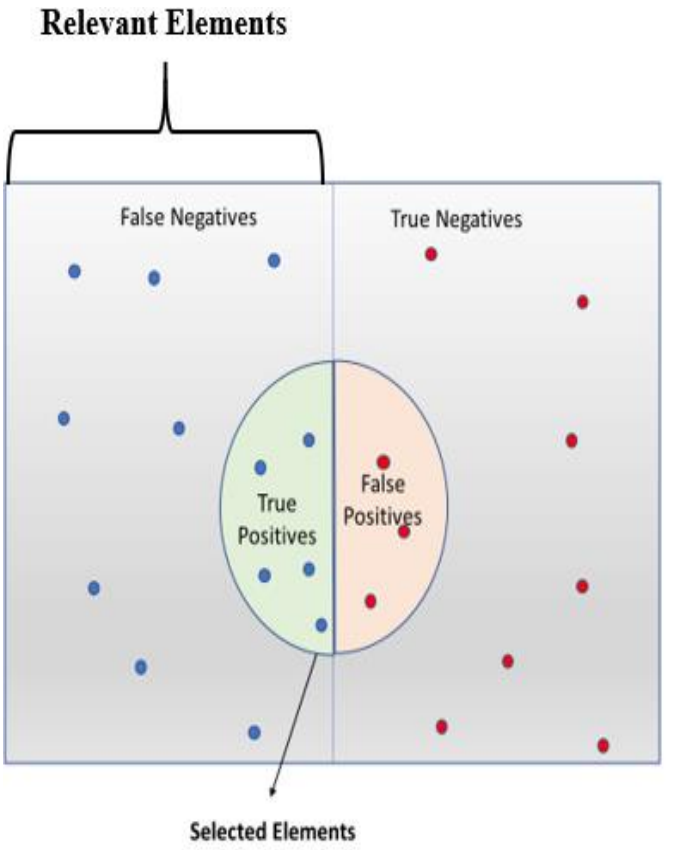
The table below shows the Confusion Matrix with the cells' entities recast as per the above analysis. The subsequent table maps the entries of the Confusion Matrix with the ones required for this study's purposes.

		Actual	
		Positive	Negative
Predicted	Positive	Covert Edges of Interest	Edges which are not of Interest but Identified wrongly
	Negative	Covert Edges of Interest but not Identified so	Edges which are not of Interest and Identified as such

Table 3.16 Confusion Matrix applied to the problem in hand.

Confusion Matrix Entries	Recast Confusion Matrix Entries
True Positives	**Covert Edges of Interest (EoIs) correctly identified.**
False Positives	**Benign or Non-Covert Edges incorrectly identified as Covert Edges of Interest**
True Negatives	**Benign or Non-Covert Edges identified correctly.**
False Negatives	**Covert Edges of Interest (EoIs) incorrectly identified as Benign or Non-Covert Edges of Interest**

Table 3.17 Table showing mapping of Confusion Matrix entries to Problem Statement.



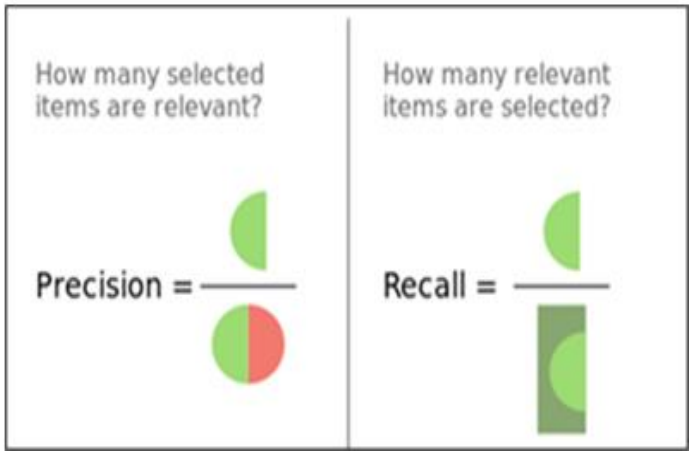


Figure 3.40 The Figure above illustrates how the values of Precision and Recall are selected from a model and how they are calculated.

3.20 Measuring Improvement in Detection

In the first instance, we have selected the first 2500 top-ranked Edge Vertices (arranged in the descending order of their values of Covertness Index). The Precision, Recall, F1 measures are all calculated to prove the better detection of the Edge Vertices of Interest (EoIs) in the overall distribution by using the index.

Case#1: 2500 Top-Ranked Edges:

Two charts are presented below. The first of the pair reflects the numbers of the Edges of Interest (EoIs), which occur in the top 2500 edges arranged along with the Covertness Index's descending values of the respective ties. The first chart has the ranks of the edges arranged in numerical order. The lower the placement of an edge on the X-axis, the better the rank. The second chart is a replica of the first, but the ranks are now shown in a logarithmic format[47].

The figures below show how the prevalence of the edges of interest (EoIs) has markedly improved after applying the Covertness Index to the ties between the nodes constituting the dyads.

We discussed using various metrics to measure improvement in the detection of covert edges, i.e., EoIs. The first of the metrics discussed was Precision, which indicates how correctly the model predicts the true positives. That is, of the Edges predicted as covert, how many are the covert edges we seek. The charts below show that there are 23 covert edges detected correctly, and the remainder out of 2500 are incorrectly predicted as covert. That is, the true positives are 23, whereas the false positives are 2500 – 23 = 2477.
Referring back to the section on the metrics, the formula of Precision was :
Precision = True Positives / (True Positives + False Positives)

[47]Though there is not much importance of depicting values on the X-axis in a logarithmic pattern with only 2500 edges to consider, if we increase the number of edges to a much larger number (say 50,000) the plotting becomes unwieldy without using log values.

In other words, the Precision of our model when we reach the count of 2500 in terms of ranking is:

Precision = 23/ (23 + 2477) = 23/ 2500 = 0.0092 or about 1%.

The figure on the face appears not too encouraging till we compare it with the Precision achieved by adopting a uniform distribution model. That is, the covert edges we want to identify and detect are distributed throughout the set of all the edges uniformly. As seen from the figure below, the number of covert edges (EoIs) detected at 2500 is hardly 2. This model's Precision is barely 0.0008 or 0.08%, which is less than 10% of what our model based on the covertness index gives us.

Let's denote Precision arising out of the Covertness Model as $\mathbf{P_C}$andthe Precision arising out of a Uniform Distribution as $\mathbf{P_U}$.

Thus,

$$\mathbf{P_C}=\frac{(\boldsymbol{NumberofEdgesofInterestDetected})}{(\boldsymbol{NumberofEdgesofInterestDetected+NumberofIncorrectCovertEdgesDetected})}$$

$$\mathbf{P_C = \frac{23}{23+2477} = \frac{23}{2500} = 0.0092 \text{ or, } 0.92\ \%.}$$

Likewise,

$$\mathbf{P_U = \frac{1.84}{2500} = 0.0008 \text{ or, } 0.08\ \%.}$$

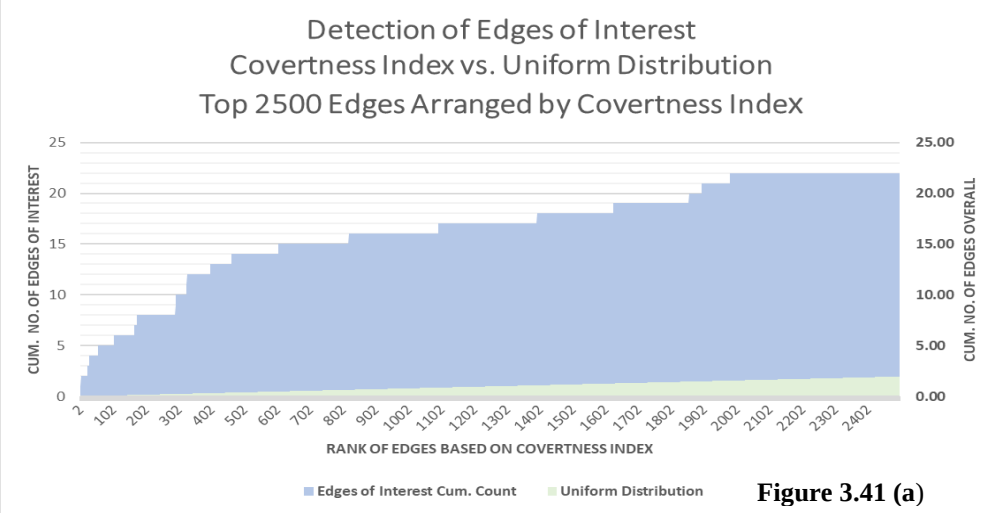


Figure 3.41 (a)

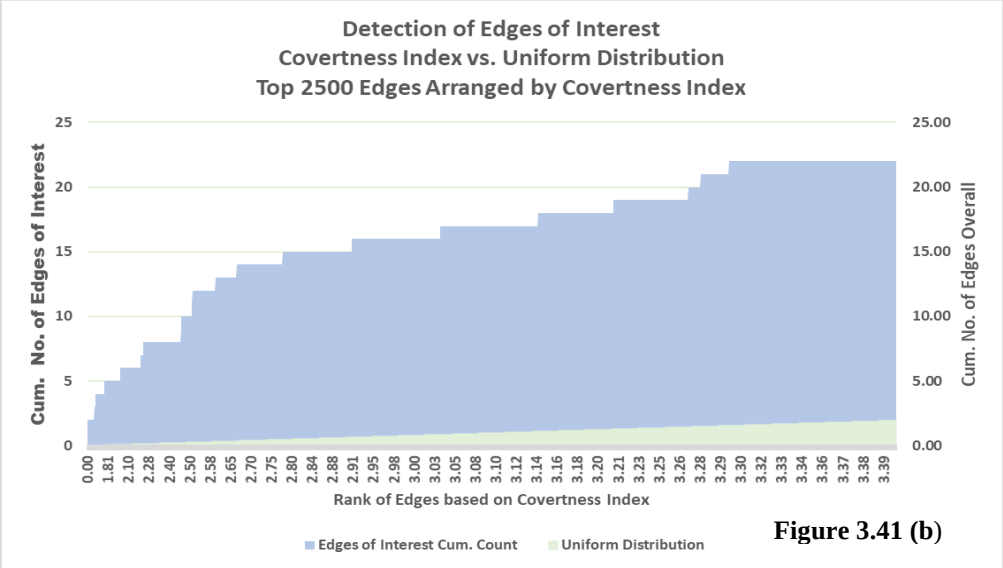


Figure 3.41 (b)

The second of the metrics we had discussed to measure performance was Recall. We may now estimate the figure for Recall for the Covertness Index model. As discussed earlier, Recall calculates how many of the Actual Positives our model captures through labeling it as Positive (True Positive). In other words, Recall calculates how many of the actual Edges of Interest (EoIs) our model was able to detect through labeling them as correct out of the total EoIs available. We've seen that in the set of the 2500 dyads ranked as per the Covertness Index measured on their ties, there were 23 EoIs correctly identified. The total number of EoIs that we have is 43. Thus, Recall as per your model comes to 23/43 or 0.53, i.e., 53 %. Again, this figure is not very impressive until we compare it with the figure of Recall obtained from a Uniform Distribution. We discussed the calculation of Precision.

In a Uniform Distribution, the Edges of Interest are presumed to be prevalent uniformly throughout the edges whose cardinality is around 55,300. Suppose we calculate the number of EoIs that will occur within the figure of 2500. In that case, it comes to 1.84, which means that using the Uniform Distribution calculation, the Recall figure is 1.84/43 0.04, i.e., 4 %. Compare this with the figure of 53 % for the model built around the Covertness Index of ties. The Recall metric increases if the model reduces the False Negatives. Here, the False Negatives are the Edges of Interest (EoIs), which the model has incorrectly identified as not covert. Since the number of detected EoIs is 23, and the total number of

EoIs is 43, 20 EoIs have been left out by employing the Covertness Index model. In contrast, the Uniform Distribution of the model identifies only about 2 EoIs and leaves out as many as 41. Our model thus scores significantly higher in the Recall metric as well.

As in the case of the Precision metric, the results are presented below notationally.

Let's denote the Recall in the Covertness Model as $\mathbf{R_C}$ and the Recall in the Uniform Distribution as $\mathbf{R_U}$.

$$\mathbf{R_C} = \frac{(\boldsymbol{Number of Edges of Interest Detected})}{(\boldsymbol{Number of Edges of Interest Detected + Number of Covert Edges Not Detected})}$$

$$\mathbf{R_C} = \frac{\mathbf{23}}{\mathbf{23+20}} = \frac{\mathbf{23}}{\mathbf{43}} \mathbf{= 0.53 \text{ or, } 53\ \%.}$$

Likewise,

$$\mathbf{R_U} = \frac{\mathbf{1.84}}{\mathbf{43}} \mathbf{= 0.04 \text{ or, } 4\ \%.}$$

Our model is far more efficient in reducing the number of covert edges not detected in the dataset. The recall is particularly important in the study of covert social networks. It is important that any detection model in this field needs to keep as many suspect actors within the surveillance system. If a significant number of the bad actors slip out of the dragnet, the results can be catastrophic. In the Uniform Distribution Model, the number of covert edges that have managed to elude scrutiny is a staggering 42 or 96 %, which, in a real-world situation, implies that almost all the malfeasant players are roaming free to perpetuate their actions.

The chart below summarizes Recall performance by invoking the Covertness Index over the Uniform Distribution model.

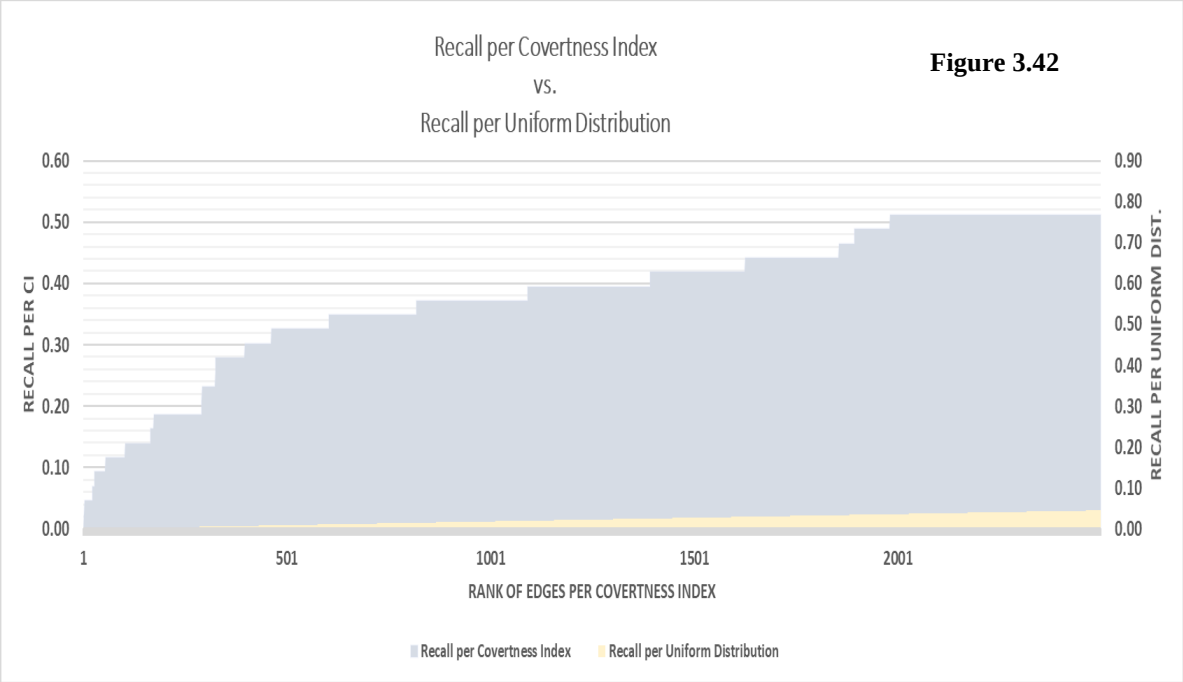


Figure 3.42

It has been discussed at some length earlier that a balanced model doesn't lean too heavily on either Precision or Recall. There are advantages in choosing a more balanced metric like the F1 Measure. F1 is an overall measure of a model's accuracy that combines Precision and Recall. A good F1 score means that there are low False Positives and low False Negatives, so the model is correctly identifying real threats, and the number of false alarms is less. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

In the Covertness Model that we have developed, the F1 score is

$$\boldsymbol{F1 = 2\,\frac{(Precision * Recall)}{(Precision + Recall)}}$$

$$\boldsymbol{(F1)_C = 2 * \frac{(0.0092 * 0.53)}{(0.0092 + 0.53)} = 0.018.}$$

$$\boldsymbol{(F1)_C = 0.018.}$$

Likewise, for the Uniform Distribution Model,

$$(F1)_U = 2 * \frac{(0.0008 * 0.04)}{(0.0008 + 0.04)} \cong 0.0016.$$

$$(F1)_U = 0.0016$$

$$(F1)_C \mid (F1)_U = 0.036 / 0.0016 = 11.3$$

Thus, we can see that the improvement in the detection of covert edges in terms of F1 Score is also considerable. This is made clear by the chart shown below.

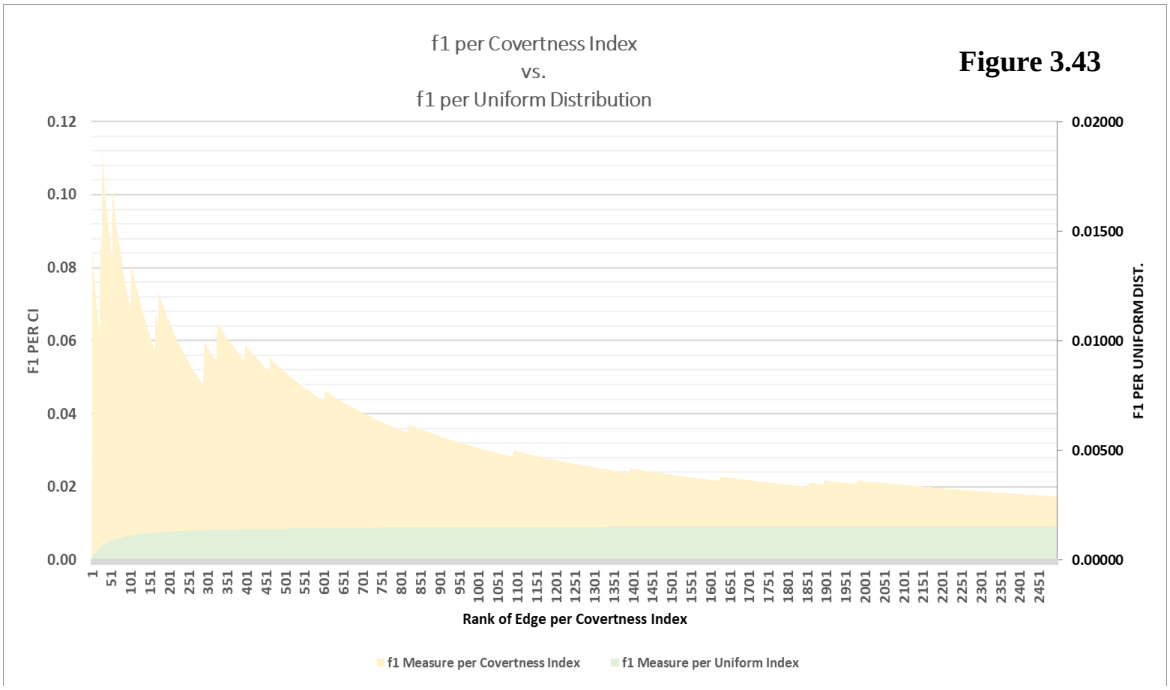


Figure 3.43

Case#2: 5000 Top-Ranked Edges:

We now compare the results for top-ranked 5000 dyads. As in the earlier instance of 2500 top-ranked pairs, we compute and compare first the prevalence[48] of the covert Edges of

[48]Prevalence as we know is same as the Precision measure for the distributions.

Interest over the entire set of 5000 edges between the Covertness Index model and the Uniform Distribution model, followed by the comparison of Recall values for both the distributions and finally by the computation and comparison of F1 Scores.

To compare the prevalence, two charts are presented below. The first of the pair reflects the numbers of the Edges of Interest (EoIs), which occur in the top 5000 edges arranged along with the Covertness Index's descending values of the respective ties. The first chart has the ranks of the edges arranged in numerical order. The lower the placement of an edge on the X-axis, the better the rank. The second chart is a replica of the first, but the ranks are now shown in a logarithmic format.

The figures below show how the Precision values of the edges of Interest (EoIs) have markedly improved after applying the Covertness Index to the ties between the nodes constituting the dyads. Precision indicates how accurate the prediction is that the model makes regarding true positives. That is, of the edges predicted as covert, how many are the covert edges we aim for? From the figures below, we may see that there are 27 covert edges detected correctly, and the remainder out of the 5000 are incorrectly predicted as covert. That is, the true positives are 27, whereas the false positives are 5000 – 27 = 4973.

Precision is calculated per the following formula: as :
Precision = True Positives / (True Positives + False Positives)

In other words, the Precision of our model when we reach the count of 5000 in terms of ranking is:
Precision = 27/ (27 + 4973) = 27/ 5000 = 0.0054 or about 0.5%.

As in the earlier case where we considered the 2500 top-ranked covert edges, the figure appears insignificant until we compare it to 0.0008 or 0.08% obtained from the Uniform Distribution. The Precision value produced by the Covertness model is more than six times that of the Precision figure from the Uniform Model.

We can now repeat the notational representation for calculating and comparing the Precision measure. The Precision arising out of the Covertness Model is denoted as $\mathbf{P_C}$, and the Precision arising out of a Uniform Distribution is represented as $\mathbf{P_U}$**.**

Thus,

$$\mathbf{P_C} = \frac{(\textbf{\textit{Number of Edges of Interest Detected}})}{(\textbf{\textit{Number of Edges of Interest Detected}} + \textbf{\textit{Number of Incorrect Covert Edges Detected}})}$$

$$\mathbf{P_C} = \frac{\mathbf{27}}{\mathbf{23+4973}} = \frac{\mathbf{27}}{\mathbf{5000}} = \mathbf{0.0054 \text{ or, } 0.54\ \%.}$$

Likewise,

$$\mathbf{P_U} = \frac{\mathbf{3.89}}{\mathbf{5000}} = \mathbf{0.00078 \text{ or } 0.08\ \%.}$$

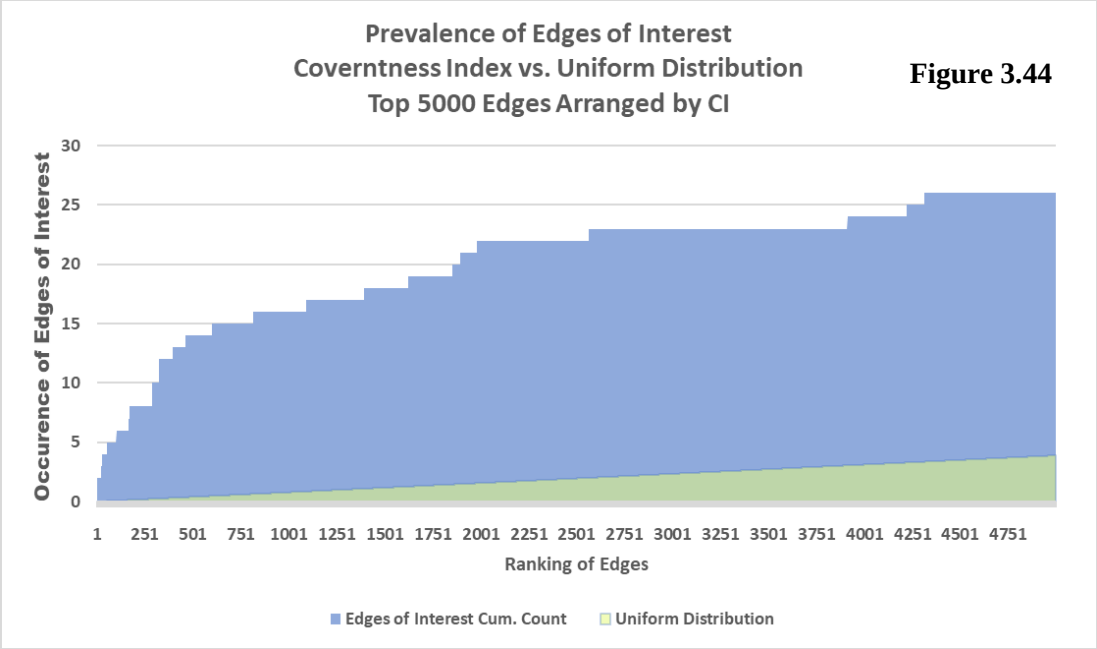


Figure 3.44

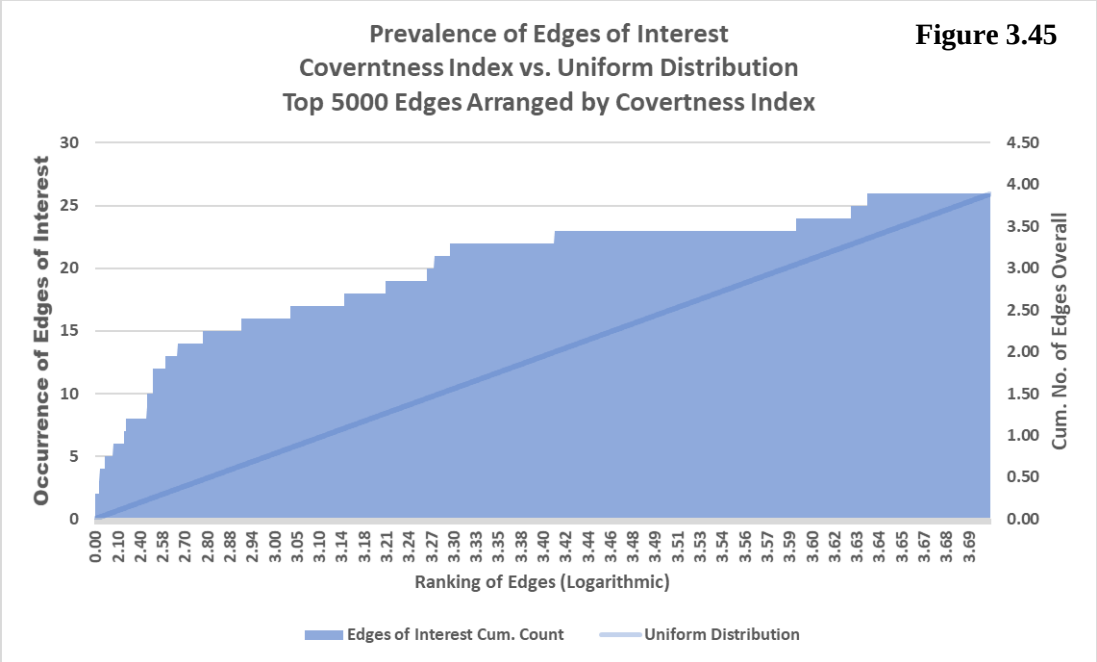


Figure 3.45

We now compare the second of the metrics we had discussed to measure performance, i.e., Recall, which calculates how many of the Actual Positives our model captures through labeling it as Positive (True Positive), which is the same as calculating how many of the actual covert Edges of Interest (EoIs) our model was able to detect through labeling them as correct out of the total EoIs available. In the top 5000 dyads ranked as per their Covertness Index, there were 27 EoIs correctly identified. The total number of EoIs that we have is 43. Thus, Recall as per our model comes to 27/43 or 0.63, i.e.,63 %. Again, this figure is not very enticing until we compare it with Recall's figure obtained from the Uniform Distribution model.

In a Uniform Distribution, the Edges of Interest are presumed to be prevalent uniformly throughout the edges whose cardinality is around 55,300. If we calculate the number of EoIs that will occur within the figure of 5000, it comes to 3.89. The figure of Recall is 3.89/43 or 0.09, i.e.,9 %. This is only a fraction of 63 % that is obtained from the Covertness Index Model. The Recall metric increases if the model reduces the False Negatives. Here, the False Negatives are the Edges of Interest (EoIs), which the model has incorrectly identified as not covert. Since the number of detected EoIs is 27, and the total number of EoIs is 43, 16 EoIs have been left out by employing the Covertness Index model. In contrast, the Uniform Distribution model identifies only about 4 EoIs and leaves out as many as 39.

Notationally, let's denote the Recall in the Covertness Model as $\mathbf{R_C}$and the Recall in the Uniform Distribution as $\mathbf{R_U}$**.**

Thus,

$$\mathbf{R_C} = \frac{(\textit{Number of Edges of Interest Detected})}{(\textit{Number of Edges of Interest Detected} + \textit{Number of Covert Edges Not Detected})}$$

$$\mathbf{R_C} = \frac{27}{27+16} = \frac{27}{43} = \mathbf{0.63 \text{ or, } 63\ \%.}$$

Likewise,

$$\mathbf{R_U = \frac{3.89}{43} = 0.09 \text{ or } 9\ \%.}$$

Thus, in the second instance where 5000 top-ranked covert dyads are considered, the Covertness Index model is far more efficient in reducing the number of covert edges not detected in the dataset. If many bad actors slip out of the dragnet in a surveillance mechanism, the entire exercise fails. In the Uniform Distribution Model, the number of covert edges that have managed to elude scrutiny is a staggering 39 or 91 %, which implies that nearly all covert players have eluded the surveillance dragnet.

The chart below summarizes Recall performance by invoking the Covertness Index over the Uniform Distribution model.

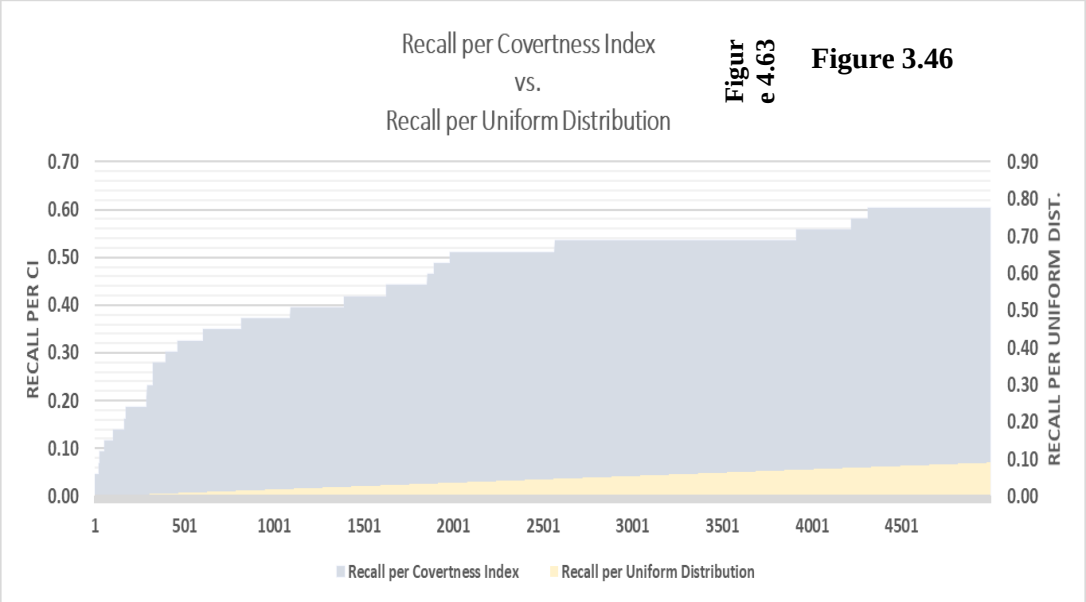


Figure 4.63 **Figure 3.46**

We now come to the third measure of performance, i.e., the F1 measure.F1 is a balanced measure of a model's accuracy that combines Precision and Recall. A good F1 score means that there are low False Positives and low False Negatives, so an optimal number of vulnerabilities is identified. The number of false alarms is also less.

In the Covertness Model that we have developed, the F1 score is

$$\boldsymbol{F1 = 2\ \frac{(Precision * Recall)}{(Precision + Recall)}}$$

$$\boldsymbol{(F1)_C = 2 * \frac{(0.0054 * 0.63)}{(0.0054 + 0.63)} = 0.0107.}$$

$$\boldsymbol{(F1)_C = 0.011.}$$

Likewise, for the Uniform Distribution Model,

$$\boldsymbol{(F1)_U = 2 * \frac{(0.0008 * 0.09)}{(0.0008 + 0.09)} \cong 0.0016.}$$

$$\boldsymbol{(F1)_U = 0.0016}$$

$$\boldsymbol{(F1)_C \mid (F1)_U = 0.011 / 0.0016 = 6.9}$$

The improvement in the detection of covert edges in terms of the F1 Score is marked. The chart below illustrates this.

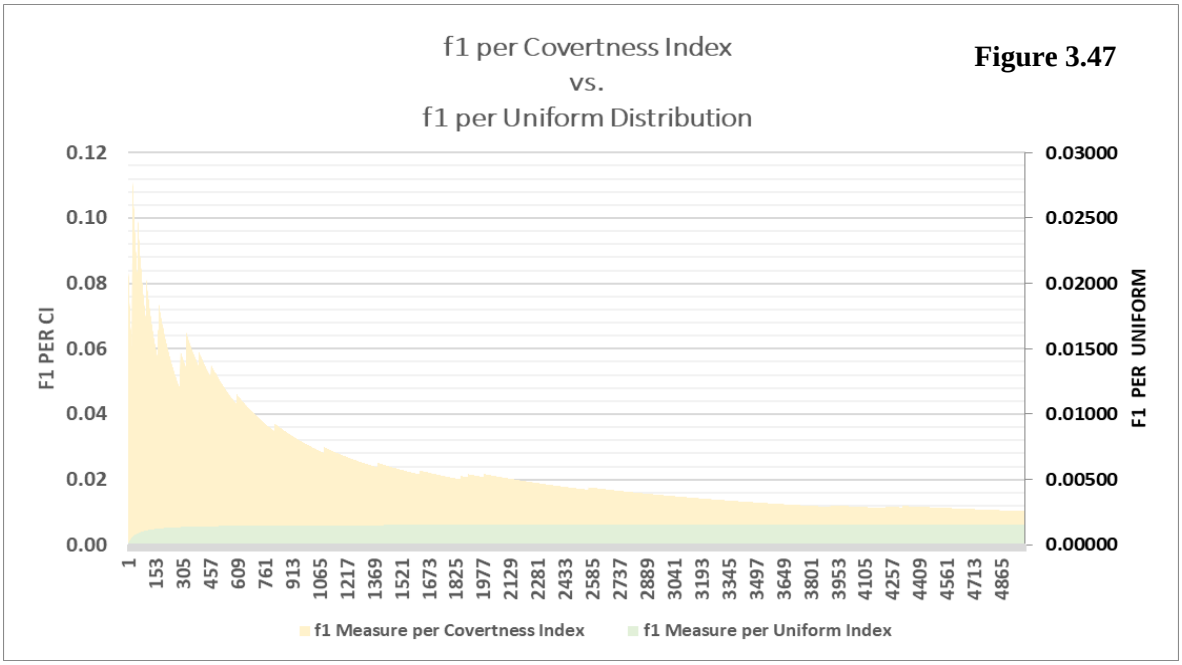


Figure 3.47

Case#3-10,000 Top-ranked Covert Dyads

The performance metrics have been calculated for this case and presented below:

Precision:

Precision = True Positives / (True Positives + False Positives)

= 35/ (35 + 9965) = 35/ 10000 = 0.0035 or about 0.35%.

As in the earlier case where we considered the 5000 top-ranked covert edges, the figure appears insignificant until we compare it to the figure of 0.0008 or 0.08% obtained from the Uniform Distribution. The Precision value produced by the Covertness model is more than four times that of the Precision figure from the Uniform Model.

We can now repeat the notational representation for calculating and comparing the Precision measure. The Precision arising out of the Covertness Model is denoted as $\mathbf{P_C}$, and the Precision arising out of a Uniform Distribution is represented as $\mathbf{P_U}$**.**

Thus,

$$P_C = \frac{(Number\ of\ Edges\ of\ Interest\ Detected)}{(Number\ of\ Edges\ of\ Interest\ Detected + Number\ of\ Incorrect\ Covert\ Edges\ Detected)}$$

$$P_C = \frac{35}{35+9965} = \frac{35}{10000} = 0.0035 \text{ or, } 0.35\ \%.$$

Likewise,

$$P_U = \frac{7.8}{10000} = 0.00078 \text{ or } 0.08\ \%.$$

The pair of figures shown below illustrates the Covertness Index model's superiority over the Uniform Distribution model in detecting the covert EoIs. The first figure has the horizontal axis laid out on a plain numerical scale, and the next one is laid out on a logarithmic scale for clarity.

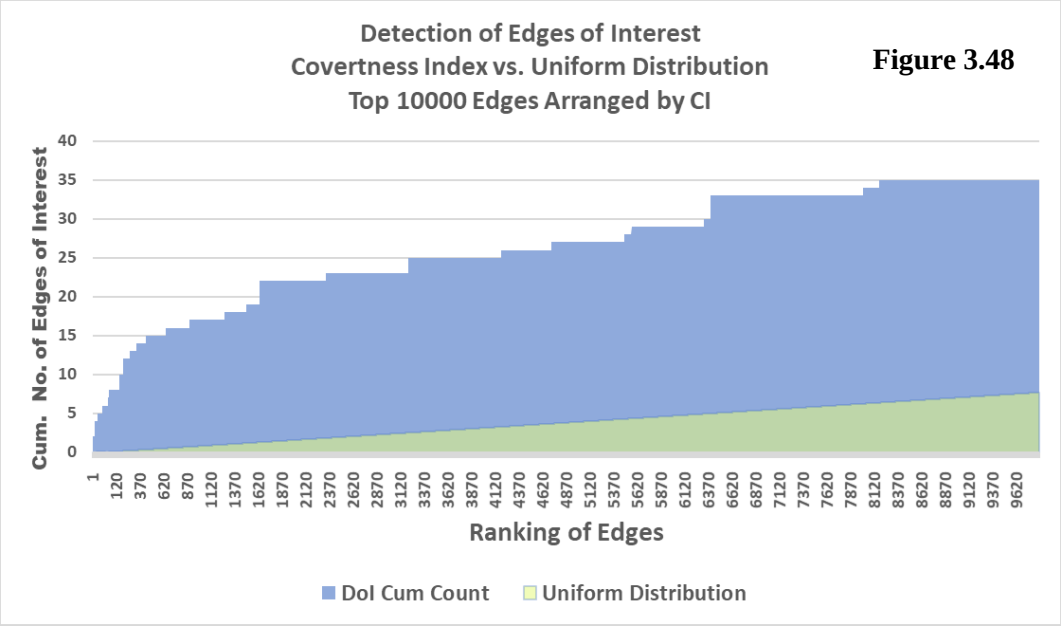


Figure 3.48

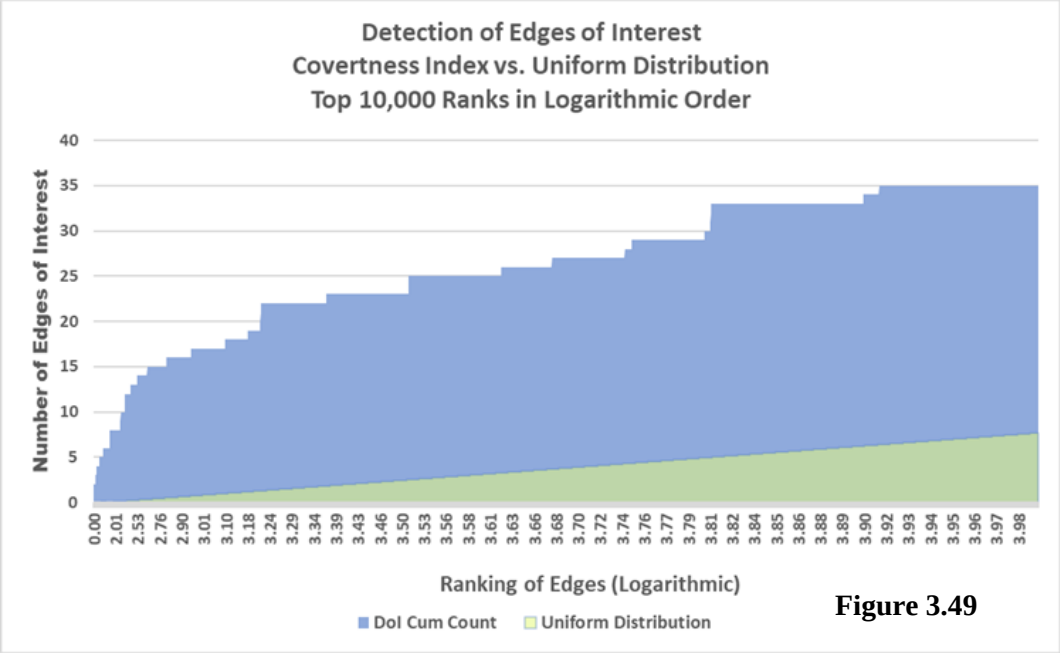


Figure 3.49

Recall:

In the top 10000 dyads ranked as per their Covertness Index, 35 EoIs are correctly identified. Thus, Recall as per our model comes to 35/43 or 0.81, i.e., 81 %.

In a Uniform Distribution, the figure of Recall is 7.8/43 or 0.18, i.e., 18 %. Again, this is only a fraction of 81 % obtained from the Covertness Index Model. The Recall metric increases if the model reduces the False Negatives. Here, the False Negatives are the Edges of Interest (EoIs), which the model has incorrectly identified as not covert. Since the number of detected EoIs is 35, and the total number of EoIs is 43, 8 EoIs have been left out by employing the Covertness Index model. In contrast, the Uniform Distribution model identifies only about 8 EoIs and leaves out as many as 35.

Notationally-

$$\mathbf{R_C} = \frac{(\textit{Number of Edges of Interest Detected})}{(\textit{Number of Edges of Interest Detected} + \textit{Number of Covert Edges Not Detected})}$$

$$\mathbf{R_C} = \frac{\mathbf{35}}{\mathbf{35+8}} = \frac{\mathbf{35}}{\mathbf{43}} = \mathbf{0.81 \text{ or } 81\ \%.}$$

Likewise,

$$\mathbf{R_U} = \frac{\mathbf{7.8}}{\mathbf{43}} = \mathbf{0.18 \text{ or } 18\ \%.}$$

$$\mathbf{R_C \mid R_U = 4.5}$$

Thus, in the third instance where 10000 top-ranked covert dyads are considered, the Covertness Index model is far more efficient in reducing the number of covert edges not detected in the dataset. In a situation where scrutiny is being carried out to detect adversarial and covert players, if many suspects go undetected, the dangers of an unwanted incident grow manifold. We may see that in the Uniform Distribution Model, the number of covert edges that have managed to elude scrutiny is a staggering 35 or 82 %, which implies that nearly all of the suspects have eluded detection.

The chart below summarizes Recall performance by invoking the Covertness Index over the Uniform Distribution model.

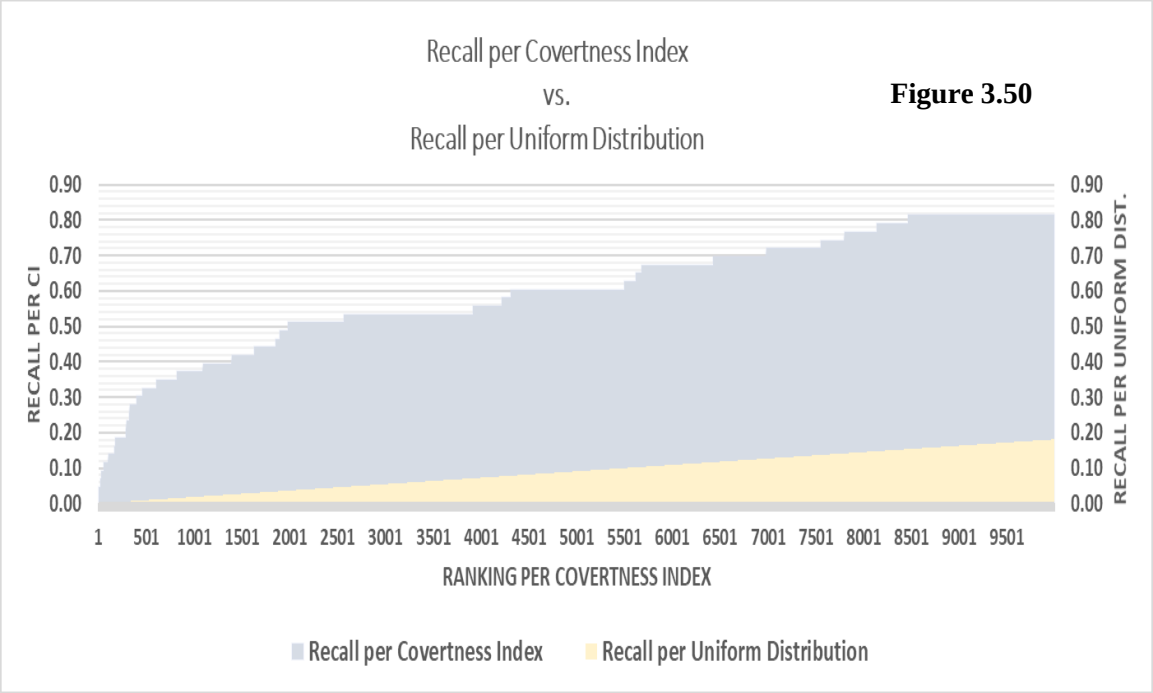


Figure 3.50

F1 Measure Scores Compared for the top 10000 ranks:

$$\boldsymbol{F1 = 2\,\frac{(Precision * Recall)}{(Precision + Recall)}}$$

$$\boldsymbol{(F1)_C = 2 * \frac{(0.0035 * 0.81)}{(0.0035 + 0.81)} = 0.007.}$$

$$\boldsymbol{(F1)_C = 0.007.}$$

Likewise, for the Uniform Distribution Model,

$$\boldsymbol{(F1)_U = 2 * \frac{(0.0008 * 0.18)}{(0.0008 + 0.18)} \cong 0.0016.}$$

$$\boldsymbol{(F1)_U = 0.0016}$$

$$\boldsymbol{(F1)_C \mid (F1)_U = 0.007 / 0.0016 = 4.4}$$

The improvement in the detection of covert edges in the F1 score is shown in the figure below.

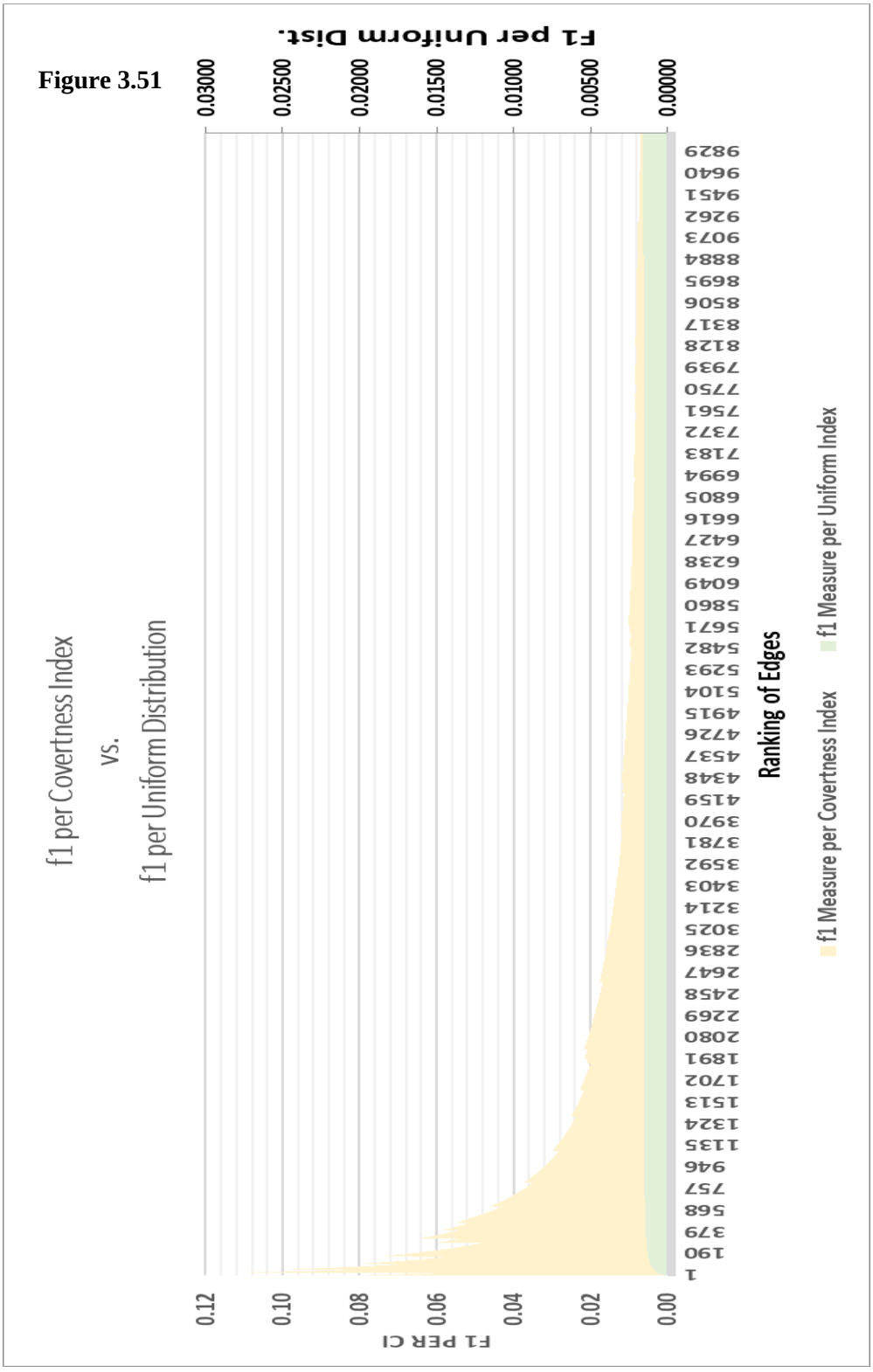


Figure 3.51

The results from all the three cases presented above are convergent. Further, even the individual metrics' performance from each of the cases compared between the two models point unequivocally towards the Covertness Index model's overall superiority. We are faced with the question of which threshold to choose – the top-ranked 2500 edges, the 5000 ranked edges, or the 10,000 top-ranked edges. The different metrics' performance for each of the cases is summarized in the table below for easier reference.

The table below proves that the choice of 2500 top-ranked edges gives the Covertness Index model. All the three performance metrics for the Covertness model, namely, Precision, Recall, and F1 Measure, are at least ten times their counterparts in the Uniform Index model. Additionally, the number of edges that are nothing but the relationships between the constituent nodes of the dyads under observation is also the least. The number of unique nodes within this set of 2500 edges is also the least.

Thus, the computational (and human) resources needed to surveil the network for covertness signs are least encumbered if we choose the smallest sized instance. It needs to be recalled at this stage that since we don’t know at the time of surveillance as to which of the actors are into covert acts or are trying to communicate in ways that allow information exchange to be confined, the entire instance of the network has to be scrutinized, that is, all the 2500 (or 5000 or 10,000) edges and the nodes between which the ties are formed need to be kept in focus. If the instance's size increases say from 2500 to 5000, the surveillance's complexity will increase in exponents of 2.

3.21 Conclusion

The concomitant complexities caused by the privacy laws, the difficulty in obtaining accurate field-based information, and the possible lack of information about the previous communication flow within the network will add to the layers of resource mobilization as the size of the instance goes on increasing. The hour's need is to optimize the bandwidth that the surveillance agency has with a decent trade-off with the results. In case we take up

the instance of 2500 top-ranked nodes, we end up with 26 of the 43 Edges of Interest (EoIs) required to complete the detection exercise. But, the edges that we have in hand, if scrutinized further and processed to produce narrower outputs regarding covertness, the loss is more than made up, as we will see in the subsequent section.

Case#	2500 edges			5000 edges			10,000 edges		
Model	Precision	Recall	F1 Measure	Precision	Recall	F1 Measure	Precision	Recall	F1 Measure
Covertness Index Model	0.0092	0.53	0.018	0.0054	0.63	0.011	0.0035	0.81	0.007
Uniform Distribution Model	0.0008	0.04	0.0016	0.0008	0.09	0.0016	0.0008	0.18	0.0016
Difference (in percent)	1150%	1325%	1130%	675%	700%	690%	440%	450%	440%

Table 3.18 Table showing the Precision, Recall and F1 Scores for the Covertness Index Model and the Uniform Distribution Model (the first two rows). The last row reflects the percentage improvement the metrics of the Covertness Model achieve over the metrics pertaining to the Uniform Model. It's important to note that the most significant improvement happens when the threshold value is 2500 i.e. the lower the threshold, the better the results in this case. This series of calculations leads the study to adopt the threshold of 2500, i.e. 2500 top ranked covert edges are selected for the next level of analysis.

Problem Statement in light of the Covertness Metric proposed

Based on the above analysis, we may revisit the problem statement that was first defined in section 1.7 of Chapter 1 of this dissertation and compare results obtained after applying the covertness index metric.

The ENRON mail corpus may be defined as a social network graph *G*, such that G = (*V*, *E*);

Where V is the set of all nodes in the graph network.
&E is the set of all edges or mail-pairs in the graph network, including those formed when copies of e-mail exchanges between pairs of nodes are marked to other nodes.

The number of edges in the network graph is represented as the cardinality of the set of edges V, i.e., $| V |$
& $| V | = 6568$.

The number of edges in the network graph is represented as the cardinality of the set of edges E, i.e., $|E |$
& $|E | = 55{,}300$.

Let's define the set of the employees of ENRON who were part of the scam as a graph G_C, such that $G_C = (V_C, E_C)$;

Where V_C is the set of all nodes of interest (NoIs) in the graph network and E_C is the set of all edges of interest (EoIs) in the graph network.

$G_C \subset G$ and $V_C \subset V$ & $E_C \subset E$;
& $| V_C | = 19$ and $|E_C | = 43$;

The ratio of the overall edges of interest (EoI) e_{ij} (i and j are nodes of interest (NoIs) in the graph network) to the set of all edges of graph G thus comes to:

$$P = |E_C| \ / \ |E| = 43/ \ 55{,}300 =$$

The problem may also be reframed in a probabilistic sense as is given below;

What is the probability of detecting at least one covert edge from amongst the overall set of edges of the ENRON e-mail network in 20 tries?

Let's define an integer $\boldsymbol{k}$, s.t., $\boldsymbol{k}$ = Number of tries; Here, $\boldsymbol{k}$ = 20.

There are 43 covert edges or Edges of Interest (EoIs).

Let's define the number of EoIs as $\boldsymbol{m}$; Here, $\boldsymbol{m}$ = 43.

The number of edges overall is 55,288 ~ 55,300

Let the total number of edges be defined as $\boldsymbol{e}$; Here, $\boldsymbol{e}$=55,288

We need to calculate the probability of not getting any covert edges in 20 tries.

Let's define the probability of detecting a covert edge as $\boldsymbol{P_c}$ and not detecting a covert edge as $\boldsymbol{P_{nc}}$.

The probability of not detecting a covert edge in the first try will be
(55,300 – 43)/ 55,300.
The probability of not detecting a covert edge in the second try will be
(55,299 – 43)/ 55,299.
In this manner, the probability of not detecting a covert edge on the 20th try will be
(55,280-43)/ 55,280.

Notationally,

$$\boldsymbol{P_{nc}} = \prod_{i=0}^{k-1} \frac{((e-i)-m)}{(e-i)}$$

$$\boldsymbol{P_c} = (1 - \boldsymbol{P_{nc}}) = \left(1 - \prod_{i=0}^{k-1} \frac{((e-i)-m)}{(e-i)}\right)$$

Hence, the probability of not getting a covert edge detected in 20 tries (k = 20) comes to:

$$\boldsymbol{P_{nc}} = \frac{(55{,}300 - 43)}{55{,}300}\ x\ \frac{(55{,}299 - 43)}{55{,}299}\ x\ \ldots .x\ \frac{(55{,}281 - 43)}{55{,}281}$$

$\boldsymbol{P_{nc}}$ = 0.984560175;$\boldsymbol{P_c}$= **(1 -** $\boldsymbol{P_{nc}}$**)** = 0.015439825

After Applying Covertness Index

After applying the Covertness Index to all the edges, the number of covert edges of interest (EoIs) comes to 23 in a selected-set of 2500 top-ranked covert edges.

Thus, after this part of the experiment, $\boldsymbol{m}$ = 23; $\boldsymbol{e}$=2500; $\boldsymbol{k}$ = 20.

Plugging in these values into equation (1) above, we get –

$$\boldsymbol{P_{nc}} = \frac{(2500 - 23)}{2500}\ x\ \frac{(2499 - 23)}{2499}\ x\ \ldots .x\ \frac{(2481 - 23)}{2481}$$

$\boldsymbol{P_{nc}}$ = 0.830638142; $\boldsymbol{P_c}$= **(1 -** $\boldsymbol{P_{nc}}$**)** = 0.169361858

Figure 4.70 below is a comparative graph that shows the enhancement in the probability of detecting covert edges in the overall set of edges in the ENRON mail corpus network.

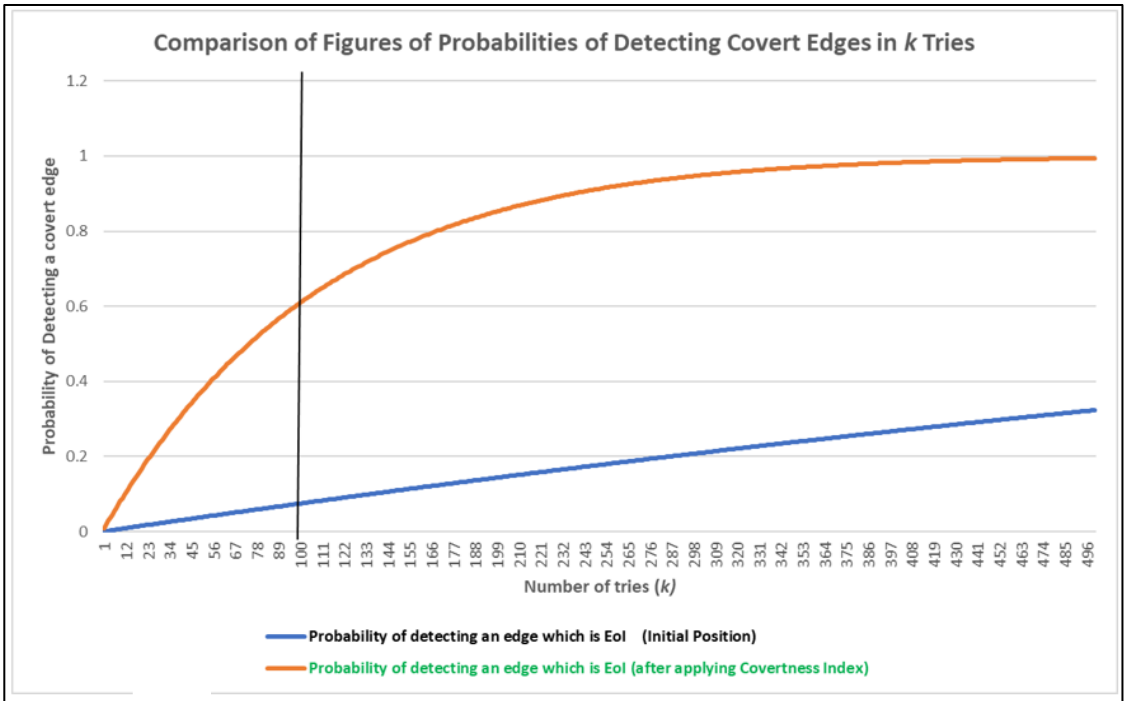


Figure 3.52 Graph showing comparison of plots reflecting the probability of identifying at least one covert edge (EoI) from amongst all the edges in the network in k tries where the value of k varies from 0 to 500. The blue line represents the plot of probabilities in a uniform distribution model and the line in orange shows the corresponding probability figures when the covertness index matrix is applied to the edges in the ENRON dataset. The vertical line shows the value of probability at $k = 20$.

Chapter 4

Detecting Collusion in Networks

4.1 Introduction

Thus far, the research has identified the basic building blocks of covertness in the form of ranking edges or ties between constituent nodes of dyads. Edges are very basic in their structure and cannot, in isolation, end the quest for identifying covert subnets or communities within the overall network. Ways need to be found to establish linkages between high ranked covert edges such that there is some commonness of purpose amongst them. The top-ranked covert edges derived in the previous section may exhibit covertness in their ties that arise from diverse causes. Some may relate to organizational policies like formulation of crucial sale strategies, for instance, which will remain confined amongst the employees who wouldn't want them to leak out to competitors; other edges might be confining information due to more mundane reasons such as marital affairs, mutual interests like gambling which might have an adverse impact once out in the open; others might simply be covert due to structural reasons, i.e., the employee nodes may be in isolated positions where there is not much chance of interaction, or due to inadequate information about information exchanges (As already analyzed earlier, the ENRON mail corpus comprises only 151 complete inboxes spanning a limited period, whereas analysis of the "To-From" headers of all emails reveals at least 4600 mail-ids). So, it's incumbent that after identifying high covertness in select edges, the next step should be to group these covert edges into groups or communities with identical covert aims. 'Similarity' in covert objectives may be termed the *commonness* of purpose for lack of a better term. A more technical term that one hears of more frequently in studies of criminal organizational networks is *common intention* (defined later in this study).

The endeavor amongst social network researchers to seek out commonness amongst actors (nodes) or even larger subgroups of actors is a well-established practice. Though the search

for conspiracy groupings in identified criminal organizations is common in law enforcement, and several field-based practices are available with policing organizations worldwide, its systematic study has started only recently. Field-based methods such as interrogation, forensic analyses of different crime scene artifacts, and compiling of history sheets containing surveillance records of known criminal elements, suspects, and other affiliates in criminal organizations have been commonplace in policing. There have been many articles on collecting and analyzing call records of suspects and potential offenders' telephones and cell phones. But the forays of social networks scientists into the domain of criminal networks and conspiracy-oriented structures are only a few decades old. The study in this domain that catches the eye is that of Baker and Faulkner (1994) who, in their landmark paper on conspiracy networks in the heavy electrical equipment industry, first dwelt upon concerted action in aid of a common (and illegal) action by groups or communities of actors have remarked in their abstract thus- *"We analyze the social organization of three well-known price-fixing* ***conspiracies*** *in the heavy electrical equipment industry. Although industrial organization economists and organizational criminologists have studied aspects of collusion, the organization of conspiracies has remained virtually unexplored. Using archival data, we reconstruct the actual communication networks involved in conspiracies in switchgear transformers and turbines. We find that the structure of illegal networks is driven primarily by the need to maximize* ***concealment****, rather than the need to maximize efficiency."*

The key terms used by them have been highlighted for effect. The mathematical contours of sociological terms such as "conspiracy," "collusion," "concealment," etc. are beginning to take shape through multi-faceted studies of criminal and terrorist networks. The terms "conspiracy" and "collusion" have, more often than not, negative and criminal connotations. Both terms have similar meanings. Merriam-Websters dictionary defines conspiracy in two different ways, (a) as an act "to join in a secret agreement to do an unlawful or wrongful act or an act which becomes unlawful as a result of the secret agreement" and (b) "to act in harmony toward a common end" and defines collusion as a "secret agreement or cooperation especially for an illegal or deceitful purpose." The terms to watch out for in the twin definitions are "secret agreement," "cooperation," "act in

harmony," and "common end," which also act as portals into a more mathematical interpretation of conspiracy and collusion. When we talk about building a computational model of conspiracy networks, what we are looking for is some sort of a formula to mathematically define the word "agreement" between the individual actors or entities within the network to work towards a common end or purpose and to ensure that the formula incorporates the fact that the "agreement" functions in a "concealed" manner. We've already come up with a way to quantify the "concealed" part, and this is the concept of the Covertness Index, which measures the confinement of information in a tie between a pair of actors in the network. The problem that now arises is how to implement the "agreement" part of it. In other words, how do we yoke together the covert ties into a set of "commonness," and all the covert ties or edge-pairs in the said set can be demonstrably proven to be working towards a "common" end.

4.2 Co-offending Networks

There has been a sustained research focus on collusion among participant nodes in instances of a covert enterprise. Recent work in identifying sub-structures of *co-offending networks* within larger criminal networks falls within this domain. Tayebi and Glasser (2016) have defined a co-offending network as " a network of offenders who have committed crimes together." The key importance of studying such networks has been increasingly at the core of academic research (Morselli. 2009, Hauck. Et al. 2002, McGloin et al. 2008, McGloin et al. 2009, Reiss 1988; Reiss, 1991). In Reiss (1988), "understanding co-offending is central to understanding crimes' etiology and the effects of intervention strategies." We may particularly note the word "together" in the definition by Tayebi and Glasser(2016). This is the term that approximates our quest for commonness or collusion; in other words, it approximates the concept of "working in harmony towards a common end"- the dictionary definition of "conspiracy."

The collusion aspect amongst actors in a criminal network has received special attention from Tayebi and Glasser (2016), who have defined it in co-offending networks. In their words,

> "A co-offending network is a network of offenders who have committed crimes together. With increasing attention to SNA, law enforcement and intelligence agencies have realized the importance of detailed knowledge about co-offending networks. Groups and organizations that engage in conspiracies, terroristic activities, and crimes like drug trafficking typically do this in a concealed fashion, hiding their illegal activities. In analyzing such activities, investigations do not only focus on individual suspects but also examine criminal groups and illegal organization and their behavior." (pp 10-11)

Having defined the act of co-offending, Tayebi and Glasser (2016, 20-27) elaborate upon the network structural properties of co-offenders' networks. They include characteristics such as **Degree Distribution**, which is the probability that a randomly selected node has a specified number of links: **Strength Distribution**, which is the number of collaborative acts (crimes) the co-offending nodes have participated in: **Connecting Paths** which determine if there are possible connections amongst the co-offenders and if so, what is the shortest path that links them using either the Shortest Path algorithm (Djikstra, 1959) or the Breadth-First Search (BFS) Algorithm: the **Clustering Coefficient**, which indicates the likelihood of an actor's collaborator to collaborate with that actor: **Connected Components Analysis**, which is premised upon the fact that there is a higher degree of connectivity within the co-offenders' group than outside the group and finally, **Network Evolution Analysis,** which reflects the dynamic nature of co-offending networks and the study of the evolutionary patterns of co-offending patterns spawned by such dynamicity helps investigators to identify these groups.

It should be noted here that Tayebi and Glasser (2016) have commented on the aspect of committing a crime together by a **network of actors in a concealed fashion**. They have also emphasized the need to **investigate entire groups of such actors rather than individuals**. Therefore, the scrutiny needs to be focused first on detecting networks of

actors co-offending with concealment being a key aspect of their actions and then identifying groups of such co-offending networks.

4.3 Community Detection

To tackle the next stage of the research question, i.e., how do we identify a group of covert entities (edges in our case rather than nodes) with "common ends" or form a conspiracy sub-network within the overall network. The problem is now akin to the class of problems referred to in Graph Theory as Community Detection. Fortunato (2010, p91), in his survey paper, defines a community as a "subgraph whose vertices have a higher probability of being connected to the other vertices of the subgraph than to external vertices." Schaeffer, in his survey on Graph Clustering (2007) defines a community as "a cluster in a graph" (p.31) and defines clustering thus – "Any nonuniform data contains underlying structure due to the heterogeneity of the data. The process of identifying this structure in terms of grouping the data elements is called clustering also called data classification. The resulting groups are called clusters. The grouping is usually based on some similarity measure defined for the data elements" (p.27).

The first problem in discovering communities in networks is looking for a quantitative formulation of the term 'community.' There is no consensus on how this term is defined in Social Network Analysis. The definition often depends on the specific network in question, and the specific domain one has in mind. However, the most common and baseline approach towards identifying communities inside a network remains rooted in the fact that there must be more edges inside the community than edges linking nodes belonging to the community with the rest of the network, which is the reference guideline at the basis of most community definitions. In his survey, Fortunato (2010) states that communities are algorithmically defined in most cases, i.e., they are just the final product of the algorithm, without a precise *a priori* definition.

A generally well-regarded property of a community that can be exploited algorithmically is connectedness. Connectedness refers to the extent, the constituent nodes within the community are linked to each other. A type of density measurement and density of a graph is perhaps the most widely used group-level index (Wasserman and Faust, 1994). It is trivial to expect that for any group of nodes to be a community, there must be links between each pair of its constituent nodes and that these links run only through nodes of the said group. Schaeffer (2007), who equates the concept of a community with a graph cluster, defines graph clustering as the task of grouping the vertices of the graph into clusters, taking into consideration the edge structure of the graph in such a way that there would be many edges within each cluster and relatively few between the clusters. He remarks that graph clustering in the sense of grouping the vertices of a given input graph into clusters is tied to the task of finding clusters within a given graph. Fortunato (2010) has categorized a community's definition into three classes, local, global, and based on vertex (node) similarity.

Local definitions focus on the sub-network in question, possibly including its immediate neighborhood but excluding the rest of the network. Wasserman and Faust (1994) identified four types of local criteria: complete mutuality, reachability, vertex(node) degree, and the comparison of internal versus external cohesion. The communities defined in this way are mostly maximal subgraphs, which cannot be enlarged with new nodes and edges without losing the property, which defines the community. Communities in a social network context can be defined in a very strict sense as subgroups whose members enjoy complete mutuality, i.e., they are all "friends to each other" (Luce and Perry, 1949), the equivalent of the definition of a *clique*[49]in graph-theoretic terms.

Global definitions of communities are those concerning the graph as a whole. Such definitions are reasonable for cases in which any prospective sub-group is an essential part of the network, which cannot be excised without seriously affecting the network

[49]a subset of a graph or network whose vertices are all adjacent to each other. In social network analysis, a clique is a maximal subgraph, i.e. which cannot be enlarged with the addition of new nodes and edges without losing the inherent property which defines it, whereas in graph theory cliques may also be non-maximal.

functioning. One of the network characteristics popularly used as a global property of communities is modularity, a concept introduced in a groundbreaking paper by Newman and Girvan (2004). In the standard formulation of modularity, a sub-network is a community if the number of edges inside the sub-network exceeds the expected number of internal edges that the same sub-network would have in the *null model*[50]equivalent to the network.

The third way that communities may be defined is based on the similarity of nodes (or vertices). It is trivial to assume that communities are groups of nodes that are, in some way, similar to each other and that the similarity between each pair of nodes can be computed based on some reference property, local or global, whether an edge connects them or not. Thus, each vertex ends up in a sub-group whose vertices are most similar to it. Several similarity measures have been used by researchers across various domains, the more popular ones of which include various *distance measures* (Euclidean, Manhattan, Mahalanobis, etc.), *dissimilarity measures* (posed by Wasserman and Faust, 1994), similarity measures such as *cosine similarity, neighborhood overlap* (whose normalized version is called a *Jaccard Index), Tanimoto coefficient*[51]*, Adamic-Adar metric, Pearson's Index*[52]*, Zhou-Lu-Zhang Index* (2009), *Preferential Attachment measure (PA)*[53], etc.

[50] In Graph Theory, a Null Model is a graph which matches the original graph which it seeks to replicate, in some of its structural features, but which is otherwise a random graph i.e. a graph in which all edges have equal probabilities of being connected to vertices. (This implies that a Random Graph doesn't have any community structure within it as density of connections is uniform throughout the graph). The null model is used for comparison with the parent-graph, to verify whether the parent-graph in question displays community structure or not. The most popular null model is the one proposed by Newman and Girvan in their landmark paper of 2004 and where they hypothesized about the null model consisting of a randomized version of the original graph, where edges are rewired at random, under the constraint that the expected degree of each vertex matches the degree of the vertex in the original graph.

[51]A variant of Jaccard Coefficient. The similarity score is the dot product of two vectors divided by the squared magnitudes of each of the vectors minus their dot product.

[52] Pearson's coefficient is a linear similarity measure which uses mean centering and normalization of profiles. It uses a best fit regression line that runs through the attributes of the two data objects, plotted on a two-dimensional plane. It is calculated by dividing the covariance of the two data objects by the product of their standard deviations.

[53] Preferential Attachment (PA) is a measure that calculates the product of the degree of two nodes, so the higher the degree of both nodes, the higher is the similarity between them.

Another sub-class of metrics in this category is the number of *edge-independent paths* between two vertices. Independent paths do not share any edges (vertices). Their number is related to the maximum flow that can be conveyed between the two vertices subject to the condition that each edge can carry only one unit of flow, which is also the formulation of the classic *max-flow/min-cut theorem* proposed by Elias, Feinstein & Shannon, (1956). This class also includes the sum of all paths between the nodes or vertices of a network, and the sum is termed as a *weighted sum of paths.* Another important class of measures of vertex similarity is based on the properties of *random walks* on graphs. (A random walk is a path across a graph or network constructed by taking random steps repeatedly. The 'walk' starts at some specific initial node or vertex; at each step of the walk, edges attached to the node or vertex are chosen randomly. The walker moves along the chosen edge to the node vertex at the other end, and the process gets iterated. Random walks are usually permitted to repeat the edges they move along more than once or retrace their steps through an edge they have just crossed.)

One of the other metrics in this class is the commute-time (Chandra et al., 1989) between a pair of vertices, which is the average number of steps needed for a random walker, starting at either vertex, to reach the other vertex for the first time and then return to the original vertex. The commute-time and then revert to the original vertex. The commute-time is closely related to the *resistance distance* introduced by Klein and Randic (1993), expressing the effective electrical resistance between two vertices if the graph is turned into an electrical network. Another quantity used to define similarity measures is the *escape probability*, defined as the probability that the walker reaches the target vertex before coming back to the source vertex. The escape probability is related to the effective conductance between the two vertices in the equivalent resistor network. (Palmer and Faloutsos, 2003; Tong et al., 2008). Fortunato (2010) describes other methods that exploit modified random walks' intrinsic properties, citing Gori and Pucci's instances and that by Tong and others who have used similarity measures derived from Google's PageRank process.

4.4 Approaches to Community Detection

The nature of communities within graphs or networks is described by Fortunato (2010); thus – "Real networks are not random graphs, as they display big inhomogeneities, revealing a high level of order and organization. The degree distribution is broad, with a tail that often follows a power law: therefore, many vertices with low degrees coexist with some vertices with a large degree. Furthermore, the edges' distribution is globally and locally inhomogeneous, with high concentrations of edges within special groups of vertices and low concentrations between these groups. This feature of real networks is called community structure". To detect such clusters within networks, three broad solutions are envisaged at this point, each with its concomitant advantages and disadvantages. The first one is the **Graph Partitioning** approach where the network is split into progressively smaller and more cohesive groups; the **Community Detection** approach is similar to graph partitioning but less structured in terms of partition size and number, and finally, the **Network Evolution** or **Link Prediction** approach which looks at how a network might look at a future point of time based on how it looked like in the immediate past. A related approach is the **Missing Link** approach, which looks for links between covert entities that might exist but not available based on available information due to deliberate deception by the participant entities or incomplete information about the network. We may look at identifying groups of edges having covertness related, i.e., groups of edges from within the set of top-ranked covert edges, which have a common aim. The resulting architecture of subgroups of covert edges may not have the structural quality to it because any path may not connect the constituents edges within the related subgroup but given substantial similarities; otherwise, their aims to keep information confined can be said to have a lot of commonness. For a better understanding of this concept, a representative network is shown in Figure[54]4.1 below:

[54] A more detailed explanation of how this figure evolved into its present structure is given in the section on Edge-Vertices which follows this discussion.

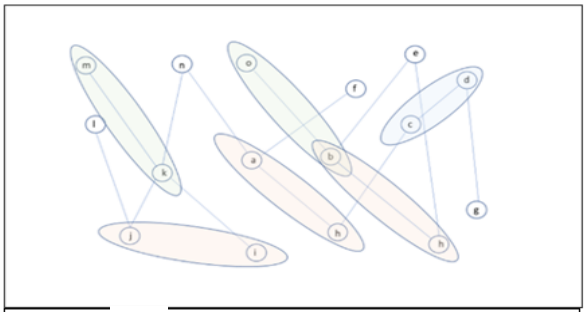


Figure 4.1 A representative network to illustrate the identification of groups of edges having covertness that is related i.e. groups of edges from within the set of top ranked covert edges which apparently have a common aim. The resulting architecture of subgroups of covert edges may not have a structural quality to it in the sense that the constituent edges within the related subgroup may not be connected by any path, but given substantial similarities otherwise, their aims to keep information confined can be said to have a lot of commonness. The pairs of nodes enveloped in different shades represent different subgroups of covertness.

The pairs of nodes highlighted in different colors represent different subgroups of covertness. Let's choose the green colored subgroup, which comprises the node pairs (*b,o*) and (*m,k*) and *vice versa*. In this sense, neither pair displays structural equivalence. Rather, their similarities belong more to the regular equivalence domain.

With this architecture in mind, we may consider constructing covert subgroups from two different approaches: a) A top-down approach and b) An aggregational approach beginning from the ground up.

A top-down approach would entail dividing the graph into smaller sized partitions and evaluating each partition for cohesiveness using different statistical methods.[55] The two main mechanisms available for this type of computational intervention are a) graph partitioning and b) community detection. The fundamental difference between the two is that in graph partitioning, the groups' number and size into which a network is divided are specified. In community detection, they are not. There is also a difference between the goals of the two types of calculations. Graph partitioning aims to divide a network into smaller, more manageable pieces, such as numerical calculations. On the other hand, community detection is more of a tool for understanding the overall structure of a network or "for shedding light on large-scale patterns of connection that may not be easily visible in raw network topology."

[55] Please see the section on Social Network Analytic methods, especially the section on Group based analytics in social networks.

4.4.1 Graph Partitioning Approaches

Graph partitioning algorithms seek to find the best division of a network, given certain conditions, regardless of whether any good division exists. If there are indeed no partitions that meet the specifications, then the least bad one must be accepted. On the other hand, with community detection, where the goal is normally to comprehend the network structure, there is no need to partition the network if there is no good division. If a network has no suitable divisions, then that in itself may be a useful piece of information, and it would be perfectly reasonable for a community detection algorithm only to divide up networks when good divisions exist and to leave them undivided the rest of the time. (Newman, 2010, pp 357-358).

Coming to the graph partitioning algorithms, as has been mentioned earlier, a **graph partition** is defined as the decomposition of any network to smaller sub-networks by dividing its set of nodes into mutually exclusive groups. Edges of the original network that cross between the resultant groups will also produce edges in the partitioned sub-network. Graph partitioning aims to ensure that the number of resulting edges is small compared to the original network. The partition may be better optimized for analysis and problem-solving than the original network. There are a variety of graph partitioning algorithms available. The simplest graph partitioning problem is the division of a network into just two parts called *graph bisection*. Most graph partitioning algorithms are, in fact, mechanisms for bisecting networks rather than for partitioning them into arbitrary numbers of sun-networks. Though this aspect may at first appear to be a disadvantage, it is not in practice since if a network can be divided into two parts, it can be divided into more than two by further partitioning one or both of the resultant partitions. This repeated bisection is the commonest approach to the partitioning of networks into arbitrary numbers of parts.

Further illumination into the bisection mechanism is provided by Newman (2010), who states:

> "Formally, the graph bisection problem is the problem of dividing the vertices of a network into two non-overlapping groups of given sizes such that the number of

edges running between vertices in different groups is minimized. The number of edges between groups is called the *cut size*. Simple though it is to describe, this problem is not easy to solve. One might imagine that one could bisect a network simply by looking through all possible network divisions into two parts of the required sizes and choosing the one with the smallest cut size. For all but the smallest of networks, however, this so-called *exhaustive search* turns out to be prohibitively costly in terms of computer time."(p. 359)

Newman dwells on the algorithm's complexity that he derives as an exponential function (2n+1) where n is the network's size. This value can escalate quickly as the network size increases. More to the point is Newman's remarks on the nature of the solution we might come to expect: "Perhaps one might find a way to limit one's search to only those divisions of the network that have a chance to be the best one."(Newman, 2010, p360). This viewpoint is a truism when we visit real-world problems of community detection in a network, and it will apply to every kind of mechanism that research may come up with it. It is also very true of the solution proposed in this study.

The graph partitioning algorithm family includes the Kernighan-Lin algorithm proposed by Ben Kernighan and Shen Lin in 1970 (Kernighan et al. 1970), which works by reducing the number of edges between the sub-networks by exchanging nodes between them. Other popular algorithms in this class are spectral partitioning proposed by Fiedler (1973) and Pothen et al. (1990), multi-level graph partitioning (Teng, 1999), (which include tools like METIS (Karypis et al. 1999), Graclus (Dhillon et al. 2007) and MLR-MCL (Satuluri et al. 2009)).

4.4.2 Community Detection Approaches

Community detection is used essentially for discovering and comprehending the overall structure of the network. It is similar to graph partitioning because the goal is to divide the network into nodes with minimum links amongst them. More importantly, the number or size of these groups or communities is not pre-decided. In this category, popular techniques

include simple modularity maximization, which is similar to the Kernighan-Lin algorithm (Newman, 2006a, 2006b), and the spectral modularity maximization algorithm, which is the equivalent of the spectral graph partitioning algorithm. Other algorithms that may be said to belong to the class of community detection are the Markov clustering mechanisms or MCL proposed by Dongen et al. (2000) and upgraded subsequently (Satuluri et al., 2009, 2010), related approaches like the local graph clustering (Speilman & Srivastava, 2008, Speilman & Teng, 2004), flow-based post-processing (maximum computing flow to improve existing partitions) proposed by Flake et al. (2000) and a related conductance- based derivative by Lang & Rao (2004), community discovery via Shingling (Broder et al., 1997), betweenness-Based methods developed by Newman et al. (2004) which look for the edges that lie between community like structures and seeks to remove them[56], short loop betweenness based algorithm (Radicchi *et al.* 2004), hierarchical clustering (and hierarchical decomposition) techniques which together comprise one of the oldest set of approaches to the subject and many others.

There are a great number of clustering techniques available in the literature. Which one(s) are the handiest? The literature has not resolved this question, and it remains an area of little consensus amongst researchers in the subject. What can perhaps be said with some certainty is that the most popular are the modularity optimization[57] techniques. Still, the results of this approach in large graphs(networks) are likely to be unreliable. Fortunato, in his survey paper on community detection (2010, p91), has commented on the popularity of this class of algorithms stating thus –"Nevertheless, people have become accustomed to using it (Modularity Optimization), and there have been several attempts to improve the measure. A newcomer, who wishes to find clusters in a given network and is not familiar

[56]Interestingly enough, this methodology is based on an attribute defined on an edge called the edge-betweenness that counts the number of geodesic paths that run along edges and edges that lie between sub-networks are expected to have high values of betweenness. The algorithm iteratively calculates the edge-betweenness of all edges and ranks them. The edge with the highest score is removed and the algorithm goes on iteratively calculating new scores of edge betweenness and progressively divides the network into two parts, then three and so on. This is somewhat similar to the Covertness Index attribute defined in this study and also has a parallel in ranking the edges as per the attribute vales.

[57]Modularity has a range of values between −0.5 (non-modular clustering) and 1 (fully modular clustering) that measures the relative density of edges inside communities with respect to edges outside communities. Optimizing the value of Modularity results in the best possible grouping of the nodes of a given network..

with clustering techniques, would not know, off-hand, which method to use, and he/she would hardly find indications about good methods in any single paper on graph clustering, except perhaps on the method presented in the paper. So, people keep using algorithms because they have heard of them, or because they know that other people are using them, or because of the reputation of the scientists who designed them. Waiting for future reliable benchmarks that may give an objective assessment of the quality of the algorithms, there are at the moment hardly solid reasons to prefer one algorithm to another" the comparative analyzes by Danon et al. (2005) and by Lancichinetti and Fortunato (2008).

4.4.3 Link Based Approaches

In the previous sections, I proposed a Covertness Index to build a rank of covert edges. The primary task was to define an attribute centered upon the confinement of information exchanged between two nodes. Thus, the study has considered the ties between the nodes as a fundamental and indivisible building block of covertness. Edges have attributes of being two-dimensional as well as behaving like an atomic unit. But, this solves the questions addressed by this research only halfway.

Suppose that we have a group of edges denoted as covert; how do we know if these edges have a type of covertness common? To rephrase the question, the methodology to cluster edges established through metrics to be oriented covertly. At the beginning of the study, examples were given where groups of covert actors in a network whose intentions[58] were not common. But they were confining information for diverse purposes. To segregate the purposes for which the constituent nodes of a dyad perform in an opaque manner is a different dimension from the original covertness issue. It is a well-known phenomenon in social networks that nodes which cluster into groups are more closely linked than nodes outside the group, which is true of the covertness property. The sequence of three figures below captures this dilemma somewhat more expressively.

[58]Common Intention may be defined as a prearranged plan and acting in concert or in some kind of social cohesion pursuant to the plan.

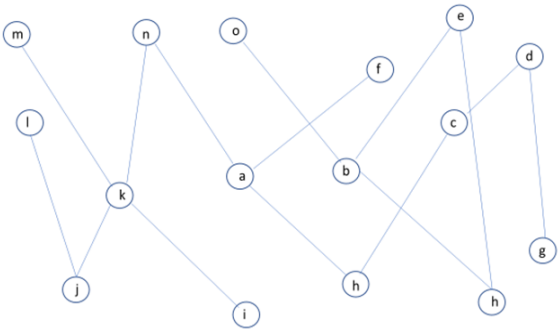


Figure 4.2 (a)

The first figure shows a general social network whose ties are under scrutiny to discover their covertness index. The second figure shows the same network's status after the exercise to calculate the covertness index of ties, and ranking them as covert or otherwise based on a certain threshold is completed. The third figure in the series represents attempts to group the covert nodes into community structures having common aims or intentions.

The set of edges (and the associated Edge-Vertex functions) which are seen in the above representative network are:

Ev={(a,f), (a,h), (a,n), (b,e), (b,o), (b,h), (c,d), (c,h), (d,g), (e,h), (i,k), (j,k), (j,l), (k,m), (k,n)}

After computing the Covertness Index on each of the above edges and then assigning each edge a rank depending on the covertness index's value, we get the following figure. In this figure, all the top ranking edges are highlighted in pale red. The set of identified covert edges are listed below:

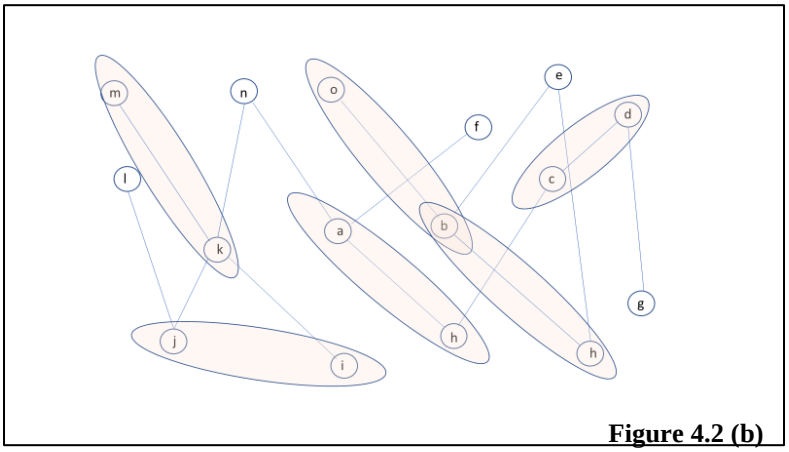


Figure 4.2 (b)

$(E_v)_{covert} = \{(a,h),(b,o),(b,h),(c,d),(i,k),(k,m)\}$

The third and last in the figures show what we aspire to do in the next step: finding links between the edges identified as covert to somehow group them in subnets representing common intention.

A third of the figures below show that there are three sub-groups of covert edges. Each sub-group is colored differently. It needs to be emphasized that some of the edges of the nodes constituting the edges are not connected structurally. The linkages have been made through other associations. The three different sets of covert edges are shown below:

$$[(E_v)_{covert}]_{Green} = \{(k,m), (b,o)\}$$

$$[(E_v)_{covert}]_{Red} = \{(a,h), (b,h), (i,j)\}$$

$$[(E_v)_{covert}]_{Blue} = \{(c,d)\}$$

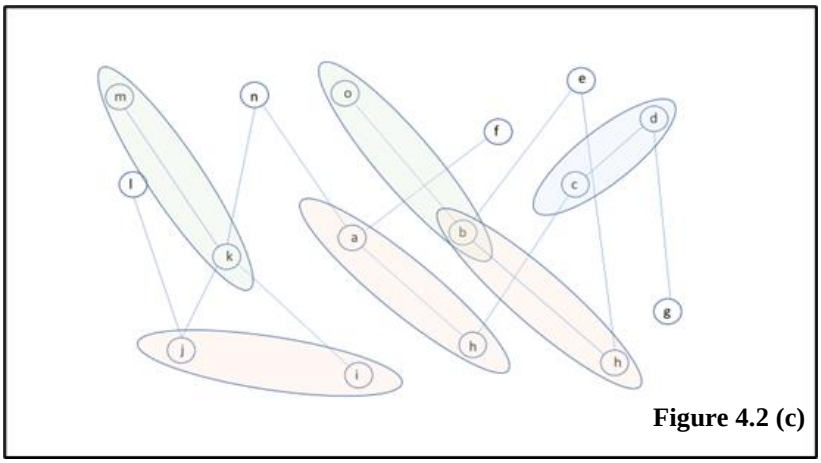


Figure 4.2 (c)

Mere confinement of information being exchanged is too wide and too shallow to measure group actors (or nodes) who display covertness characteristics but whose aims to maintain covert attitudes are widely variable. So the task now is to make groups (communities) out of these covert edges so that the groups have common intentions, i.e., their covertness display has a common aim. Community detection already exists as a topic within the overall research on social networks. It is a widely studied and researched topic, and there is a correspondingly wide variety of algorithms for community detection (Newman, 2015, p354).

4.5 Link-Prediction in Social Networks

Social networks are highly dynamic; they grow and evolve quickly by adding new nodes and edges, indicating new interactions in the existing social structure. As discussed in some previous sections, quantifying the mechanisms through which such networks evolve is still not fully comprehended. The second part of this research is to define and convert into a measurable format a basic mathematical and computational problem underpinning the evolution of social networks, the link-prediction problem. Link prediction in social networks is defined as the probability of establishing a link between two disparate (un-connected) nodes.

Liben-Nowell and Kleinberg (2007) were one of the first to propose the Link Prediction model. Given a social network snapshot, the Link Prediction approach seeks to predict the edges added to the network accurately. In their words – "In effect, the link-prediction problem asks: to what extent can a social network's evolution be modeled using features intrinsic to the network itself? Consider a co-authorship network among scientists, for example. There are many reasons exogenous to the network why two scientists who have never written a paper together will do so in the next few years. For example, they may happen to become geographically close when one of them changes institutions. Such collaborations can be hard to predict. But one also senses that many new collaborations are hinted at by the topology of the network: two scientists who are *close* in the network will have colleagues in common and will travel in similar circles; this social proximity suggests that they are more likely to collaborate shortly. Our goal is to make this intuitive notion precise and understand which proximity measures in a network lead to the most accurate link predictions. We find that several proximity measures lead to predictions that outperform chance by factors of forty to fifty, indicating that the network topology does indeed contain latent information from which to infer future interactions. Moreover, certain fairly subtle measures—involving infinite sums over paths in the network—often outperform more direct measures, such as shortest-path distances and numbers of shared neighbors."(Liben-Nowell et al., 2007)

Link prediction modeling is not limited to predicting social-network evolution; it is also applicable across a wide variety of application areas; to begin with, it is very relevant to several interesting, current applications of social networks (Cao, Z., Zhang, Y., Guan, J., & Zhou, S., 2018). To give an example, in online social networks, such as Twitter, Facebook, and Weibo, link prediction is used to recommend registered users to connect with someone they are acquainted with but were not able to recognize in the network (Zhang, Y., Zheng, Z. and Lyu, M. R., 2014; Aiello, L. M. 2012; Zhang, Z. K., Zhou, T. & Zhang, Y.C., 2011). Accurate recommendations through link predictions are likely to promote user loyalty in personalized services (Zhang, Y., Zheng, Z. & Lyu, M. R. 2014).

In Bioinformatics[59], it has been used in protein-protein interaction (PPI) prediction (Airodi et al.,2006) or to annotate the PPI graph (Freschi, 2009).

Accurate link prediction based on known network structure and some specific biological information help design targeted experiments, which may substantially reduce experimental time and cost (Bu, D. B., Zhao, Y. & Cai, L. 2003, Stumpf, M. P. H., 2008, Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. 2013; Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T., 2013). The Internet and web science area can be used in tasks like automatic web hyperlink creation (Adafre et al., 2005) and web site hyperlink prediction (Zhu et al., 2002). In e-commerce, one of the most prominent usages of link prediction is to build recommendation systems (Huang et al., 2005; Liu et al., 2007; Li et al., 2009). It also has various applications in other scientific disciplines. For instance, in bibliography and library science, it can be used for de-duplication (Malin & Karley, 2005) and record linkage (Ahmed et al., 2007).

4.6 Link-Prediction in Missing Link Problems

In the application to monitoring the network of criminals, link prediction is used to discover the possible connections among criminals (including potential criminals), which is useful for locating some specific criminals and thus detecting and disrupting terror attacks (Knoke, 2015; Li, 2014).

Increasingly, for example, researchers in artificial intelligence and data mining have argued that a large organization, such as a company, can leverage the interactions within the informal social network among its members; these ties serve to supplement the official hierarchy imposed by the organization itself (Krautz et al., 1997; Raghavan, 2002). Liben-Nowell and Kleinberg (2007) have noted that effective methods for link prediction could

[59] However, in the implementation of Link Prediction modeling in biological networks, including inprotein-protein interaction networks and metabolic networks, the links are found to be typically neither complete (high false positives) nor highly reliable (high false negatives)(Cao, Z., Zhang, Y., Guan, J., & Zhou, S., 2018; Maslov, S. & Sneppen, K. 2002; Yu, H. Y., 2008; Jeong, H., Mason, S. P. & Barabási, A. L.; 2001).

be used to analyze such a social network to suggest "promising interactions or collaborations that have not yet been identified within the organization."

It can be used in security-related applications to identify terrorists and criminals (Al Hasan & Zaki, 2011). Liben-Nowell and Kleinberg (2007) have gone a step further to include the analysis of covert networks; in their words, "In a different vein, research in security has recently begun to emphasize the role of social network analysis, largely motivated by the problem of monitoring terrorist networks; link prediction in this context allows one to conjecture that particular individuals are working together even though their interaction has not been directly observed." which is typically the approach that has been adopted in several studies on clandestine networks, specifically the ones involving the 9/11 attacks. (Krebs, 2002).

This approach to link prediction in networks has also been extrapolated by Liben-Nowell and Kleinberg(2007) by applying it to the problem of inferring missing links from an observed network:

> "in several domains, one constructs a network of interactions based on observable data and then tries to infer additional links that, while not directly visible, are likely to exist. This line of work differs from our problem formulation in that it works with a static snapshot of a network, rather than considering network evolution; it also tends to take into account specific attributes of the nodes in the network, rather than evaluating the power of prediction methods that are based purely on the graph structure." (p.2)

Link prediction modeling is only one of the mechanisms to tackle the larger problem of network evolution. There are a plethora of models dealing with network evolution in recent research on networks. Prime examples would be the work of Barabasi et al. (2002), Davidsen et al. (2002), Jin et al. (2001) on collaboration networks, and the survey of Newman (2003). It is difficult to evaluate and compare the different models used in these works. They have generally been evaluated only in terms of how well they reproduce certain global structural features observed in real networks. However, link prediction

modeling offers a very natural basis for such comparative evaluations: a network model is useful to the extent that it can support meaningful inferences from observed network data (Liben-Nowell & Kleinberg, 2007). Newman (2001) has a similar approach in as much as he has considered the correlation between certain network-growth models and data on the appearance of edges of co-authorship networks. Hasan et al. (2006) have extended this work in two ways. First, they showed that data external to the scope of graph topology could significantly boost the prediction results. Secondly, they employed different similarity measures as features in a supervised learning setup where the link prediction problem is binary. This supervised classification approach to link prediction modeling gained further momentum in work by Bilgic et al. (2007), Wang Chao et al. (2007), and Doppa Janardan et al. (2009).

A similar but not identical area of application of link prediction modeling (and thus where explicit graph representations were not used) has been in the domain of relational data (Tasker et al., 2003, Popescul et al., 2003, Popescul et al., 2003) and also in the Internet domain (Sarukkai et al., 2000). According to Hasan et al. (2008), the prediction system proposed in these works can accept any relational dataset. The objects in the dataset are related to each other in any complex manner. The system's task is to predict the *existence* and the *type* of links between a pair of objects in the dataset. Proximal applications have been seen in Stochastic relational models (Airodi et al. 2006, Weiss et al. 2004, Xu et al. 2008, Fu et al., 2009), Probabilistic relational models (Getoor et al. 2002), graphical models (Nallapati et al.,2008), and other variations of such studies. The advantages of these approaches include the genericity and ease with which they can incorporate the model's entities' attributes. On the downside, they are usually complex and have too many parameters, many of which may not be intuitive to the user (Hasan et al., 2008).

4.7 Challenges in Using Link-Prediction

Clearly, there is a close relationship between the study of network evolution and link prediction modeling. Both research areas have seen much progress in their respective fields. An evolution model predicts the edges of a network at some future point in time, factoring in the widely available social network attributes, such as the small-world phenomenon (Kleinberg, 2000) and the power-law degree distribution (Barabasi et al., 1999). The two are not the same; however, the main difference between the social network evolution model and the link prediction model is that the former concentrates on the global properties of the network, while the latter focuses on the network's local parameters to predict the existence of a link between a certain pair of nodes in the network. Nevertheless, the ideas in the network evolution model have been key to research works that have directly addressed the problem of link prediction (Kashima et al., 2006).

One of the main challenges in link prediction is the evolution of huge networks in size and highly dynamic for which earlier algorithms may not scale well. For example, internet-scale social networks like WhatsApp, Instagram, Linked-In, Facebook, mySpace, Flickr, and so on pose special challenges during their extremely fast evolution. More direct approaches are required to address these challenges. For example, using the timestamps of past interactions, which explicitly use the lineage of interactions, can significantly improve link prediction efficiency (Tylenda et al., 2009). Another direct approach, matrix factorization, has been employed to estimate the similarity between the nodes in real-world social networks having approximately 2 million nodes and 90 million edges (Song et al., 2009). Matrix-based factorization mechanisms have also been applied to higher-order models, such as tensors. Any traditional algorithm that aims to compute pair-wise similarities between such a big graph's vertices is doomed to fail (Hasan et al., 2008).

4.8 Network Evolution vis-a-vis Link-Prediction

We have seen the link prediction mechanism is inextricably linked with the dynamics of network evolution. Essentially, any analysis of link prediction applications is predicated upon a time-based approach tied to the truth that network evolution is itself a time-dependent process. If link prediction is used to predict the network's future shape, time becomes an essential factor in the calculations. A modified example from Liben-Powell and Kleinberg (2007) to work using their notation illustrates the importance of time in link prediction applications.

Example: Given a social network G (V, E) where V is the set of vertices and E the set of edges, in which an edge $e = (u, v) \in E$ represents some form of interaction between its endpoints at a particular time t (e), we can record multiple interactions by parallel edges or by using a complex timestamp for an edge. For time $t \leq t'$, we assume that G [t, t'] denotes the subgraph of G restricted to the edges with time-stamps between t and $t.'$ In a supervised training setup for link prediction, we can choose a training interval[t_0, t_o'], and a test interval[t_1, t_1'] where$t_0' < t_1$. The link prediction task is to output a list of edges not present in$G$ [t_0, t_o'], but are predicted to appear in the network G [t_1, t_1'].

Hasan et al. (2008) extrapolated this approach and modeled the link prediction problem as a supervised classification task, where each data point corresponds to a pair of vertices in the social network graph. They proposed using the information from the training interval ([t_0, t_0']). The model belonging to this timestamp seeks to predict the future links in the test interval ([t_1, t_1']) are sought to be made. Assuming that $u, v \in V$ are two vertices in the graph G (V, E) and the label of the data point$\langle u, v \rangle$ is $y_{u,v}$ and further assuming that the interactions between u and v are symmetric, the pair$\langle u, v \rangle$and$\langle v, u \rangle$represent the same data point: hence, $y_{u,v} = y_{v,u}$.

$$y^{\langle u,v \rangle} = \begin{cases} +1, & \text{if } \langle u, v \rangle \in E \\ -1, & \text{if } \langle u, v \rangle \notin E \end{cases}$$

Using the above labeling for a set of training data points, a classification model is built that can predict the unknown labels of a pair of vertices$\langle u, v\rangle$where$\langle u, v\rangle \in E$ in the graph G [t_1, t_1']. This is a typical binary classification task. Any popular supervised classification tool, such as naive Bayes, neural networks, support vector machines, or k nearest neighbors, can be used. The major challenge in this approach is choosing a set of features for the classification task. Feature sets that have been used successfully for supervised link prediction tasks are discussed next.

In essence, the model proposed above is dedicated to predicting how a social network will evolve over time. An earlier timestamp snapshot of the network is taken as the basis for working out the same network structure slightly later. This approach may not fit for the study here as we are not looking for a network's future evolution. Rather, we are trying to determine whether two entities (edges and their associated Edge-Vertices in this case) are linked in some manner that may not be structurally discernible, i.e., there may not be any visible ties amongst them. Hence, though the context of the problem studied in this paper is somewhat similar, it is not identical. The situation that presents itself is somewhat more static in that *we are trying to infer the present state of the network, presuming that many of the links or ties which should otherwise have been there are missing*.

What we have here is an e-mail network of an organization (ENRON) where the covert dyads (or edges) have been identified through the mechanism of ascertaining how much of the information exchanged between the constituent nodes of the dyad has been confined (the Covertness Index of the tie between them essentially). Further analysis is required to determine if these edges have any links with each other, i.e. if there is any common intention that binds together several of these covert edges. As explained earlier, one of the keys and defining features of covert networks are pre-existing ties or relationships that may have existed between the actors at some point in the past but that are not observable now. A related feature is a deception inherent to the relationships between the actors where they try and hide their information exchange. In this study, these features which exist between nodes in a covert relationship are extrapolated to the edges, which are considered the basic units instead of nodes.

4.9 Link-Prediction in Covert Networks

Earlier, we saw link prediction models applied to identify criminals and terrorists in clandestine or covert social networks. The second part of this study attempts to apply the strategy of individual node-based link prediction modeling to the edges derived in the first part. That is to say; we've selected a definite number of edges between the constituent nodes of dyads based on their Covertness Index scores and also based on a heuristic threshold value of the index. The challenge arises when using the link prediction techniques to build links between these covert edges and thereby build larger covert subgroups of related edges. There is no longitudinal dimension (i.e., timestamp-based analysis) to this problem, unlike the one illustrated above by Liben-Nowell and Kleinberg (2007). Covert network analysis necessarily relies on the information already at hand to essentially fill in the structural holes (missing links) in the network and present a complete picture. The objective is not to predict how the network might look at some future point in time, but rather how it should look if the missing ties are reinstated.

Berlusconi et al. (2016) used this type of link prediction application in a non-timestamp based network analysis where a similarity index was used to construct missing links between the actors in a criminal network (the Oversize drug trafficking network in Italy, to be precise). Links that merited study were successfully predicted between actors with strong commonness, which is largely the direction this study takes. The only difference, it needs to be emphasized is that this work uses Edge Vertices rather than nodes (or actors) as the basic building block of covert subnets.

As discussed above, there are many different techniques to implement link prediction. Broadly, however, such techniques may be divided into feature-based and probabilistic mechanisms. Each technique has inherent strengths and weaknesses. The cardinal principle in applying models for link prediction, according to Liben-Nowell and Kleinberg (2007), is that "*a network model is useful to the extent that it can support meaningful inferences from observed network data*." This statement has motivated several studies in link prediction modeling, most notably the one carried out by Kashima et al. (2006), who

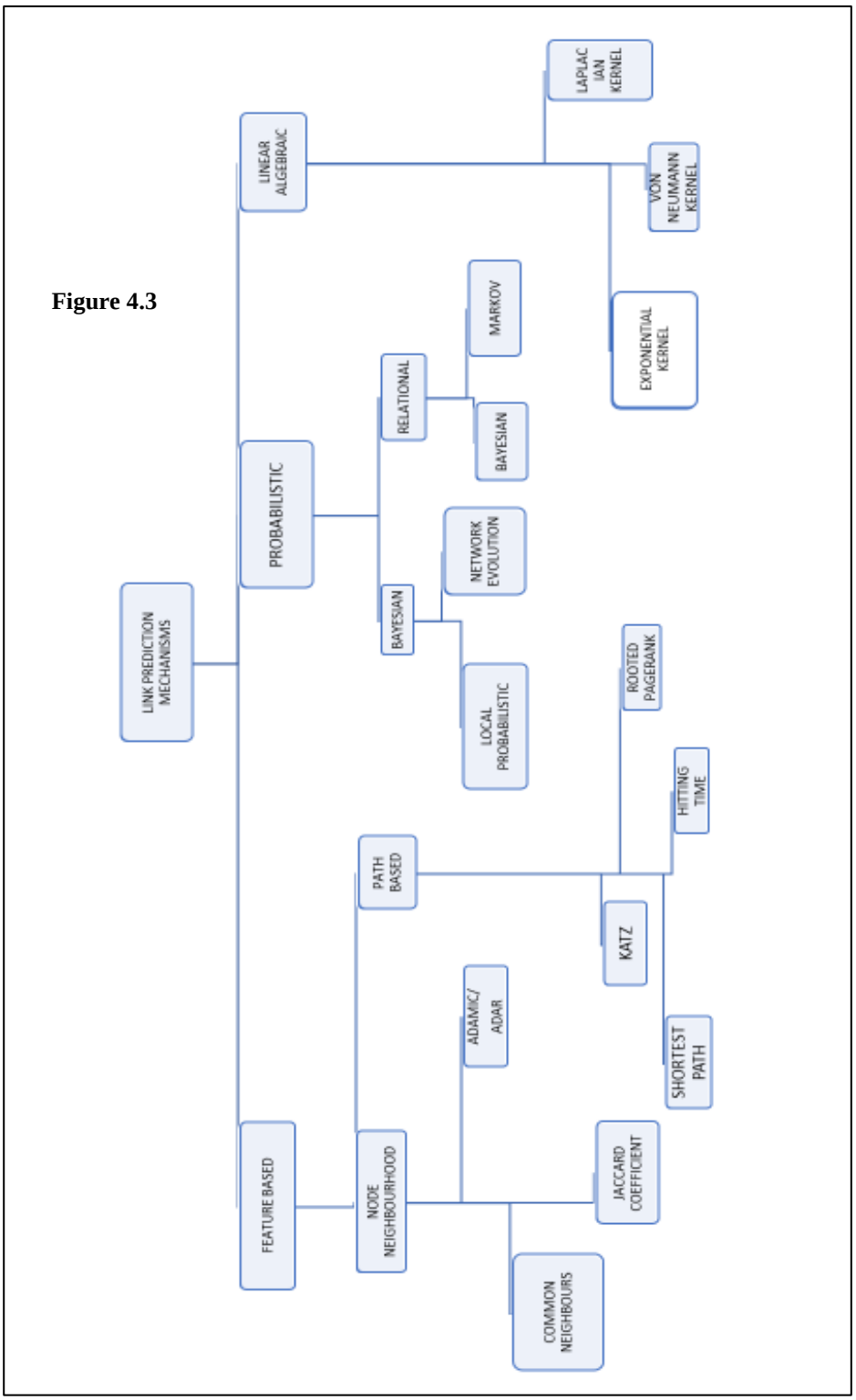
proposed the development of tunable parameters within the network model, which allows the construction of learning algorithms for link prediction, leading to improved accuracy of prediction. The link prediction problem is usually described in terms of a binary classification issue or a ranking issue on node pairs. Thus, link prediction is based on two types of information available in the network, first, on the information about the nodes themselves and second, information on the network topology as a whole. Probabilistic approaches to link prediction, on the other hand, involve supervised models that use Bayesian concepts. The main idea here is to obtain a posterior probability that denotes the chance of co-occurrence of the node pairs of interest. An advantage of such a model is that the score itself can be used as a classification feature. Figure 4.3 shows the different techniques in a high-level view.

4.10 Feature Based Link-Prediction

The notations used in describing the feature-based mechanisms are all adapted from Hasan & Zaki (2011)[60], who described the methodologies in some detail in their exhaustive survey of link prediction in social networks. This research reviews some of the more popular methods that could potentially be used to design a similarity measure to detect links between top-ranked (i.e., candidate covert) pairs of edges.

[60]**Notation**. Typically, small letters, like *x, y, z* are used to denote a node in a social network, the edges are represented by the letter *e*. For a node *x*, *Γ(x)* represents the set of neighbors of *x*. *degree(x)* is the size of the *Γ(x)*. The letter *A* is used for the adjacency matrix of the graph.

Figure 4.3


LINK PREDICTION MECHANISMS
LINEAR ALGEBRAIC
LAPLACIAN KERNEL
VON NEUMANN KERNEL
EXPONENTIAL KERNEL
PROBABILISTIC
RELATIONAL
MARKOV
BAYESIAN
BAYESIAN
NETWORK EVOLUTION
LOCAL PROBABILISTIC
PATH BASED
ROOTED PAGERANK
HITTING TIME
KATZ
SHORTEST PATH
FEATURE BASED
NODE NEIGHBOURHOOD
ADAMIC/ ADAR
JACCARD COEFFICIENT
COMMON NEIGHBOURS

Features are network attributes that are based on graph topology[61]. Features usually represent some sort of proximity between pairs of vertices or nodes as nodes are deemed to be the fundamental unit of a network upon which models need to be built. Node and edge attributes play an important role in link prediction. It needs to be noted that in a social network, the links are directly motivated by the utility of the individual representing the nodes, and the utility is a function of vertex and edge attributes. Many studies (Hasan et al. .2006, Doppa et al., 2009) have shown that using node or edge attributes as proximity features could significantly improve link prediction performance. Forexample, Hasan et al. (2006) showed that attributes such as the degree of overlap among the research keywords used by a pair of authors were the top-ranked attributes for link prediction in a co-authorship social network dataset. In this study, the vertex (node) attribute was the research keyword set, and the presumption was that two authors are proximal to each other if their research work centers around a large set of common keywords.

Similarly, the Katz metric computes the similarity between two web pages by the degree to which they have a larger set of common words where the words in the web page form the node attributes. The advantage of such a feature set is that it is generally computationally less complex and relatively easier on the resources. On the flip-side, the features are highly domain-specific, requiring very good domain knowledge to identify them. Link prediction based on feature sets computes similarity based on node neighborhoods or ensembles of paths between two nodes.

A feature-set based approach usually offers generic advantages, and no domain knowledge is necessary to compute the values of these features from a social network. Many significant studies are completely based around feature-sets (36, 29,22). Graph topological features fall into two broad categories – (a) node neighborhood-based and (b) path-based.

[61]Network Topology is the way in which the nodes and edges are arranged within a network. Topological properties can apply to the network as a whole or to individual nodes and edges. Some of the most used topological properties and concepts are Degree, Shortest Path, Scale Free Networks, Transitivity and the Centralities.

4.11 Node-Neighborhood based Features

4.11.1 Common Neighbors

The size of the set of common neighbors for two nodes, *x* and *y* defined as $|\Gamma(x) \cap \Gamma(y)|$. The idea of using the size of common neighbors is just an attestation to the network transitivity property. In simple words, it means that in social networks, if vertex *x* is connected to vertex *z* and vertex *y* is connected to vertex *z*; then, there is a heightened probability that vertex *x* will also be connected to vertex *y*. Thus, as the number of common neighbors grows, so does the chance that *x* and *y* will have a link between them. Newman (2001) has computed this quantity in the context of collaboration networks to show that a positive correlation exists between the number of common neighbors of *x* and *y* at time *t* and the probability that they will collaborate in the future.

4.11.2 Jaccard Index

The common neighbors metric is not normalized, so one can use the Jaccard Coefficient, which normalizes the size of common neighbors as below:

$$\textbf{Jaccard-coefficient}(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Conceptually, Jaccard defines the probability that a common neighbor of a pair of vertices *x* and *y* would be selected if the selection were made randomly from the union of the neighbor-sets of *x* and *y*. So, for a high number of common neighbors, the score would be higher. However, from their experimental results on four different collaboration networks, Liben-Nowell et al. (2007) showed that the Jaccard coefficient performs worse than the number of common neighbors.

4.11.3 Adamic/Adar

Adamic and Adar (2003) proposed this score as a metric of similarity between two web pages. For a set of features, it is defined as:

$$\sum_{z:\,feature\ shared\ by\ x,y} \frac{1}{log(frequency(z))}$$

This measure was customized for link prediction as per the formula below (Liben-Nowell & Kleinberg, 2007), where the number of common neighbors is considered a feature of nodes *x* and *y*.

$$\textbf{adamic/adar}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

This version of the formula allows the Adamic/Adar measure to weigh common neighbors with a smaller degree more heavily.

4.12 Path-Based Features

4.12.1 Shortest Path Distance

The fact that the neighbors of a node can themselves become neighbors suggests that the path distance between two nodes in a social network can influence the formation of a link between them. The shorter the distance, the higher the chance that this could happen. But, also note that, due to the small world phenomenon (Watts et al., 1998), mostly every pair of nodes is separated by a small number of vertices. So, this feature sometimes does not work that well. Hasan et al. (2006) found this feature to have an average rank of 4 among nine features used in their work on link prediction in a biological co-authorship network. A similar finding of poor performance by this feature was also reported by Liben-Nowell & Kleinberg (2007).

4.12.2 Katz Measure

The Katz measure is a variant of the shortest path distance, proposed by Leo Katz (1953), but generally works better for link prediction. It directly sums over all the paths that exist between a pair of vertices *x* and *y*. But, to penalize the contribution of longer paths in the similarity computation, it exponentially dampens the contribution of a path by a factor of β^l, where *l* is the path length. The exact equation to compute the Katz value is as follows:

$$\textbf{Katz}(x,y) = \sum_{l=1}^{\infty} \boldsymbol{\beta^l} \, . \, |paths_{x,y}^{(l)}|,$$

where $|paths^{(l)}_{x,y}|$ is the set of all lengths *l* from *x* to *y*. Katz generally works much better than the shortest path since it is based on an ensemble of all paths between nodes *x* and *y*. The parameter β ($\leqslant$1) can be used to regularize this feature. A small value of β considers only the shorter paths for which this feature very much behaves like features based on the node neighborhood. A problem with this feature is that it is computationally expensive. It can be shown that the Katz score between all pairs of vertices can be computed by finding $(I - \beta A)^{-1} - I$, where *A* is the adjacency matrix, and *I* is an identity matrix of proper size. This task has roughly cubic complexity (i.e., to an exponent of 3), which could be very expensive, even for medium-sized social networks.

4.12.3 Hitting Time

The concept of hitting time comes from the random walks on a graph. For two nodes, *x* and *y*, in a graph, the hitting time, $H_{x,y}$ defines the expected number of steps required for a random walk starting at *x* to reach *y*. Shorter hitting time denotes that the nodes are similar to each other, so they have a higher chance of linking in the future. Since this metric is not symmetric, the commute time, Cx,y = Hx,y +Hy,x, can be used for undirected graphs. The benefit of this metric is that it is easy to compute by performing some random trail walks. On the downside, its value can have high variance; hence, this feature's prediction can be poor (Liben-Nowell & Kleinberg, 2008). For instance, the hitting time between *x* and *y*

can be affected by a vertex *z*, which is far away from *x* and *y*; for instance, if *z* has a high stationary probability, it could be hard for a random walk escapes from the neighborhood of *z*. To protect against this problem, random walks can be fitted with a restart. The random walk is periodically reset by returning to *x* with a fixed probability α in each step. Due to a social network's scale-free nature, some of the vertices may have very high stationary probability (π) in a random walk. To safeguard against this, the hitting time can be normalized by multiplying it with the stationary probability of the respective node, as shown below:

$$\textbf{normalized-hitting-time}(x,y) = H_{x,y} . \pi_y + H_{y,x} . \pi_x$$

4.12.4 Rooted Page Rank

Pagerank measures are mostly used for web-page ranking(Brin et al., 1998). These measures have an inherent relationship with the hitting time. So, PageRank value can also be used as a feature for link prediction(Chung and Zhao, 2010). However, since PageRank is an attribute of a single node, it requires modification to represent the similarity between a pair of nodes *x* and *y*. PageRank's original definition denotes a node's importance under two assumptions: for some fixed probability α, a consumer at a particular web page jumps to a random web page with probability α and follows a linked hyperlink with probability $(1 - \alpha)$. Under this random walk, any web page's importance of any webpage w is the expected sum of the importance of all the web pages *p* that link to *w*. In random walk terminology, one can replace the term importance with the term stationary distribution. For link prediction, the random walk assumption of the original PageRank can be altered as follows: the similarity score between two vertices *x* and *y* can be measured as the stationary probability of *y* in a random walk that returns to *x* with probability $(1 - \beta)$ in each step, moving to a random neighbor with probability β. This measure is asymmetric and can be made symmetric by summing with the counterpart result where *x* and *y* are reversed. Liben-Nowell and Kleinberg (2008) call this "rooted PageRank." The rooted PageRank between all node pairs (represented as *RPR*) can be derived as follows. Let *D* be a diagonal *degree*

matrix with $D[i, i] = \sum_j A[i, j]$. Let $N = D^{-1}A$ be the adjacency matrix with row sums normalized to 1. Then,

$$\boldsymbol{RPR = (1 - \beta)\,(1 - \beta N)^{-1}}$$

4.13 Probabilistic Bayesian Models

Probabilistic Bayesian-based models employ an *a posteriori* probability that denotes the chance of co-occurrence of the node pairs (edges) of interest. There are **three** algorithms. The first one, proposed by Wang, Satuluri, and Parthasarathy (2007) and Kashimaand Abe (2006), uses a Markov random field (MRF)[62] based **probabilistic local model** in which (the output is itself used as a feature in addition to other features like Katz, common neighbors, and vertex attribute similarity and the output is invariably a binary one.) The second algorithm uses a **network evolution based parameterized probabilistic model**. The third model proposed by Clauset et al. (2008) is a **probabilistic hierarchical model,** which considers the hierarchy in an organizational network. The nodes divide into groups, and the groups are further divided into sub-groups and so on.

4.14 Probabilistic Relational Models

A probabilistic relational model (PRM) is a technique that incorporates both node and edge attributes to model the joint probability distribution of a set of entities and the links that associate them. In the methods discussed earlier, the node attributes play a significant role in the link prediction problem. The involvement of node-specific attributes in these

[62]A Markov Random Field (**MRF**) is a graphical model of a joint probability distribution. It consists of an undirected graph $G = (N,E)$ in which the nodes N represent random variables. Let X_S be the set of random variables associated with the set of nodes S. Then, the edges E encode conditional independence relationships via the following rule: given disjoint subsets of nodes A, B, and C, X_A is conditionally independent of X_B given X_C if there is no path from any node in A to any node in B that doesn't pass through a node of C. (http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0809/ORCHARD/#:~:text=A%20Markov%20Random%20Field%20(MRF,the%20set%20of%20nodes%20S.)

approaches makes them non-generic and not useful in all scenarios. The benefit of a PRM is that it considers the object-relational nature of structured data by capturing probabilistic interactions between entities and the links themselves. So, it is better than a flat model that discards such relational information (Hasan et al., 2011). It needs to be underlined that the only entities that other (non-relational) models consider are the node. However, in PRM, heterogeneous entities can be blended. There are two main approaches; one, relational Bayesian-based network considers the relationship (edge) links to be directed (Getoor et al. 2002). The other relational Markov-based network considers the related links undirected (Tasker et al., 2003). Although both are suitable for link prediction tasks, an undirected model seems to be more appropriate for most networks due to its flexibility (Hasan et al. 2011).

4.15 Linear Algebraic Models

A linear algebraic model is a method that generalizes several graph kernels and dimensionality reduction methods to solve the link prediction problem(Kunegis et al., .2009). This method is unique because it is the only method that proposes to learn a function *F*, which works directly on the graph adjacency or the graph Laplacian matrix. Any function *F* that accepts a matrix and returns another matrix is suitable for link prediction,i.e., the entries in the returned matrix should encode the similarity between the corresponding vertex pairs. Many graph kernels can be used for F. A few of the more popular are the **exponential kernel**, which uses the exponential value of the adjacency matrix *A*, i.e., A^n; the **Von Neumann kernel,** and the **Laplacian kernels**.

Chapter 5

Collusion Metric: Design and Applications

5.1 The Choice of a Similarity Function

As the discussions in the preceding chapter show, implementing the link prediction mechanism is numerous. A small but growing subset of this vast corpus of research on link prediction techniques is to find missing links in an existing network. Recent research has focused on feature-based mechanisms or approaches using machine learning (ML) algorithms with feature-based attributes as inputs. Less visible in this field are the applications of probabilistic (Bayesian, relational) or linear algebraic models, possibly owing to the complexity of their use. Feature-based link prediction mechanisms to reconstruct missing links are often the simplest and most effective in their results. Zhou et al. (2009) empirically investigated a simple link prediction framework based on node similarity. They compared nine well-known local similarity measures on six real networks spanning a wide array of network categories, including protein-protein interactions (PPI), co-authorships, electrical grids, political blogs, router level topology of the Internet, and finally, the US Air transportation network. Their results indicate that the simplest measure, namely the common neighbors, has the best overall performance. Similar results were reported by Dong et al. (2011).

This study has employed the Jaccard Index, a normalized version of the common neighbor model, and is an equally simple technique. The importance of the number of mail exchanges between the constituent nodes of a dyad was discussed in-depth previously. It has been factored into the formula of the modified Covertness Index. This strand of logic has also been baked into the considerations of selecting a proper feature-based similarity measure. In the Jaccard Index, the common neighbor formula is divided by the union of sets of common neighbors that both nodes in question may possess.

This work diverges from most other studies. It considers edges between the constituent nodes of a dyad to be the fundamental units of analysis and not the nodes themselves, i.e., the concept of Edge Vertex has been enunciated. In the first part of this study, Edge Vertices were arranged in descending order of their Covertness Index value, and a certain number of them selected based on a heuristic (i.e., a threshold). The challenge is to find pairs of dyads that are linked by some common intention. The problem of finding links signifying common intentions that exist between related Edge Vertex pairs (or dyads) is tied to the structure of the ENRON network itself. Information exchanges between nodes in the ENRON network are usually of two types: through the direct exchange of mails (direction of mail is not relevant to this study) and the copies of these emails marked by different nodes to other nodes. Mail copies are usually mirror images of the emails that a pair of nodes exchange with each other.

In the first part of the study, a Covertness Index was developed to serve as a function of the total emails exchanged between the constituent nodes of a dyad and the copies marked out by this pair from amongst these exchanges. We may reasonably assume that the nodes receiving mail copies' are privy to at least some of the information exchanges that might be happening between the nodes of the dyad, which has sent the copies. If so, these recipient nodes can be reasonably assumed to be partners of a part of the covert proceedings if the dyad is deemed covert. Extending this logic strand, we may look at the recipient nodes receiving mail copies from several dyads to be a common repository of their knowledge.

In the common neighbor model, the number of neighbors that a particular node within the network possessed through direct (structural) links was the statistic used to determine the node's similarity with another node. Since this study has substituted the edge of the tie between the constituent nodes of a dyad, the concept of a neighborhood node changes quite a bit. The ENRON network is an e-mail network, and an edge is the fundamental entity. A neighborhood is now defined for our purposes as the nodes that have received copies from constituent nodes of a dyad whose common neighbor index is being evaluated. The set of neighbors that a dyad has is the set of nodes which have received mail copies from it; in

other words, the Neighborhood Relationship Set that we defined earlier. Thus, for a pair of edges or dyads, the common neighbor will be the intersection of each edge's Neighborhood Relationship Set.

Consider two dyads, the first one comprising nodes i and j, and the second, the pair of nodes p and q. That is, we are considering the associated Edge-Vertices $((E_v)_{ij}$ and $(E_v)_{pq}$. Then for the two edges and their Edge Vertices, the size of their common neighbors is defined as $|\Gamma(Ev)_{(ij)} \cap \Gamma(Ev)_{(pq)}|$. Accordingly, the Jaccard Index, which normalizes the size of the set of common neighbors, is shown below:

$$\textbf{Jaccard-Index } ((E_v)_{ij},(E_v)_{pq}) = \frac{|\Gamma((Ev)ij) \cap \Gamma((Ev)pq)|}{|\Gamma((Ev)ij) \cup \Gamma((Ev)pq)|}$$

We may also define the similarity function between the edges (ij) and (pq) as the Jaccard Index between them described in the equation above.
Thus,

$$(S)_{ij} \leftrightarrow_{pq} = \frac{|\Gamma((Ev)ij) \cap \Gamma((Ev)pq)|}{|\Gamma((Ev)ij) \cup \Gamma((Ev)pq)|} \text{----------(6.1)}$$

For the present case, the edges' neighborhood comprises the nodes to whom the copies of the emails exchanged between the dyad's constituent nodes have been marked. Figure 5.1 shows how the neighborhood is defined for an edge. Observe the edge between nodes a and b in the figure. All the nodes other than a and b themselves have received copies of emails from instances of mails exchanged between a and b comprise the Edge Vertex neighborhood defined on edge between nodes a and b. We may observe from the figure that nodes c, d, e, f, g, and h have all received copies of emails exchanged between a and b. Thus, the neighborhood set of the Edge Vertex between a and b will include these nodes. Notationally speaking:

$$(Ev)_{ab} = \{c,d,e,f,g,h\}$$

$$\Gamma(Ev)_{ab} = |\{c,d,e,f,g,h\}| = 6$$

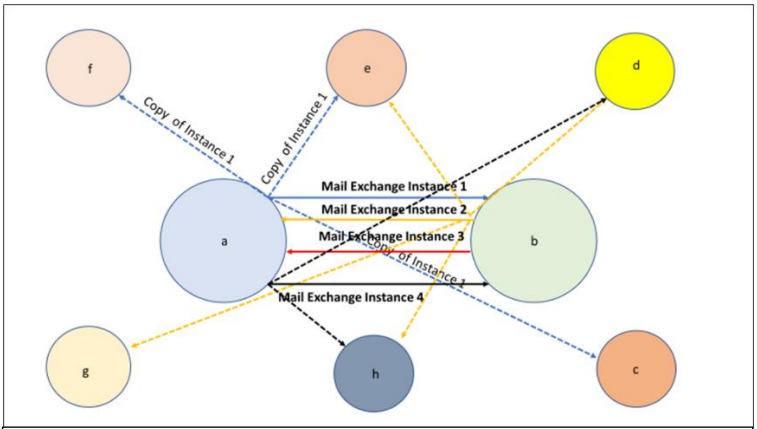


Figure 5.1 Diagram showing how the neighborhood of the edge between nodes *a* and *b* is defined.

5.2 Building Links Using Collusion Index

A pair of edges is considered to build the common neighbors from edges (and their associated Edge-Vertices). There is a set of nodes for each of the edges to which copies of emails have been marked. This set forms the Union Set when the nodes' details to which copies are marked find a mention. As we have seen, the Set of Union comprises the total number of unique nodes that have received copies of mail exchanges from either of the Edge Vertices in question. The cardinality of the Set of Union becomes the denominator during the Jaccard Index of correlation calculation.

To calculate the numerator, we need first to determine the cardinality of the Set of Intersections. The Set of Intersection comprises those nodes which have received copies from both the edges. The fraction represents the Jaccard correlation between the edges and populates the corresponding cell in the correlation matrix.
An illustration is provided below:

Let's define an Edge-Vertex function on the edge formed between the following mentioned employee mail-ids (nodes, basically):

jeff.dasovich@enron.com and richard.shapiro@enron.com.

Suppose that we want to calculate the Jaccard Index between this edge-pair and the edge-pair formed by a second pair of mail-ids:
tana.jones@enron.com and stacy.dickson@enron.com

The calculation is enumerated as follows:

Step#1: Compute the set of nodes which have received copies of emails from jeff.dasovich@enron.com and richard.shapiro@enron.com

The set of recipient nodes is:

{james.steffes@enron.com, janine.migden@enron.com, karen.denne@enron.com, mark.schroeder@enron.com, mona.petrochko@enron.com, mpalmer@enron.com, paul.dawson@enron.com, aleck.dadson@enron.com, david.parquet@enron.com, ginger.dernehl@enron.com, james.steffes@enron.com, linda.robertson@enron.com, paul.kaufman@enron.com, susan.mara@enron.com**}**

Let's name this set as $\boldsymbol{A}$.

Step#2: Calculate the cardinality of set $\boldsymbol{A}$,
It comes to 14 i.e. $|\boldsymbol{A}| = 14$.

Step#3: Determine the set of nodes which have received copies of emails from the second edge, namely:
tana.jones@enron.com and stacy.dickson@enron.com

Set of recipients:

{karen.lambert@enron.com, david.parquet@enron.com, dale.neuner@enron.com, jeffrey.hodge@enron.com**}**

Step#4: Calculate the cardinality of the above set, which we may name set B. The cardinality of set B is 4, i.e., $|B| = 4$.

Step#5: Effect a union of set A and set B, which we may name Set U. The set of the union looks like:

{james.steffes@enron.com, janine.migden@enron.com, karen.denne@enron.com, mark.schroeder@enron.com, mona.petrochko@enron.com, mpalmer@enron.com, paul.dawson@enron.com, aleck.dadson@enron.com, david.parquet@enron.com, ginger.dernehl@enron.com, james.steffes@enron.com, linda.robertson@enron.com, paul.kaufman@enron.com, susan.mara@enron.com, karen.lambert@enron.com, dale.neuner@enron.com, jeffrey.hodge@enron.com}

Step#6: Compute the cardinality of the set of the union or U;
It comes to 17.
$|U| = 17$

Step#7: Effect an intersection of the sets A and B and name it set I.

Thus, I = {david.parquet@enron.com}, since this is the only email node common to both sets A & B.

Step#8: Compute the cardinality of the set of the intersection or set I;
It comes to 1, i.e., $|I| = 1$.

Step#9: Compute the Jaccard Index[63] (JI) of the pair of edges;

$$\text{JI (pair of Edges)} = \frac{\textbf{Cardinality of the set of Intersection}}{\textbf{Cardinality of the set of union}}$$

[63]For all purposes from here on, Jaccard Index and Collusion Index carry the same meaning and formulation. Both terms are used interchangeably.

$= |I| / |U| = 1/17 = 0.0588$

Thus, the similarity values that link a pair of edges (Edge Vertices) is reflected in the JI value that gets computed in the manner indicated above. Continuing for all the pairs of edges in the network, the JIs of all the edge pairs can be computed, and the pairs of Edge Vertices can then be ranked in a descending order based on their JI values. The higher the JI value, the more cohesive or similar the correlation is between that pair of edges. The conjugation of the edges' pairs and their associated Edge-Vertex functions is the key step in building larger covert edge-pairs communities with common intentions.

In the sections on the computation of the Covertness Index, we had selected a threshold value of covertness above which the Edge Vertices (Dyads of nodes) were selected for the next computation step, i.e., the task of link prediction between similar or cohesive pairs of edges again, after we compute the similarity co-efficient $S_{ij \leftrightarrow pq}$ where $(i,j,p,q) \in V_{JI}$, where V_{JI} is the set of all nodes present in the chosen selection and the parameter JI denotes the Jaccard Index value chosen as the threshold.

Obviously,

$V_{JI} \subset V$, where V is the set of all nodes in the network.

Similarly, all the edges that fall within the scope of the selection belong to E_{JI}, where $E_{JI} \subset E$ (the set of all edges in Graph G).

In the previous section, we had selected three threshold values in serial order for selecting the top-ranked covert edges, namely, $T=2500$, $T=5000$ & $T=10000$.

The top-ranked 2500 covert edges have been chosen for the generation of the Edge Pair similarity coefficient (the Jaccard Index or the Collusion Index basically) between every edge pair present in the set of 2500 top-ranked covert edges. The pairs of the edges are again ranked in a descending manner on their similarity coefficient scores (i.e., the Jaccard Index value); the higher the Jaccard Index value, the higher ranked the edge-pair. The

higher values mean that the pair of edge-pairs that occur higher in the Collusion Index ranking have a better correlation. A higher value of similarity implies that the pair of edges have a higher degree of correlation. This can be simply stated as the measure of Common Intention between the edges constituting the edge-pair in question. Stated differently, the ranking of the edge-pairs based on their Collusion-Index values is a ranking of how covertly related the pair of edges are; in fact, the Collusion-Index scores accentuate the initial edge attribute values of the Covertness Index, which were earlier used to rank edges.

Two tables (see table 5.1 and table 5.2) are given below to show how the pairs of edges line up in the Collusion Index rankings. The first table (table 5.3) shows the top 50 ranked edge-pairs. In the second table (table 5.4), which also contains the same top 50 ranked edge-pairs, the status of edge-pairs of interest in these rankings is reflected. The pairs comprising edges, both of which are of interest for covertness classification (i.e., both the edges comprising the pair are EoIs), are highlighted in yellow. It needs to be mentioned here that the Collusion Index Values of some of the edge-pairs which exist between pairs of nodes will be zero if no emails have been exchanged. These pairs are the perfectly covert edges with an unmodified Covertness Ranking of one. Such pairs have been taken out of the selected-set of edge-pairs in consideration since they yield a Collusion Index value of zero, as the example below illustrates:

Let's consider node pairs (a,b) and (c,d) such that $a, b, c, d \in V_{JI}$, where JI is the threshold value selected. Let's further presume that of the emails exchanged between nodes a and b, not a single instance has been marked out as a copy to any outside node. That is, the unmodified Covertness Index values of the edges (a,b) and (c,d) are one. We plug in these values into equation 6.1, defining the similarity coefficient.

$$(S)_{ab} \leftrightarrow_{cd} = \frac{|\Gamma((Ev)ab) \cap \Gamma((Ev)cd)|}{|\Gamma((Ev)ab) \cup \Gamma((Ev)cd)|}$$

$$(S)_{ab} \leftrightarrow_{cd} = \frac{0}{|\Gamma((Ev)ab) \cup \Gamma((Ev)cd)|} = 0;$$

While computing the Collusion Index values of the links between edge-pairs and then selecting a set of edge-pairs ranked based on their JI values, we need to be careful about removing edge-pairs with 0 values that may be perfectly covert dyads with a high value of Covertness Index. In the present experiment, there are not that many EoIs that have Covertness Index values of 1, and the loss of such perfectly covert EoIs is negligible. In instances where there may be a greater proportion of such perfectly covert edges, it will be a good practice to keep the edge-pairs discarded due to null Jaccard Index values in a separate container for further evaluation. One way to avoid deleting such edges from being selected would be to define exceptions within the algorithm and add them back into the zone of consideration after the collusion index values are calculated for the remaining members of the set of edge-pairs.

Another possibility of edge-pairs generating null values arises if the pairs don't have any common nodes to which copies of the emails exchanged have been marked. These edge-pairs can be safely eliminated from contention as they have no meaningful information to offer in terms of collusiveness between the edges in question.

The third category of edge-pairs that can be eliminated is the diagonal entries in the matrix formed when the similarity or collusion index is calculated between the constituent edges. These entries occur due to interaction between the same pair of edges essentially, but which are ordered differently. For instance, the edge-pairs (a,b) and (b, a) are the same and will have both the numerator and denominator common. However, there will be rare instances when both edges may have an identical intersection and the union sets, i.e., when all the mail exchanges between the constituent nodes of both the dyadic-pairs in question have been marked out as copies to the same set of outside nodes. The collusion index value will be 1 in these cases. But such instances will be aberrations rather than the rule and can be safely eliminated from consideration(See the color matrix in Figure 5.2 below).

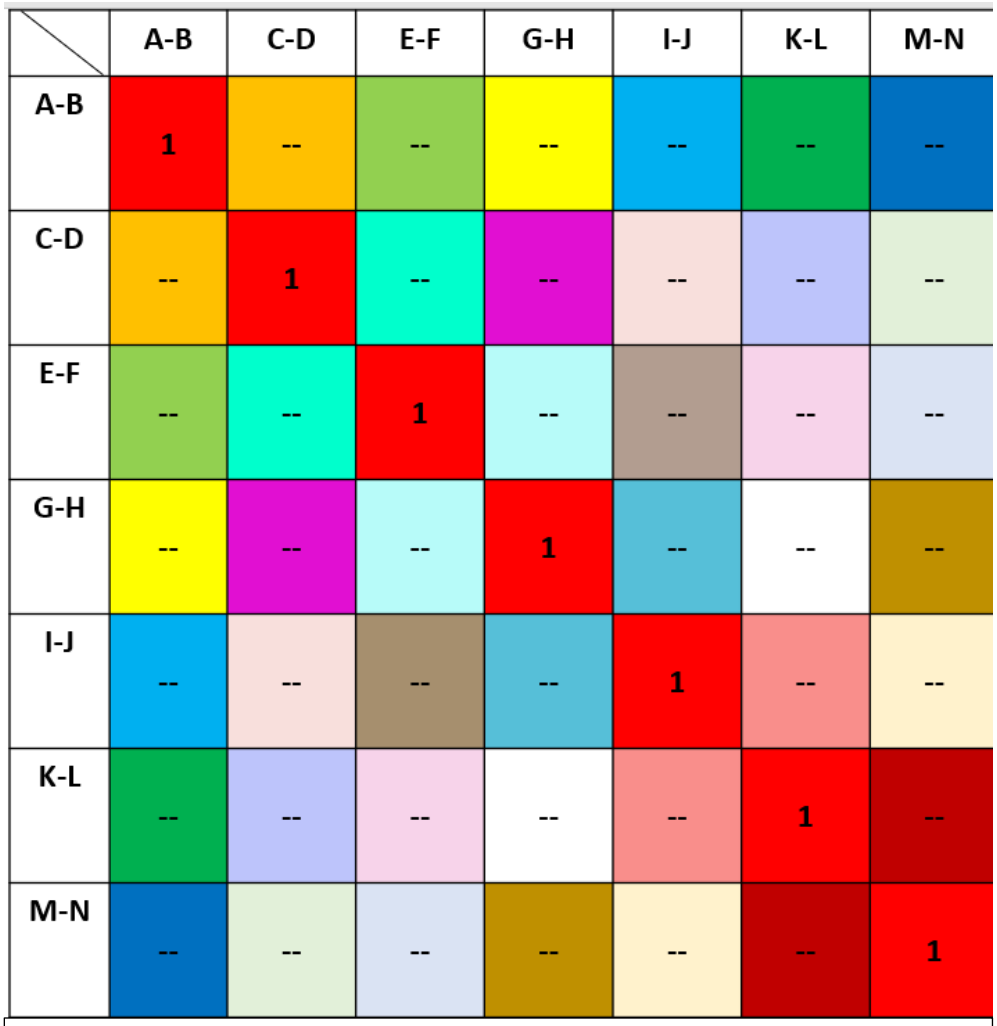


Figure 5.2 A representative Adjacency Matrix showing the entries obtained after applying the similarity metric (Jaccard Index) to pairs of edges (shown along the rows and columns). It may be noticed that the diagonal entries are all1's since the edge pairs are identical and will have all their values in common. The matrix itself is symmetric and diagonal in nature. The cells with the same colors have the same similarityvalues.

2	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	beverly.aden@enron.com	0.994764398
3	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	mpalmer@enron.com	0.994764398
4	jeff.dasovich@enron.com	beverly.aden@enron.com	jeff.dasovich@enron.com	mpalmer@enron.com	0.994764398
5	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.994764398
6	jeff.dasovich@enron.com	beverly.aden@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.994764398
7	jeff.dasovich@enron.com	mpalmer@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.994764398
8	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	john.moore@enron.com	0.993527508
9	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	lisa.alfaro@enron.com	0.993527508
10	kay.mann@enron.com	john.moore@enron.com	kay.mann@enron.com	lisa.alfaro@enron.com	0.993527508
11	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	matthew.berry@enron.com	0.993527508
12	kay.mann@enron.com	john.moore@enron.com	kay.mann@enron.com	matthew.berry@enron.com	0.993527508
13	kay.mann@enron.com	lisa.alfaro@enron.com	kay.mann@enron.com	matthew.berry@enron.com	0.993527508
14	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
15	kay.mann@enron.com	john.moore@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
16	kay.mann@enron.com	lisa.alfaro@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
17	kay.mann@enron.com	matthew.berry@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
18	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	smara@enron.com	0.992167102
19	jeff.dasovich@enron.com	beverly.aden@enron.com	jeff.dasovich@enron.com	smara@enron.com	0.992167102
20	jeff.dasovich@enron.com	mpalmer@enron.com	jeff.dasovich@enron.com	smara@enron.com	0.992167102
21	jeff.dasovich@enron.com	smara@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.992167102
22	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	alvin.thompson@enron.com	0.992125984
23	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	anita.patton@enron.com	0.992125984
24	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	anita.patton@enron.com	0.992125984
25	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	brad.bangle@enron.com	0.992125984
26	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	brad.bangle@enron.com	0.992125984
27	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	brad.bangle@enron.com	0.992125984
28	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
29	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
30	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
31	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
32	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
33	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
34	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
35	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
36	chris.germany@enron.com	cindy.vachuska@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
37	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
38	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
39	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
40	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
41	chris.germany@enron.com	cindy.vachuska@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
42	chris.germany@enron.com	dana.daigle@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
43	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
44	chris.germany@enron.com	jesse.villarreal@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
45	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
46	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
47	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
48	chris.germany@enron.com	cindy.vachuska@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
49	chris.germany@enron.com	dana.daigle@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
50	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	matthew.fleming@enron.com	0.992125984

Table 5.1 Table showing edge pairs whose similarity co-efficient (Jaccard Index values) are the highest. The co-efficient is termed as the Collusion Index and measures how related are a pair of covert edges.

2	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	beverly.aden@enron.com	0.994764398
3	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	mpalmer@enron.com	0.994764398
4	jeff.dasovich@enron.com	beverly.aden@enron.com	jeff.dasovich@enron.com	mpalmer@enron.com	0.994764398
5	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.994764398
6	jeff.dasovich@enron.com	beverly.aden@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.994764398
7	jeff.dasovich@enron.com	mpalmer@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.994764398
8	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	john.moore@enron.com	0.993527508
9	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	lisa.alfaro@enron.com	0.993527508
10	kay.mann@enron.com	john.moore@enron.com	kay.mann@enron.com	lisa.alfaro@enron.com	0.993527508
11	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	matthew.berry@enron.com	0.993527508
12	kay.mann@enron.com	john.moore@enron.com	kay.mann@enron.com	matthew.berry@enron.com	0.993527508
13	kay.mann@enron.com	lisa.alfaro@enron.com	kay.mann@enron.com	matthew.berry@enron.com	0.993527508
14	kay.mann@enron.com	john.ayres@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
15	kay.mann@enron.com	john.moore@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
16	kay.mann@enron.com	lisa.alfaro@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
17	kay.mann@enron.com	matthew.berry@enron.com	kay.mann@enron.com	roger.ondreko@enron.com	0.993527508
18	jeff.dasovich@enron.com	angela.schwarz@enron.com	jeff.dasovich@enron.com	smara@enron.com	0.992167102
19	jeff.dasovich@enron.com	beverly.aden@enron.com	jeff.dasovich@enron.com	smara@enron.com	0.992167102
20	jeff.dasovich@enron.com	mpalmer@enron.com	jeff.dasovich@enron.com	smara@enron.com	0.992167102
21	jeff.dasovich@enron.com	smara@enron.com	jeff.dasovich@enron.com	steven.kean@enron.com	0.992167102
22	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	alvin.thompson@enron.com	0.992125984
23	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	anita.patton@enron.com	0.992125984
24	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	anita.patton@enron.com	0.992125984
25	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	brad.bangle@enron.com	0.992125984
26	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	brad.bangle@enron.com	0.992125984
27	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	brad.bangle@enron.com	0.992125984
28	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
29	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
30	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
31	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	cindy.vachuska@enron.com	0.992125984
32	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
33	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
34	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
35	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
36	chris.germany@enron.com	cindy.vachuska@enron.com	chris.germany@enron.com	dana.daigle@enron.com	0.992125984
37	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
38	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
39	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
40	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
41	chris.germany@enron.com	cindy.vachuska@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
42	chris.germany@enron.com	dana.daigle@enron.com	chris.germany@enron.com	jesse.villarreal@enron.com	0.992125984
43	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
44	chris.germany@enron.com	jesse.villarreal@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
45	chris.germany@enron.com	alfonso.trabulsi@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
46	chris.germany@enron.com	anita.patton@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
47	chris.germany@enron.com	brad.bangle@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
48	chris.germany@enron.com	cindy.vachuska@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
49	chris.germany@enron.com	dana.daigle@enron.com	chris.germany@enron.com	mark.friedman@enron.com	0.992125984
50	chris.germany@enron.com	alvin.thompson@enron.com	chris.germany@enron.com	matthew.fleming@enron.com	0.992125984

Table 5.2 Same table showing edge pairs whose similarity co-efficient (Jaccard Index values) are the highest. One of the edge-pairs of interest occurs at rank 4.

The results obtained after applying the Collusion Index (Jaccard Index) to calculate links between all the top-ranked covert edges have been represented graphically in the figures below (Figures 5.3, 5.4, and 5.5). The first figure (Figure 5.3) is a chart plotting the prevalence of the pairs of edges of interest (EoIs) against the overall ranking of all possible edge-pairs (i.e., which have been considered as per discussions above) based on the value of the Collusion Index of the links binding each pair. The second of the charts (Figure 5.4) is the same but shows the rankings' logarithmic value. The last of the charts (Figure 5.5) shows the comparative performance in detecting the pairs of the edges of interest (EoIs) between the model where Collusion Index values are calculated and applied to the links between the edge-pairs and a Uniform Distribution model where the pairs of EoIs are assumed to populate the entire set of edge-pairs in a uniform manner, which is similar to the comparisons that were done between the Covertness Index model and the Uniform Distribution model in the earlier section.

After these charts, a set of three more plots (Figures 5.6, 5.7, and 5.8) are presented, showing how the use of the Collusion Index as a similarity measure between pairs of edges improves the detection performance measurably. As per the previous practice, the performance improvement has been presented in Precision, Recall, and F1 Score metrics. The calculations are shown in the paras following the charts. The horizontal axes in all charts denote the edge-pairs' rankings based on the Jaccard Index values of the links between them. The vertical axes denote the count of the covert or the desired edges in the distribution.

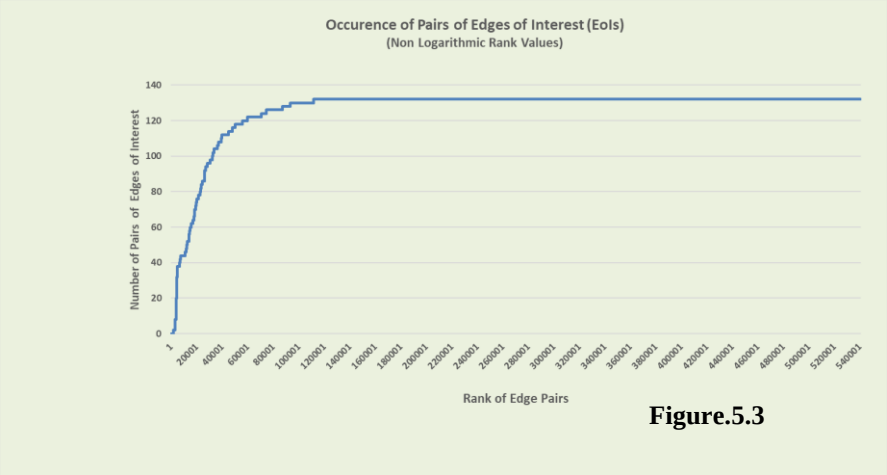


Figure.5.3

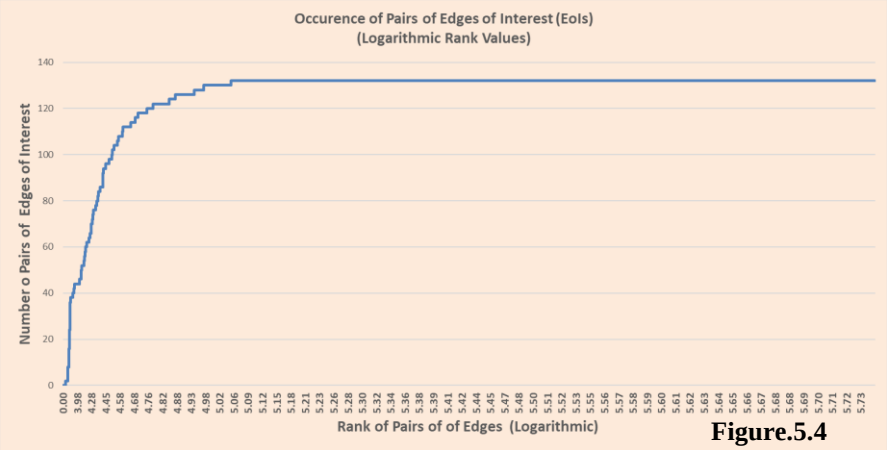


Figure.5.4

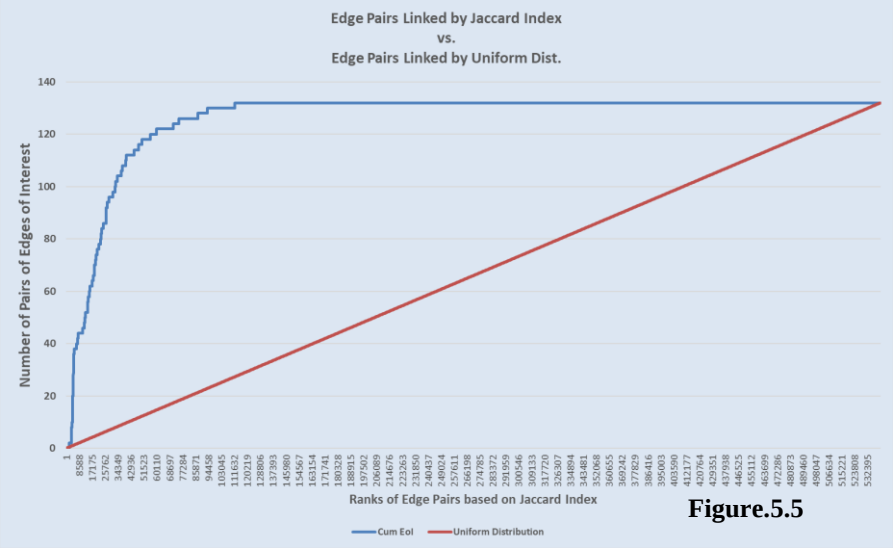


Figure.5.5

Figure.5.6

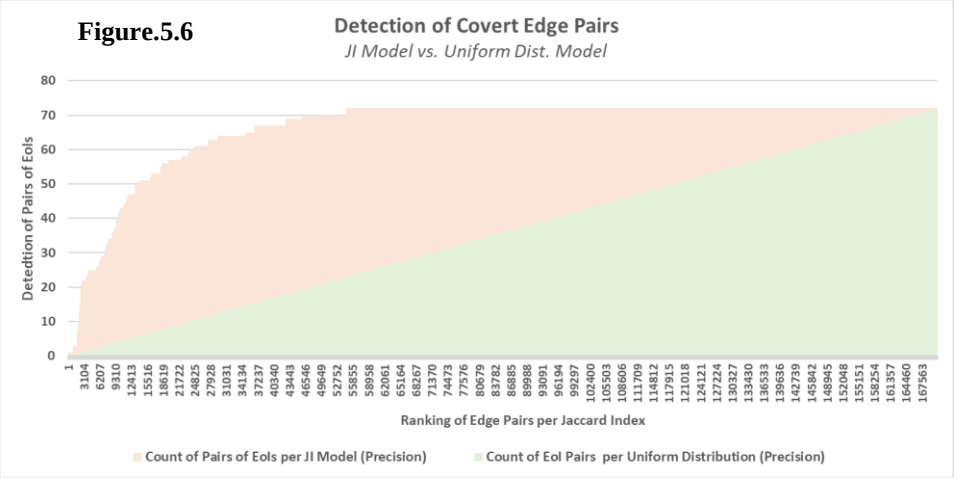


Figure.5.7

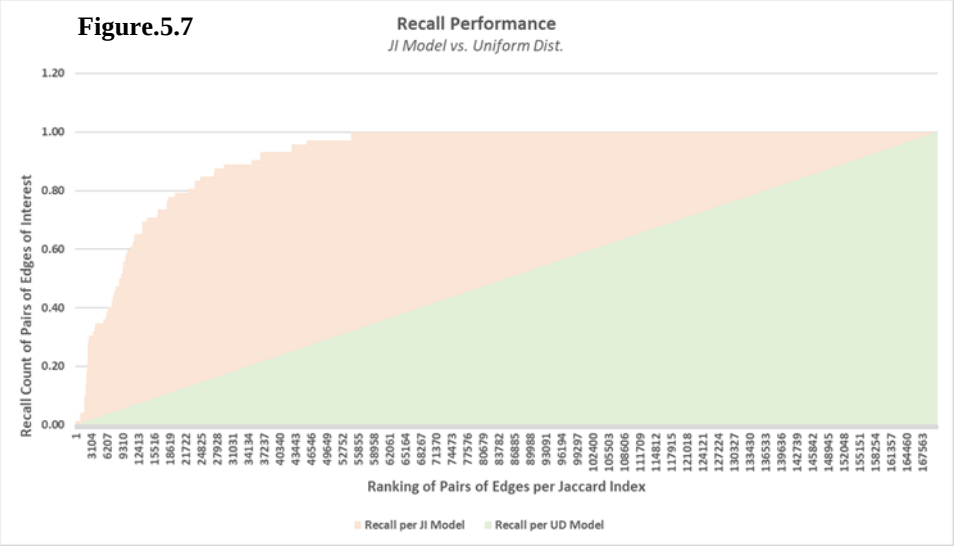
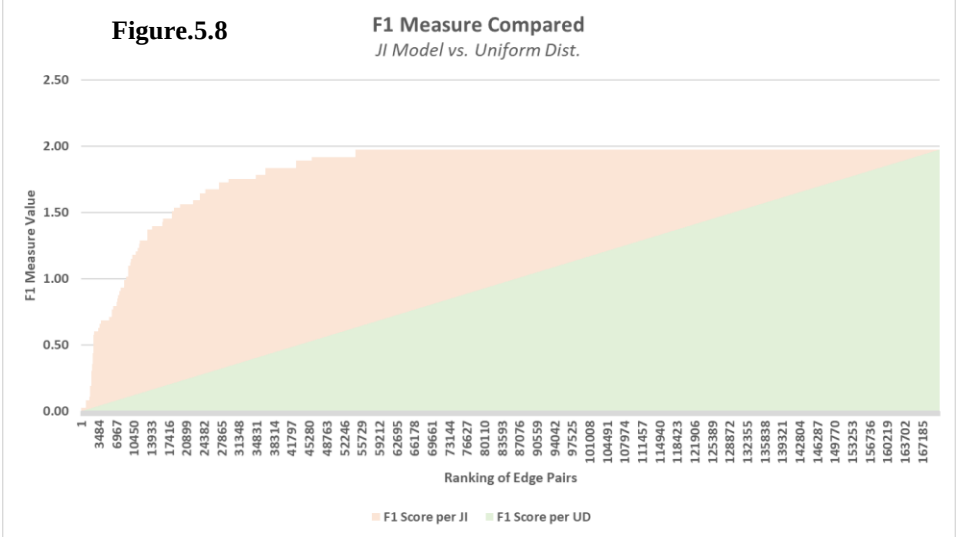


Figure.5.8



5.3 Selecting a Threshold of Collusion

We may observe that there are as many as 170,000 edge-pairs or more in the theatre of consideration. Of these, only 61 are pairs that have been formed out of the edges of interest (EoIs), which we may recall are only 43 in number. If we consider keeping the entire set of 170,000 plus entities within the investigation purview, the task becomes a surveillance nightmare. There needs to be a *selected-set* of some convenient size that will keep within its fold, a substantial number of EoIs pairs while not being so large as to be unmanageable for investigators.

This discussion again raises the question of the heuristic or threshold value that needs to be chosen to evaluate the Collusion Index (Jaccard Index) metric's success. In this study, I have selected the top-ranked 2500 edge-pairs[64] with ranks based on the Jaccard Index values of the links between the edge-pairs. As was discussed earlier, this value is strictly based on the surveillance circumstances, including parameters of human resources availability, computational resources, information availability, and, more crucially, the time available to predict the network. More often than not, the agencies tasked with such responsibilities are deficient in all these resources. Often, there is a race against time to produce actionable results. There is an element of error, no doubt, given the constraints of incomplete information. The savings made in terms of resources and the fact that this approach doesn't impinge on communication privacy more than makeup for the losses. Also, the outputs vastly optimize the scope of surveillance. For instance, there is no need to study all 170,000 plus edge-pairs. The results from just the 2500 edge-pairs (the top-ranked 0.6% of the overall set of edge-pairs) taken up for study seem to justify the efforts.

[64]This threshold is deliberately chosen to prepare a common ground to compare performances between the first stage of the detection performance i.e. after the application of the Covertness Index on the ties between the actors and the performance after applying Jaccard Index to link pairs of edges between actors.

5.4 The Impact of Applying the Collusion Index

After computing the Collusion Index values of the links that form between the pairs of edges and then ranking them in descending order, the top-ranked 2500 pairs of edges are observed to contain 14 pairs of edges of interest (EoIs), i.e., pairs of dyadic node pairs, in which all 4 participant nodes are nodes of interest (NoIs), 74 pairs of edges, in which 3 out of 4 constituent nodes are NoIs. Hence, out of the top 2500 most correlated pairs of edges, 88 pairs are either completely constituted by nodes of interest (i.e., 4 out of 4) or nearly so (3 out of 4 constituent nodes). The table shows how many interest nodes are present in each edge-pair or cluster obtained after calculating the Collusive Index. The proportion of accuracy in constructing the desired communities of edges (edge-pairs of interest) comes to 3.52% (88/2500), which is much better than the initial accuracy figures arrived at after the CI rankings were worked out (23/2500 or 0.92 %).

Number of Nodes of interest (NoIs)	Count of Clusters having the NoI's
4	14
3	74
2	562
1	562
0	1288
Total	2500

Table 5.3 Table showing the edge-pairs and the number of NoIs in each edge-pair. It needs to be mentioned that each edge-pair can have up-to 4 distinct nodes.

It's also noteworthy that only 2500 pairs of edges have been considered out of a possible 170,000 plus edges. The threshold value of 2500 is a heuristic and can be fine-tuned depending on the degree of accuracy desired in the results.

But, what is significant is that the proportion of edge-pairs that are of interest to the study in the *selected set*[65]is quite high. Recall that the initial exercise was to narrow down the scope of surveillance from the entire network to a manageable subnet without intruding into the emails' content. At this stage of the operations, we can see that nearly one in every

[65]The Selected Set is defined as the set of edge-pairs which are having the highest values of Collusion Index which are ranked in a descending order.

twenty of the communities of actors (nodes), i.e., the edge-pairs available after applying the similarity coefficient (Collusion Index), is of interest. This is to say that if surveillance happens on all 2500 selected edge-pairs, the investigation will succeed with a likelihood of 3.52% in detecting covert edges.

The second finding of significance is the steep reduction in the number of nodes that are not of interest to our study. An analysis of the pairs of edges populating the top 2500 ranks by the Collusion Index value reveals 12 nodes of interest (NoI) and 221 nodes that are not of interest (nNoI) in our study. The ratio is approximately 1: 20 (12:221), i.e., about 5.4 % of the nodes that are filtered out turn out to be nodes of interest for the study's purposes. Compared to the previous ratio of NoI to nNoI arrived at after computing and ranking the 2500 top-ranked edges based on their Covertness Index, this ratio is far more favorable (To recall the exact figures, the NoIs numbered 16 and the nNoIs numbered 1780, which yields a ratio of 1:1100 or 0.89 %). Thus, the exercise of correlating pairs of edges has not only improved the chances of picking up the correct subnetwork for more granular analysis (3.52% from 0.92%), the level of *noise*[66] has also decreased significantly since. We now have 221 nodes that are of no interest (nNoIs) compared to 1780 nNoIs, which were present in the set of 2500 top-ranked covert edges. If the experiment had stopped at the stage of applying the Covertness Index, the resources allocated to surveillance would have been some multiple of 1780 rather than 221, which is eight times less. So far, the Collusion Index as a clustering mechanism has successfully enhanced the probability of selecting a subnet where the constituent nodes have a greater degree of covert affiliation. It has eliminated nodes that are not spoken, related in terms of their covert affiliations.

[66]Here, noise refers to the information that is not necessary for scrutiny and will likely consume valuable resources which otherwise would be allotted to the assets of interest (covert nodes or interest).

5.5 Improvements due to the Collusion Index

To buttress the argument further, we apply the same performance metrics for measuring the Covertness Index model's performance, namely, the precision, recall, and F1 measures. These metrics are applied for the limited set of 2500 top-ranked edge-pairs (going by their Collusion Index values), and the comparisons are made against the Uniform Distribution model.

In a previous section, I discussed using various metrics to measure improvement in detecting covert edges, i.e., EoIs. The first of the metrics discussed was Precision, which indicates how correctly the model predicts the true positives. That is, of the covert edge-pairs predicted to be related, how many are part of a conspiracy with common intentions. The charts below show 88 edge-pairs whose linkages are predicted correctly (we include edge-pairs with three out of four constituent nodes in the NoI list). The remainder out of 2500 is incorrectly predicted as being related. That is, the true positives are 88, whereas the false positives are 2500 – 88 = 2412.

Precision is calculated as follows :
Precision = True Positives / (True Positives + False Positives)
In other words, the Precision of our model when we reach the count of 2500 in terms of ranking is:

Precision = 88/ (88 + 2412) = 88/ 2500 = 0.0352 or about 3.52%.

This figure doesn’t appear encouraging until we compare it with the uniform distribution model's figure. The edge-pairs that are related through commonness are assumed to be distributed throughout the set of all the edges uniformly. In this model, the number of related pairs of EoIs which are detected at the count of 2500 is less than 1 (0.56 to be exact). This model's Precision is barely 0.000224, or 0.0224%, which is about 5% of our model's performance based on the Collusion Index.

Figures 5.9, 5.10&5.11 are graphs that show the betterment in performance wrought by applying the Collusion Index to bring out linkages between different edge-pairs obtained from the Selected Set. To recall matters briefly, the Selected Set is a set of edges with the Covertness Index's highest values. A certain number of these edges is selected based on a heuristic figure. In this study, the heuristic threshold of Covertness Index value was so chosen that 2500 highly covert edges were selected, i.e., the Selected Set comprised these 2500 edges. The Collusion Index was applied to find how strongly linked each of these selected covert edges was. The higher the value of the collusion metric of the link between two edges, the more common they are in terms of covertness. The results at the end of this part of the experiment give us pairs of highly linked edges and, by inference, tightly linked set of nodes (since the edge is associated with two nodes), which we may look at as small covert communities or conspiracy subnetworks, which was the original intent of their research.

The graph in Figure 5.9 shows the improvement in the detection of the covert entities that the Collusion Index model brings about compared to a Uniform Distribution model, which allows the covert edges to be present throughout the set of all edges in a uniform manner. In a sense, the Uniform Distribution model serves as the null-set in this study. Figure 5.10compares the Recall metric readings between the Collusion Index model and the Uniform Distribution model. Even here, the performance after applying the Collusion Index is marked. The graph in Figure 5.11 pertains to the F1 metric, which shows the Collusion Index model's clear superiority.

Let’s denote the precision arising out of the JI Model as $\mathbf{P_J}$ and the precision arising out of a Uniform Distribution as $\mathbf{P_U}$**.**

Thus,

$$P_J = \frac{(Number\ of\ Pairs\ of\ Edges\ of\ Interest\ Detected)}{(Number\ of\ Edge\ Pairs\ of\ Interest\ Detected + Number\ of\ Incorrect\ Edge\ Pairs\ Detected)}$$

$$P_J = \frac{88}{88+2412} = \frac{88}{2500} = 0.0352 \text{ or, } 3.52\ \%.$$

Likewise,

$$P_U = \frac{0.56}{2500} = 0.000224 \text{ or, } 0.0224\ \%.$$

We may now recall the original ratio of covert edges we had while starting. The result was expressed as Equation (1.1) i.e.

$P = |E_C| / |E| = 43/ 55{,}300 = 0.000778$ (or 0.08 %)

We may now compare this to the figure of covert edge-pairs within the overall set of edge-pairs after applying the Collusion Index similarity measure to the edge-pairs, which is –

$P_J = 0.0352 >>> 0.000778$

The results obtained after the experiments are nearly 45 times better than what we started with.

Figure 5.9

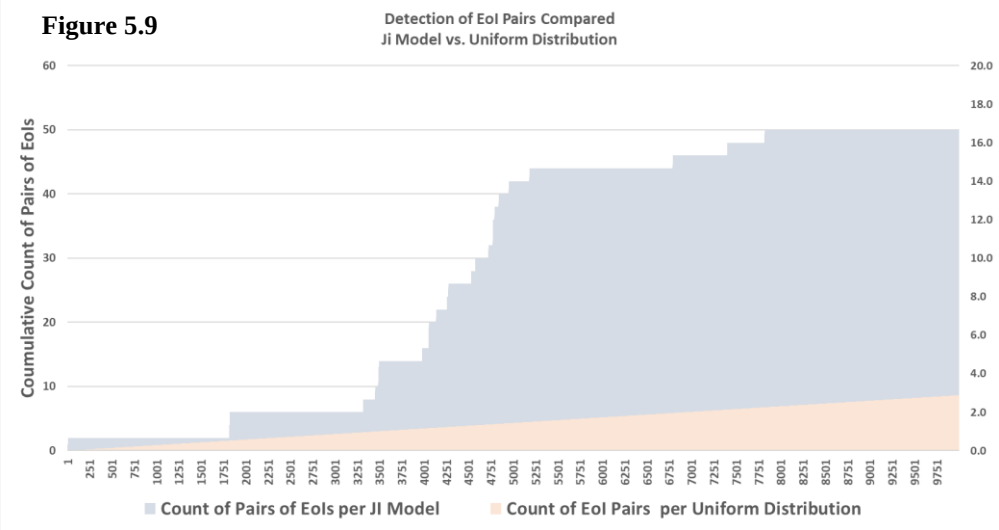


Figure 5.10

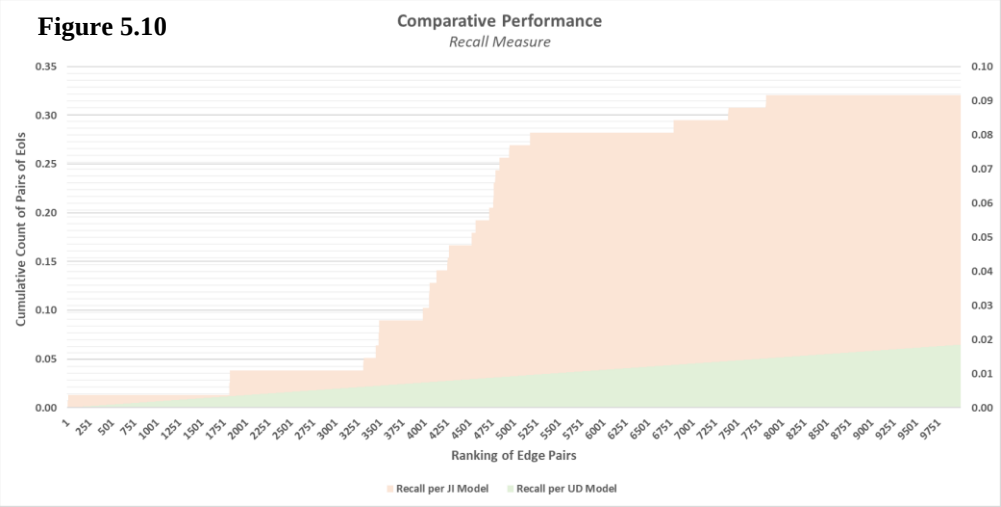
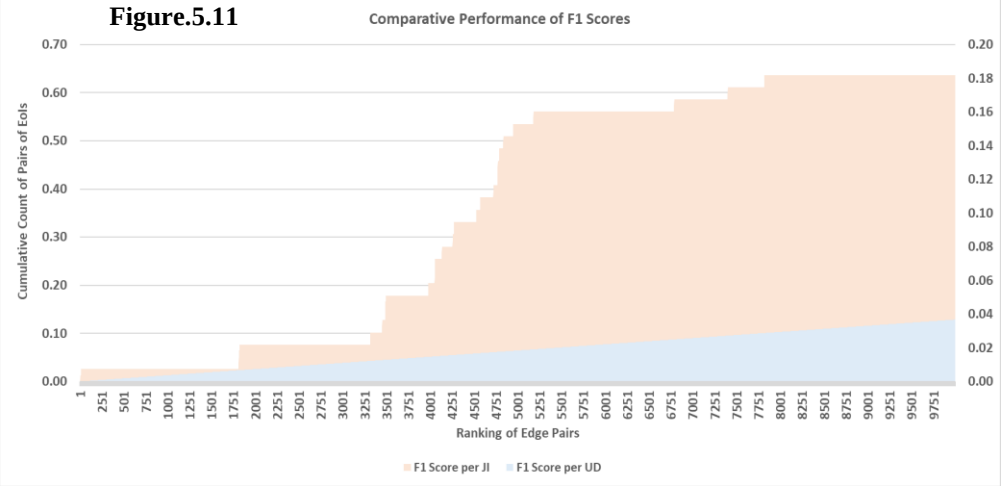


Figure.5.11



Chapter 6

Summary and Analysis of Results

6.1 Summary of the Results

6.1.1 Background

The dissertation has focused on the email corpus of the ENRON Company, which went bankrupt following a major financial scam in 2002. The ENRON dataset is a corpus of emails collected from the inboxes of 151 employees covering a limited period relevant to the investigation of the insider trading scam and other concomitant illegal business practices that led to it. A total of 517,431 mail exchanges are available from these inboxes, and information about 6568 distinct employee email ids is available in the mail corpus. These mail-ids are not unique, as many of the employees had more than one mail-id. Preliminary data cleaning was undertaken to combine all the email-ids of a single employee into a single mail-id entity. This mail-id takes the place of a node or vertex in the ENRON company's social network universe.

The dissertation has treated the ties or edges between nodes or mail-ids as comprising all the mail exchanges that have taken place between the mail-ids. Thus, if a pair of mail-ids has exchanged a thousand mails, it still translates to a single edge, as does the single mail exchange between a pair of mail-ids. Following this methodology, the figure of 517,431 mail exchanges has been condensed to approximately 55,300 unique mail pairs, i.e., the number of mail ids that have exchanged at least one mail. The research has focused on the roles played by 19 ENRON employees who were either indicted or otherwise privy to the proceedings as witnesses or recipients of the information. Forty-three unique mail pairs exist between different mail-id pairs from amongst these employees of interest. The solution proposed in this dissertation has pared down the existing edges in the mail corpus of

ENRON (55,300 approximately) to a small enough selection of edges. The problem was notationally defined as per the statement below:

Let's define the ENRON mail corpus as a social network graph G, such that $G = (V, E)$,

where V is the set of all nodes in the graph network.

And E is the set of all edges or mail-pairs in the graph network.

Hence, any node v_i belonging to the network belongs to set V.

In other words, $V = \{v_i : 1 \leq i \leq |V|\}$,

where the number of nodes in the network graph is represented as the cardinality of the set of nodes V, i.e., $|V|$

and $|V| = 6568$.

The number of edges in the network graph is represented as the cardinality of the set of edges E, i.e., $|E|$

& $|E| = 55{,}300$.

$E = \{e_{ij} : 1 \leq i \leq |V|, 1 \leq j \leq |V| \text{ and } i \neq j\}$

Let's define the set of the ENRON employees who were part of the scam as a graph G_C, such that $G_C = (V_C, E_C)$,

where V_C is the set of all nodes of interest (NoIs) in the graph network, and E_C is the set of all edges of interest (EoIs) in the graph network.

$G_C \subset G$ and $V_C \subset V$ & $E_C \subset E$;

& $|V_C| = 19$ and $|E_C| = 43$;

The probability of finding an edge of interest (EoI) e_{ij} (i and j are nodes of interest (NoIs) in the graph network) in the set of edges of graph G thus becomes[67]:

$$P = |E_C| / |E| = 43/ 55{,}300 = 0.000778$$

The task undertaken by the study was to increase this value of the probability of detection, i.e., boost the chances of detecting an edge of interest (EoI) in the set of edges of the graph network comprising the mail corpus of ENRON.
Thus we may reframe the problem statement in Section 1.6 Chapter 1 by applying a probabilistic formulation to the solution above.

Problem Statement (Probabilistic Perspective)

What is the probability of detecting at least one covert edge from amongst the overall set of edges of the ENRON e-mail network in 20 tries?

Let's define an integer $\boldsymbol{k}$, s.t., $\boldsymbol{k}$ = Number of tries; Here, $\boldsymbol{k} = 20$.
There are 43 covert edges or Edges of Interest (EoIs).
Let's define the number of EoIs as $\boldsymbol{m}$**;** Here, $\boldsymbol{m} = 43$.
The number of edges overall is 55,288 ~ 55,300
Let the total number of edges be defined as $\boldsymbol{e}$; Here, $\boldsymbol{e}$=55,288.

We need to calculate the probability of not getting any covert edges in 20 tries.
Let's Denote the probability of detecting a covert edge as $\boldsymbol{P_c}$and not detecting a covert edge as $\boldsymbol{P_{nc}}$.

The probability of not detecting a covert edge in the first try will be
(55,300 – 43)/ 55,300.
The probability of not detecting a covert edge in the second try will be
(55,299 – 43)/ 55,299.

[67]*Ibid* p.26.

In this manner, the probability of not detecting a covert edge on the 20th try will be (55,280-43)/ 55,280.

Notationally,

$$\boldsymbol{P_{nc}} = \prod_{i=0}^{k-1} \frac{((e-i)-m)}{(e-i)}$$

$$\boldsymbol{P_c} = (1 - \boldsymbol{P_{nc}}) = \left(1 - \prod_{i=0}^{k-1} \frac{((e-i)-m)}{(e-i)}\right)$$

Hence, the probability of not getting a covert edge detected in 20 tries (k = 20) comes to:

$$\boldsymbol{P_{nc}} = \frac{(55{,}300-43)}{55{,}300}\ x\ \frac{(55{,}299-43)}{55{,}299}\ x.x\ \frac{(55{,}281-43)}{55{,}281}$$

$\boldsymbol{P_{nc}}$ = 0.984560175; $\boldsymbol{P_c}$ = (1 - $\boldsymbol{P_{nc}}$) = 0.015439825

Step#1: After Applying the Covertness Index

After applying the Covertness Index to all the edges, the number of covert edges of interest (EoIs) comes to 23 in a selected-set of 2500 top-ranked covert edges.

Thus, after this part of the experiment, $\boldsymbol{m}$ = 23; $\boldsymbol{e}$=2500; $\boldsymbol{k}$ = 20.

Plugging these values into the equation above, we get:

$$\boldsymbol{P_{nc}} = \frac{(2500-23)}{2500}\ x\ \frac{(2499-23)}{2499}\ x\ \ldots.x\ \frac{(2481-23)}{2481}$$

$\boldsymbol{P_{nc}}$ = 0.830638142; $\boldsymbol{P_c}$ = (1 - $\boldsymbol{P_{nc}}$) = 0.169361858

Step#2: After Applying the Collusion Index

After applying the Collusion Index to all the edge-pairs, the number of covert edge-pairs of interest (EoIs) comes to 88 in a selected-set of 2500 top-ranked covert edge-pairs.

Thus, after this part of the experiment, $\boldsymbol{m}$ = 88; $\boldsymbol{e}$=2500; $\boldsymbol{k}$ = 20.

Plugging these values into the equation above, we get –

$$\boldsymbol{P_{nc}} = \frac{(2500-88)}{2500}\ x\ \frac{(2499-88)}{2499}\ x\ \ldots.x\ \frac{(2481-88)}{2481}$$

$$\boldsymbol{P_{nc}} = 0.48701;\ \boldsymbol{P_c} = (\mathbf{1} - \boldsymbol{P_{nc}}) = 0.51299$$

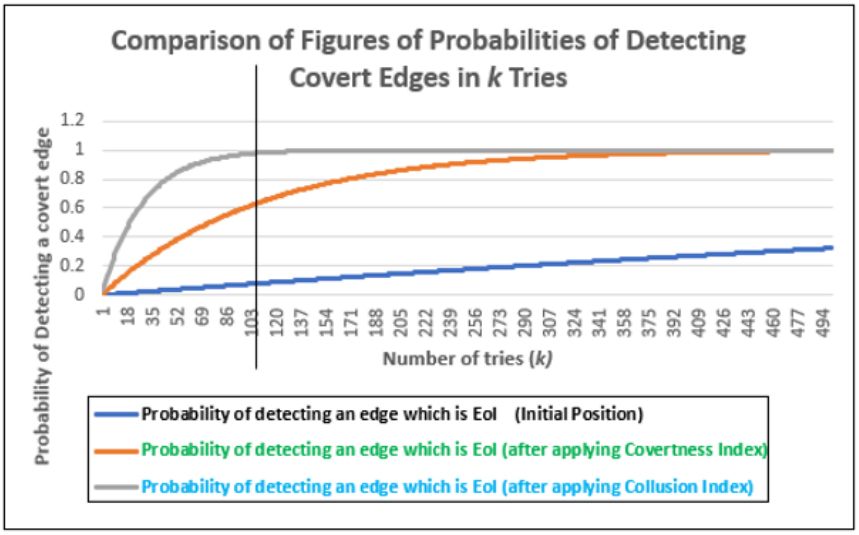


Figure 6.1 Plot showing the comparison between the probabilistic outcomes of detecting at least one covert edge in a fixed number of tries. The enhancements brought about through the application of the Covertness Index metric (orange line) and the Collusion Index (grey line).

The plot in Figure 6.1 shows the comparison of probabilities of detecting at least one covert edge or edge of interest or EoI in a fixed number of tries. The line colored blue-gray is the plot of probabilities of detecting at least one covert edge (EoI) in *k* tries.

In the graph shown above, the value of k varies between 1 to 500($1 \leq k \leq 500$).

As can be seen, the best chance of detecting at least one covert edge occurs at the value of k = 100 (approx.) after the application of the Collusion Index to find linkages between pairs of covert edges (see the vertical line cutting across the line plots at k =100 on the horizontal axis).

Thus, the efficacy of the metrics is established.

6.1.2 Covertness Index Metric

As a first step, a Covertness Index was defined on the edges connecting the nodes constituting the network's dyads. The Index was designed to measure the confinement of information between the nodes in the pair. Confinement of information is one of the principal methods to detect if the nodes communicating between themselves hide any information. Since the network in question is an email-based network, the methodology adopted in this study to measure covertness or confinement of information between a pair of nodes was to ascertain how much of the information exchanged was going out of the pair of nodes to other nodes. In other words, since the information exchange is happening via email, how many of the mails were being copied out to the others. The proportion of emails copied out represents the 'overt' portion of the exchange based on the trivial assumption that two entities will not share confidential information with third parties. Going by this logic, the retained information exchange (without being copied out) comprises the 'covert' portion. This interpretation was mapped out mathematically as follows:

Let e_{ij} (V_i and V_j being two nodes constituting a dyad) represent an edge between nodes *Vi* and *Vj*.

The cardinality of e_{ij} or, $|e_{ij}|$ = Number of mails exchanged between nodes Vi and Vj.

Let $(e_{ij})_c$ represent the mails copied out from the set of emails exchanged between nodes Vi and Vj.

Thus, the cardinality of $(e_{ij})_c$ or, $|(e_{ij})_c|$ = Number of mails copied out by the dyad with the edge $(e_{ij})_c$, i.e., between nodes Vi and Vj.

The Covertness Index of edge e_{ij} is defined as per the formula below:

$$(\boldsymbol{CI})_{ij} = \mathbf{1} - \frac{|(\boldsymbol{e_{ij}})_{\boldsymbol{c}}|}{|\boldsymbol{e_{ij}}|}$$

This formula is modified to include the number of emails exchanged between two nodes in a dyad. The volume of exchange is akin to a centrality measure for an edge, and it adds context to the calculation of the Covertness Index. Examples have been provided in the concerned section to show how the exclusion of the volume of emails exchanged within the pair skews the covertness results and allows dyad having fewer emails exchanged to dominate the rankings.

The modified Covertness Index is defined as follows:

$$\boldsymbol{Modified}\ (\boldsymbol{CI})_{ij} = |\boldsymbol{e_{ij}}| * ((\boldsymbol{CI})_{ij}) = |\boldsymbol{e_{ij}}| * (\mathbf{1} - \frac{|(\boldsymbol{e_{ij}})_{\boldsymbol{c}}|}{|\boldsymbol{e_{ij}}|})$$

The Covertness Index(only the modified Covertness Index formula is used in this study) is calculated for each of the dyads, and the edges are ranked in descending order. A selection of edges is made based on a threshold value of covertness chosen heuristically. In this research, three such thresholds are chosen, yielding sets of 2500 numbers of edges, 5000 numbers, and 10,000 edges, respectively. The following formula represents the smaller sized sets based on thresholds of Covertness Index values:

The ENRON mail corpus is defined as a graph G,

such that $G = (V, E)$,

where V is the set of all nodes in the graph network.

And E is the set of all edges or mail-pairs in the graph network.

Then, the Selected Set of edges based on a threshold value of Covertness Index is defined as graph G_T,

such that $G_T = (V_T, E_T)$;

(T is the threshold value of the Covertness Index).

Thus, $E_T = \{e_{ij} : (\boldsymbol{CI})_{\boldsymbol{ij}} \leq T \text{ and } i \neq j\}$

And, $V_T = \{ v_i, v_j : (\boldsymbol{CI})_{ij} \leq \mathsf{T} \text{ and } i \neq j\}$;

Obviously, $E_T \subset E$ & $V_T \subset V$

And $|E_T| \ll |E|$ and $|V_T| \ll |V|$,

The selected sets were examined for the presence of edges of interest (EoIs). Three evaluation metrics were calculated: Precision (number of true positives, i.e., the actual EoIs to the number of positives detected by applying the index), Recall (number of true positives as a proportion of the actual number of EoIs present in the dataset, i.e., 43) and finally, the F1 measure, which is the harmonic mean of the Precision and Recalls figures.

6.1.3 Effectiveness Examined

To ascertain if the application of the Covertness Index has been effective in detecting the EoIs in the selected set, there is a need for a reference model with which the results can be compared. The model chosen was the Uniform Distribution model, where it is assumed that the EoIs are distributed uniformly. The comparison establishes clearly that the results are far better for all three threshold values chosen in terms of all the three metrics.

Case#	2500 edges			5000 edges			10,000 edges		
Model	Precision	Recall	F1 Measure	Precision	Recall	F1 Measure	Precision	Recall	F1 Measure
Covertness Index Model	0.0092	0.53	0.018	0.0054	0.63	0.011	0.0035	0.81	0.007
Uniform Distribution Model	0.0008	0.04	0.0016	0.0008	0.09	0.0016	0.0008	0.18	0.0016
Difference (in percent)	1150%	1325%	1130%	675%	700%	690%	440%	450%	440%

Table 6.1 Table showing the comparative results between the model based on the application of the Covertness Index to the constituent edges of all dyadic pairs in the ENRON mail based network and a uniform distribution model wherein the probability of finding an edge of interest or EoI is the same across the distribution of edges. The results clearly show how the application of a Covertness Index to the edges improves the chances of detecting the EoIs substantially. The comparison is undertaken with three metrics, namely, Precision, Recall and F1 (shown on the row side) and also across three threshold values chosen heuristically, i.e. 2500, 5000 and 10,000 edges. The results are the best for 2500 edges.

A summary of the comparative results is given in Table 6.1. The table's values bear ample evidence that the Covertness Index application measurably enhances the probability that edges of interest (EoIs) will be detected. Furthermore, a comparison between the thresholds selected shows that the smallest selection set (2500 top-ranked edges) works efficiently across all metrics. Hence, the index's application improves the detection of the edges of interest and acts to minimize the size of the selected set (i.e., smaller thresholds seem to offer better results).

A second takeaway from the experiment is the vast reduction in the set of edges that an investigator needs to examine. When the experiment started, 55,300 edges were examined, and 6568 nodes needed to be accounted for. If these many mails were to be examined for concluding the covert nature of the dyads or the communities of such pairs, it would be extremely resourced intensive. By reducing the number of edges to be examined to 2500, the complexity decreases by a factor of 20 (This figure is arrived at by dividing the number of edges we started with by the number we ended up with, i.e., $55{,}300 \div 2500 \sim 20$.)

6.1.4 Collusion Index Design & Efficacy

This part of the experiment successfully reduces the set size needed to be examined for the presence of edges of interest (EoIs) and boosts the chances of detection. However, the mere identification of covert edges is not the result desired in this study. The stated aim was to identify covert communities sharing common aims and striving to produce the same output. The differentiation between the identification of covert edges and their grouping into common-aim based communities is crucial. There is always a possibility that the covert edges detected will belong to different sub-groups having varied aims and separate outputs. In such an eventuality, the results of surveillance will be sub-optimal. The second part of the experiment was undertaken to address this aspect, which seeks to agglomerate covert pairs into bigger communities of pairs of edges. The linkages between the pairs of edges selected based on the first part of the experiment are based on a similarity measure that seeks to bring out the commonalities between pairs of edges, particularly covert (EoIs). The higher the value of the similarity index, the better the chances that the pair's edges are

collaborating (or colluding) in a common covert enterprise. It may be pointed out that the linkages may be purely benign and may not have anything to do with any clandestine outcomes. But we need to realize that *all the edges* selected as inputs for the second part of the experiment are *deemed to be covert* and, therefore, worthy of further investigation and scrutiny. The study does not purport to know which of the edges are covert within the selected set and not of interest. Nevertheless, even a set of limited sizes still causes problems of computational complexity.

For example, even if the smallest set of edges is selected from the first experiment (i.e., 2500 edges), the potential number of edge-pairs which will need to be examined comes to 2500x2500 = 6,250,000, i.e., more than 6 million! Thus, a similar exercise of pruning edge-pairs by selecting a threshold value of the similarity index needs to be carried out. Small sets of edges are based on heuristic thresholds for the Covertness Index values selected for scrutiny. There are several candidates available for linking the edges into edge-pairs.

This study has selected the Jaccard Index as the similarity measure based on its simple applicability, robustness, and its property of normalization, i.e.; it accepts the sum of the number of features being linked as a denominator (in this case, the union of the sets of nodes which are receiving copies from a pair of edges). The Jaccard Index defines the similarity between two edges by determining how many of the outside nodes which have received copies (other than the constituent nodes of the edge-pairs in question) are common to them (intersection of the sets of copied nodes in respect of each of the edges in the pair) and dividing this value by the cardinality of the union of the sets of copied nodes of each of the edges in the pair in question. The question of just why this measure brings out the element of collusion between a pair of edges may be answered from the observation that edges with common aims tend to have common neighbors, which is essentially similar to the established concept in community detection algorithms that similar nodes tend to have similar neighbors[68]. Following this logic, if two edges have many common nodes with

[68]This is akin to the concept of *regular equivalence* which states that two regularly equivalent nodes are equivalent if they have similar neighbors who are themselves similar. In this case, similarity means a larger shared set of common nodes which have received copies from each of the pairs of edges. This has been discussed more comprehensively in Chapter 5.

received copies, they are similar. But this aspect needs to be looked at from a normalization perspective, in the sense that if each of the edges has large sets of nodes that have received copies of their mail exchanges, then there is a greater likelihood of the set of intersections being large as well. This is not a desirable outcome since we started to assume that covert edges tend to be secretive about their information exchanges and are, therefore, likely to have comparatively smaller sized sets of copied nodes. Suppose we don't use a normalized measure. In that case, we are likely disincentivizing edges that have been opaque about information sharing. This runs counter to the stated purpose of detecting edges (or edge-pairs), which have been opaque about their communications and thus covert.

We now continue with the second part of the solution definition as follows:

The ENRON mail corpus is defined as a graph G,

such that $G = (V,E\ )$,

where V is the set of all nodes in the graph network.

And E is the set of all edges or mail-pairs in the graph network.

The Selected Set of edges based on a threshold value of Covertness Index is defined as graph G_T,

such that $G_T = (V_T,\ E_T\ )$;

(T is the threshold value of the Covertness Index).

$E_T = \{\ e_{ij}: (\boldsymbol{CI})_{\boldsymbol{ij}} \leq T$ and$i \neq j\}$

And, $V_T = \{\ v_i,\ v_j : (\boldsymbol{CI})_{\boldsymbol{ij}} \leq T$ and$i \neq j\}$.

We may recall the formula for the similarity coefficient (Jaccard Index), which was defined in equation 6.1:

$$(\boldsymbol{S})_{\boldsymbol{ij}} \leftrightarrow_{\boldsymbol{pq}} = \frac{|\boldsymbol{\Gamma}((\boldsymbol{Ev})\boldsymbol{ij}) \cap \boldsymbol{\Gamma}((\boldsymbol{Ev})\boldsymbol{pq})|}{|\boldsymbol{\Gamma}((\boldsymbol{Ev})\boldsymbol{ij}) \cup \boldsymbol{\Gamma}((\boldsymbol{Ev})\boldsymbol{pq})|}$$

where, $v_i, v_j, v_p, \& v_q$ are nodes belonging to the Select Set of edges, which is the result after enforcing a threshold value on the ranking of edges based on their values of the Covertness Index.

Based on the above, the Selected Set of edge-pairs linked by the given *similarity coefficient* (Jaccard Index) may be defined as a graph network$(G_T)_{JI}$:

Such that $(G_T)_{JI} = ((V_T)_{JI}, (E_T)_{JI})$;

($\boldsymbol{JI}$ being the threshold value of the similarity coefficient or the Jaccard Index indicating a linkage between two edges).

$(E_T)_{JI} = \{((\boldsymbol{EP})_{\boldsymbol{ij}\leftrightarrow\boldsymbol{pq}} : (\boldsymbol{S})_{\boldsymbol{ij}\leftrightarrow\boldsymbol{pq}} \leq \boldsymbol{JI} \text{ and } (\boldsymbol{i},\boldsymbol{j}) \neq (\boldsymbol{p},\boldsymbol{q})\}$

where $\boldsymbol{EP}$ represents an edge pair between dyads (edges) $\boldsymbol{(i, j)}$ and $\boldsymbol{(p, q),}$ and $\boldsymbol{S}$ is the *similarity coefficient* (also called the Collusion Index) defined on the linkage between the pairs of edges.

$\&(V_T)_{JI} = \{ v_i, v_j, v_p, v_q : (\boldsymbol{S})_{\boldsymbol{ij}\leftrightarrow\boldsymbol{pq}} \leq \boldsymbol{JI} \,\&\, (\boldsymbol{i},\boldsymbol{j}) \neq (\boldsymbol{p},\boldsymbol{q})\}$.

We saw earlier that the number of pairs of edges of interest (EoIs) detected within the top 2500 edge-pairs ranked in a descending order comes to 21, which works out to a precision value of 0.8 %. When this value is compared with the precision value obtained from a Uniform Distribution model, it's nearly ten times. However, if we compare the detection accuracy on covert edge-pairs arrived at after applying the Collusion Index to the links between edges constituting a pair with the detection accuracy of EoIs arrived at following the application of the Covertness Index, we don't find much of a difference. The detection percentage seems to have slipped marginally from 0.92 % (23 EoIs in a Select set of 2500) to 0.8%. But this doesn't tell the entire story. Detection in the second part of the experiment (application of the Collusion Index) pertains to pairs of edges or four nodes rather than pairs of nodes or two nodes, as in the first part. The detection of communities of four nodes or edge-pairs at any given iteration is a major gain over the detection of just one covert edge. With the Collusion Index application, we have in hand a set of developing communities of covert edges that have common intentions or are likely colluding with each other to pursue some clandestine enterprise, and this is what we had primarily endeavored to achieve in this study. These communities can be increased in size, with further formulations of similarity and collusion in future work.

6.2 Analysis of Results

6.2.1 Data Loss Evaluation

One of the proposed solutions' challenges is the loss of potentially useful data that is an inevitable part of the result as detection performance improves. For example, after the Covertness Index was applied and a threshold value was defined on the ranked edges, we could detect 23 EoIs within the top 2500 edges ranked per their Covertness Index values, which implies that of the 43 edges of interest, 20 were not detected. This loss of potentially useful information has to be looked at as a trade-off between achieving 100% accuracy and expending huge resources in the process and achieving accuracy that is less than 100% but enough to launch a successful surveillance process. To illustrate, we may compare the results obtained with 10,000 top-ranked edges versus those obtained from a set of 2500 (after applying the Covertness Index).

When 10,000 edges were selected, 35 EoIs were correctly detected within the set, an increase of 12 EoIs over the 23 correctly detected in the 2500 set. But, when we consider the resources needed to keep surveillance over 10,000 ties (each tie representing potentially several email exchanges), the cost is at least four times greater than what is required to keep an eye on 2500 ties only. Suppose the covert communities are further distilled out, as was done in the second part of the experiment using the Collusion Index. In that case, all we need to do is to keep a surveillance on the groups instead; however, if we had gone in for a selected set of 10,000 edges after the first stage (i.e., application of the Covertness Index), the potential number of linkages we would have needed to deal with would have come to 10,000 x 10,000, i.e., $10^{8,}$ i.e., a 100 million, rather than the 2500 x 2500 = 6,250,000 or approximately 6 million which is less by a factor of sixteen. This implies that the resources needed to keep surveillance would also increase 16-fold, an unattractive prospect given the very marginal gains in identifying suspect (covert) pairs of EoIs (a gain of only 12 from 23 to 35).

It needs to be mentioned that the discussions on trade-offs between giving up the surveillance on important assets or artifacts are very subjective. There is a possibility that the 'one that got past' might have been of vital importance, and the ones who remained in the dragnet are of relatively less importance, and this is a decision that the field units (or the oversight mechanisms in business organizations) need to take. But we need to keep in mind that when a surveillance process begins, there is no information on who is a covert actor and who isn't. There is just the network structure, the nodes (or actors), their links or ties, and other topological features (metadata basically) that are available initially. The idea is to detect covert activities within what appears to be a benign network (or is a benign network if it's a business organization) without any prior information whatsoever other than the basic network structure. There may be arguments to the contrary, stating that there is invariably some knowledge available *a priori* about any network that needs to be studied. It needs to be emphasized that this *a priori* knowledge should serve to enhance the model developed in this study rather than the other way around. Conversely, suppose the covert structure has completed its payload delivery, and the covert actors are more or less known. In that case, this methodology can only be used for training and testing the model.

There is a need to consider if we are to proceed with another layer of linking the edge-pairs pairs. If we decide to proceed further and operate on the resultant dataset with more similarity measures, the complexity level will increase exponentially. Hence, by losing 12 EoIs through the choice of a lesser sized dataset, we've gained by keeping the complexity at a manageable scale. Table 6.2 shows the trade-off between losing some of the Edges of Interest (covert edges) against the prospect of increasing resources (both computational and human) to enforce proper surveillance. It is presumed that each edge requires at least one unit of resource to watch over it. For instance, 2500 edges will require 2500 units of surveillance, and 5000 edges will require 5000 units, and so on.

Threshold Value	Detection of Covert Edges	Percentage of Loss/Gain	Computational/ Resource Optimization	Percentage of increase in Resource Usage
0	0	0.00%	0	0.00%
2500	23	100.00%	2500	100.00%
5000	26	11.54%	5000	100.00%
10000	35	25.71%	10000	100.00%
55300	43	18.60%	55300	453.00%

Table 6.2 The table plots the losses or gains made in detecting covert assets (in this case, covert edges of interest or EoIs against the prospects of increasing the computational and human resources required to mount surveillance over the edges. An assumption of convenience that is made here is that each edge requires at least one unit of surveillance.

The same table is represented as a graph in Figure 6.2. The interpretation from the plot shown here is clear about the choices that need to be made. Though there is a small loss of a few edges of interest which will remain outside the net of scrutiny, the savings in resources that need to be deployed is quite significant. There should be no hesitation in choosing the smallest sized selection set to begin the scrutiny.

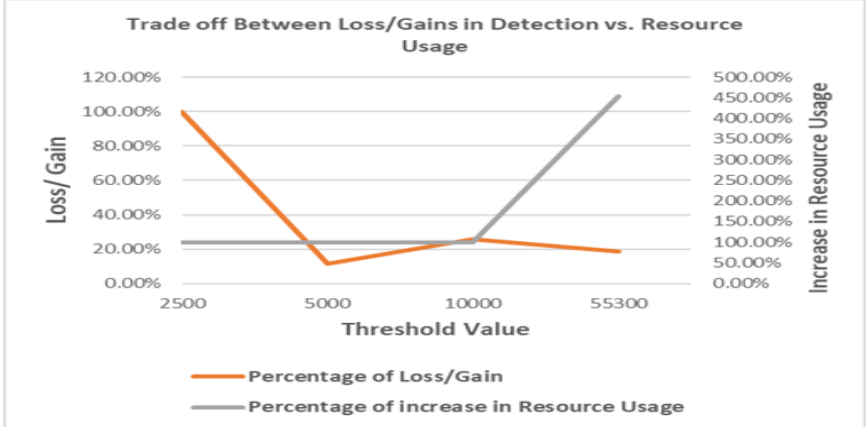


Figure **6.2** The graph illustrates the trade offs between losing assets of interest (in this case, the Covert Edges of Interest or EoIs) and the gains made in optimising resources to be used in surveillance (both computational and human). For the sake of convenience, the assumption made here is that one unit of surveillance resource is required for every edge that needs to be kept under scrutiny.

Threshold Value	Detection of the Constituent Nodes in Covert Edges	Percentage of Loss/Gain in Detection	Computational/ Resource Optimization	Percentage of increase in Resource Usage
0	0	0.00%	0	0.00%
2500	16	100.00%	2500	100.00%
5000	17	5.88%	5000	100.00%
10000	17	0.00%	10000	100.00%
55300	19	10.53%	55300	453.00%

Table 6.3 The numbers in the Table reflect the gain/ loss pattern in the detection of the constituent nodes in the covert edges i.e. the edges of Interest or EoIs. It can easily be seen that the biggest gains in detection occur around the threshold value of 2500 and thereafter, the gains plateau somewhat till the last rank is reached. The massive increases in surveillance resources to achieve the marginal gains in detection of covert nodes from the value of 2500 upwards suggests the need to keep the number of nodes to be kept in the net of scrutiny at 2500 or so.

6.2.2 Node Centric Analysis

Although this dissertation is focused on the edges formed between the constituent nodes in dyads, there is an interesting corollary benefit obtained related to node-centric analysis, which is the more common analytic approach used in social network analysis. Recall that there were 19 employees of ENRON identified as persons of interest (or nodes of interest or NoIs). From these nodes, we extracted 43 edges of interest (EoIs), which became the pivots around which this study developed. It should be interesting to analyze how these nodes fare in the trade-offs just discussed above. Table 6.3 displays the accuracy gains and losses in detecting the (constituent) nodes compared to surveillance resource requirements. As the earlier analysis of covert edge detection, we assume that each node requires at least one unit of surveillance resource (human and computational). It is easily concluded, based on these figures, that the observations made regarding the optimization of surveillance resources for the detection of covert edges (EoIs) are also true for the constituent nodes of

those edges. As in the case of the EoI detection, the table is also represented as a graph in Figure 6.3 to put the bigger picture in perspective. The gains made in terms of covert asset detection after the threshold value of 2500 are marginal and do not warrant the corresponding massive increases in surveillance resource deployment.

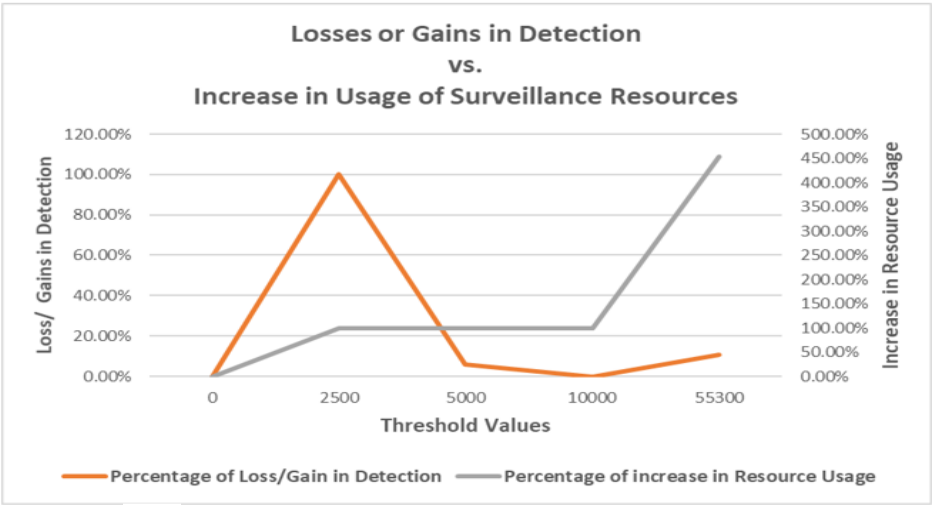


Figure **6.3** Graph showing the trade-off between making gains in detection of constituent nodes (Nodes of Interest or NoIs) against increasing surveillance resources at the assumed rate of one resource per node. It is clear that the gains in detecting covert nodes made after the threshold value of 2500 are marginal and for the sake of these small gains, the massive increases required in surveillance resources are not desirable.

6.2.3 Collusion Index Results Analyzed

The second step in the dissertation was to detect covert communities within the network building on the Covertness Index metric developed in the first part. The mechanism to achieve this was the design of a Collusion Index. The overall methodology was adapted to detect covert communities in which the covert edges (EoIs) with common intentions are grouped using the said Collusion Index. The approach is agglomerative rather than divisive (e.g., using graph partitioning methods). The Collusion Index is a similarity coefficient based on the Jaccard Index, a common similarity measure. The index aims to identify

features common to pairs of covert edges selected based on their high Covertness Index values. The stronger the links between a pair of edges, the more likely they are colluding towards a common (and likely covert) goal. When the Collusion Index was applied to the selected set of 2500 highly covert edges, each pair was assigned a link value, and the edge-pairs were given rankings based on these values. Again, as in the earlier case, it was felt necessary to cull edge pairs down to a manageable number, which could then be kept under the surveillance lens rather than keep all the pairs. (There were 170,000 plus such pairs; in fact, the potential number of pairs can reach six million-plus, just that many pairs will have no common features, and their links will have null or zero values and can be eliminated swiftly).

An examination of the edge-pairs ranked highly based on the values of their Collusion Index revealed that 14 edge-pairs whose constituent edges were EoIs and 74 edge-pairs one of whose constituent edges was an EoI and the other edge had one of its constituent nodes in the set of NoIs (That is, 3 out of 4 nodes in the linked edge-pairs were of interest). Thus, there are 88 edge-pairs of interest in the selected set of 2500 pairs. If 100 units of surveillance resources were to be deployed to keep a watch, at least 3 of them would come across communities that are covert in their transactions and collusion with each other. Communities of 4 nodes are not very large. Hence, one may argue that it is inappropriate to invest resources to track small communities that are likely to produce undesirable outcomes. But, we do well to remember here that bigger communities comprise these smaller units of edge-pairs, and they are likely to be detected in the longer run through observation. The presence of such ‘conspiracy’ subgroups though small in size, offers a valuable window into predicting undesirable activities in a larger, more benign network (in most cases, although networks completely devoted to criminal activities also exist); which is achieved by using widely available structural features without invoking the need to be intrusive. Note that if these cohesive covert communities' activities arouse suspicion in the minds of the investigators, they can always call for methods that will access content through legal methods. By doing so for a very limited number of communities of actors, they will dispel any allegations of widespread surveillance and disruption of privacy.

6.2.4 Reduction in 'Noise.'

We now tackle the 'needle' part of the 'needle in the haystack' problem. The twin applications of the Covertness Index and the Collusion Index have succeeded in identifying a limited set of suspected covert and cohesive 'conspiracy' communities within the network, and how many of these turn out to be malfeasant will be revealed through more intensive surveillance. But the question that arises at this point is if these indices are also playing some role in thinning out the 'haystack.' To answer this question, we need to understand what a haystack would correspond to. Suppose the aim is to identify nodes that are part of covert communities within the network. In that case, the nodes that are completely uninvolved in covert activities and thereby, no interest in the study are equivalent to the 'haystack' in this research. The word 'noise' has already been used to refer to such nodes, also termed nNoIs (nodes NOT of Interest). Table 6.4 shows how the nNoIs are eliminated measurably through the use of the indices.

The above result is another of the unique contributions of this study. The metrics developed therein, in as much as the metrics succeed in reducing the overall set of nodes from which the nodes of interest need to be extracted. The reduction in the number of nodes that are not of interest is fairly precipitous, indicating that the metrics of the Covertness Index and Collusion Index tend to incentivize nodes that are part of covert formations with high ranking scores. The nodes which are likely not covert or opaque in their transactions are correspondingly disincentivized and eliminated from contention. This property of the metrics results in a significant increase in the chances of covert detection formations of which the NoIs are part of in the sense that the denominator (overall set of nodes or entities in the network) of the equation decreases and the numerator (the number of covert nodes or entities) increases.

	Nodes of Interest (NoIs)	Nodes not of Interest (nNoIs)	Total Number of Nodes
ENRON Dataset before experiment	**19**	**6551**	**6570**
After Application of Covertness Index (Select Set of 2500 edges)	**16**	**1780**	**1796**
After Application of Collusion Index (Select Set of 2500 edge-pairs)	**12**	**221**	**233**

Table 6.4 The Table shows the progression of the pruning of the nodes in the ENRON network as the experiment progresses. There is a small decrease in the numbers of the nodes of interest or NoIs from 19 which were originally present in the dataset to 16 after the application of the Covertness Index and then a marginal drop to 12 after the application of the Collusion Index. There is a total reduction of 7 NoIs which is roughly 37%. In contrast, the fall in the numbers of the nodes outside of interest (nNoIs) is very steep, from 6551 at the outset to 1780 after applying Covertness Index and thence to 221 after the application of the Collusion Index. A fall of 97 % approximately. The metrics have thinned out the 'noise' in form of nNoIs very effectively, thereby enhancing detection

6.2.5 Dealing with Data Incompleteness

Another test that the metrics need to undergo is that of resistance to the incompleteness of information. One of the principal features of covert networks is their incompleteness (Sparrow, 1991), and incomplete information may lead to wholly erroneous results regarding the nature of edges and pairings of edges. Dyads that seem innocuous based on the readings of their currently observable parameters may be malfeasant. But, due to lack of information resulting from unavailable data on pre-existing ties or deception and camouflage of exchanges, they emanate signals that may confuse the observer. From this standpoint, the metrics evolved to measure covertness should be resistant to partial or inadequate information.

Table 6.5 shows the status of information available about the ENRON dataset at its inception. We should note that there were only 151 mail in-boxes, to begin with. As shown in the table, 8 of 19 nodes of interest (NoIs) had in boxes associated with them, while 11 NoIs had none. This implies that the email-based information for these NoIs was successfully reconstructed from the available inboxes, and 8 of them were in-boxes belonging to nodes of interest (NoIs). In contrast, the remaining 11 (out of a total of 19 NoIs in the dataset) is seen not have in-boxes, and their email-based information is reconstructed from the available in-boxes. It is also interesting to note from the table that the nodes of interest (NoIs) that don't have in-boxes are not significantly affected by the metric operations. This result indicates that the metrics have resistance to incomplete information and can detect covertness despite this hurdle.

	NoIs with in-boxes	NoIs without in-boxes	Total NoIs
At Inception	**8**	**11**	**19**
After applying Covertness Index (Top 2500 edges)	**8**	**8**	**16**
After applying Collusion Index (Top 2500 edge-pairs)	**6**	**6**	**12**

Table 6.5 Table showing the effect of the application of the two metrics on the numbers of nodes of interest (NoIs). While there is very marginal reduction in the numbers of NoIs with inboxes after the applications (only a reduction of 2 from to 6), the numbers of NoIs which don't have in-boxes to begin with, i.e. have incomplete information about their nature, are also impacted marginally with the numbers reducing from 11 to 6. This reflects indicates the property of resistance to inadequate information inherent to both the metrics proposed.

6.3 Characteristics of the Metrics

Two metrics have been proposed in this dissertation: 1) an edge-based attribute called a Covertness Index, which captures the confinement of information within the constituent nodes of a dyad and 2) a Collusion Index or similarity coefficient based on the Jaccard Index, which measures the strength of the links between pairs of covert edges. It is time to recall the seven desirable attributes of a covert network metric defined in Chapter 1. Each attribute is reproduced below and discussed in terms of how well the proposed covertness

attribute and the similarity coefficient satisfy its description. In this regard, we may recall the desirable attributes defined for the proposed metrics at the beginning of this study. These are now reproduced below and are also discussed from the compliance perspective to show how well the covertness attribute and the similarity coefficient developed during the dissertation fit the requirements.

(a) The attribute needs to be based on easily observable topological features of a network akin to more popular centrality measures.

The Covertness Index is an easily calculated metric with minimal requirements of information from the topology. During the experiments in this study, we've seen that only 151 email inboxes were available in the public domain, to begin with. Much of the information available about the employees who had either been indicted or had been aware in some manner about the insider trading that happened in ENRON was absent (of the 19 employees who were of interest to the study, only 8 had inboxes and regarding the other eight information was gathered from the study of mail headers in the available inboxes). If we consider the total number of employee mail-ids that were part of the corpus and analyzed at some point in the study, the figure comes to 6570 mail-ids. Of this figure, complete information was available for 151 mail-ids. In other words, 6419 mail-ids had incomplete information.

The aspect of incomplete information was factored into the experiment, and the Covertness Index was not much affected by this fact. However, nodes (mail-ids) that didn’t represent the 151 inboxes (whether they were nodes of interest) tended to have higher attrition rates when the index was applied.

The Collusion Index was also designed on the available structural information and was found to be robust enough to yield meaningful results in covert community detection. As has been discussed, this index is more of a regular equivalence feature than a structural one, and its computation is based on the intersection of the sets of nodes that have received copies from the edge-pairs. This information is easily available and, more importantly, doesn’t require any intrusive content-based analysis.

(b) The attribute should be a *minimalistic* one that will not require many components or complicated manipulations of input variables; in other words, it shouldn't consume too many computational resources or bandwidth to calculate.

Both metrics proposed in this study are simple, and their calculations are relatively uncomplicated. The computational complexity required is minimal, and the components of their formulae are also easily extracted from the network topology.

(c) The attribute must be capable of being used for *non-intrusive analysis*, i.e., there should be no need for the contents of the information exchanges to be known, which is especially important for bypassing existing data privacy laws, encryption mechanisms, varying policies across countries which allow differential access to information, legal strictures, inadequate information about the network and covert channels of communication which may not be apparent to the surveillance team.

As was pointed out in (a) above, the Covertness Index calculation relies on how the constituent nodes mark many copies in a dyad. This information is based on the available meta-data of the network independent of content in the emails. Similarly, the Collusion Index also relies on the commonality of the nodes receiving copies of emails from the edge-pairs' constituent edges, which has no dependency on the emails' content. Thus, this property is also satisfied with the proposed metrics.

(d) It should be easily *applied* across all networks (node or edge or even higher groupings). That is, the attribute should be of such a nature that it can be combined easily with existing metrics.

We have seen that the Covertness Index was applicable at the level of edges between nodes, which is the most basic structure in a network. Later, when the Collusion Index was applied, the Covertness Index was used in the input selected set to rank the edges by covertness value. The Collusion Index, which is applied to bring out the strength of the

linkages between covert edge-pairs, is also amenable to being repeated for higher-order structures in the networks, e.g., pairs of edges (or quartets) and so on. Thus, both metrics are easily adaptable for application at any layer within the network.

(e) The attribute should have the ability to act as a *linkage* mechanism, tying together disparate nodes, edges, triads, or higher group formations based on some formulation of commonness (or cooperation as the term is viewed in this study). A corollary of this characteristic is that the attribute should exhibit some form of *structural transcendence*, i.e., there need not be any tangible structural links between entities that might be grouped through its application.

Both the metrics have a tie or edge-based mechanism. Although the Covertness Index doesn't link any structures, its subsequent usage as a ranking parameter allows the constituent nodes of the dyad to be yoked into larger community structures using the second metric, the Collusive Index. This is a linkage mechanism that fits the regular equivalence metric description rather than a structural one. That is, the edge-pairs to be linked may not have any linkages amongst themselves structurally. Edge-pairs, which are in similar neighborhoods, are liable to form strong collusive bonds.

(f) More importantly, it should allow the investigator to *reduce* the sheer volume of data that is not interesting and increase the chances of obtaining positive results.

This aspect was discussed in detail earlier in this chapter. Both metrics succeed in reducing the size of the overall dataset considerably. Ultimately, the 6780 nodes that we began with was reduced to just 221 by the time the experiment ended. Thus, the size of the haystack was pared down considerably. While this happened, the size of the set of edges of interest (EoIs) remained more or less unaffected, a development that significantly increased the chances of detecting covert edges.

(g) The attribute should be *dynamic*, i.e., it should change in time and still retain its usefulness despite evolutionary changes in the network structure. If there are changes in

the network's topology, changes in the attribute should be able to track the changes and even predict the future shape of the network.

Although this dissertation is a non-longitudinal one and doesn't factor in time as a parameter, it can well be concluded that the Covertness Index, in particular, can be key in detecting unusual patterns of communication over time. For example, suppose the value of the Covertness Index of a particular edge increases sharply in a particular period. In that case, it signifies the start of some covert collaboration by the constituent nodes. Similarly, if an edge that had a high Covertness Index value suddenly comes down within a short timespan, this would indicate that the covert enterprise has either completed its formalities successfully or has transferred the task of payload delivery elsewhere in the network. Both possibilities are indicators of probable covert collaboration, and the variations in Covertness Index values over time capture this fact.

Thus, both the Covertness Index and the Collusion Index meet the above criteria and provide better than reasonable results to detect a much higher proportion of covert edges than would be possible through uniform distribution.

6.4 Recapitulation of the Steps

Figure 6.8 recapitulates the steps taken in this dissertation. The study started with the ENRON email corpus, which had 570,000 plus mail exchanges amongst 151 employee email ids. This figure was rationalized to 55,300 unique mail pairs, which represented pairs of employee mail-ids who had exchanged emails, and each pair of nodes (mail-ids) which had exchanged at least one mail was assigned an edge value of 1. The study defined the concepts of Relation Sets, Shared Relationship Sets, and Neighborhood Sets, designed to bring out the nature and quantity of mail exchanges and mail copies. These three sets were defined on the edges between each dyad's constituent nodes and were made to convey the nature of the pair's neighborhood.

Layered on top of these three concepts is the concept of an Edge-Vertex, a function defined on edge between the constituent nodes of a dyad that had exchanged at least one mail. An Edge-Vertex of a dyad is a list containing the following elements:

(i) The identities of the constituent nodes of the dyad.

(ii) The sum of the Relationship Set elements conveys the total number of emails exchanged via the edge between the constituent nodes.

(iii) The sum of the Shared Relationship Set elements, i.e., the total number of emails copied from the mail exchanges between the constituent nodes to nodes outside the dyad.

(iv) The Covertness Index, i.e., the ratio of the sum of the Shared Relationship Set elements and the sum of the Relationship Set elements deducted from 1.

(v) The Neighborhood Relationship Set, whose elements are the sets of nodes that are the recipients of copies from each mail exchange between the constituent nodes.

The Edge-Vertex function is a unique concept that allows multiple operations to take place with the edge or tie, acting as the fundamental unit. This entity calculates the Collusion Index between pairs of edges in the experiment's second leg. The Collusion Index allows us to determine if two covert edges are linked in a meaningful manner, i.e., whether they have common covert intentions.

Figure **6.4** Chart showing the sequence of steps followed by the study to arrive at the detection of covert communities which have common aims (intentions) within the overall architecture of the social network in question (the ENRON mail corpus).

Chapter 7

Conclusions and Future Work

The dissertation started out answering Krebs's challenge in his groundbreaking study of the 9/11 attacks (2001), who states thus –"The big question remains – why wasn't this attack predicted and prevented? Everyone expects the intelligence community to uncover these covert plots and stop them before they are executed. Occasionally plots are uncovered, and criminal networks are disrupted. But this is very difficult to do. How do you discover a network that focuses on secrecy and stealth?"

The three big questions that this quote expresses are quite apparent. First, why are the covert actors not detected before the incident happens? Second, how does one recognize a covert actor and, by extension, a community of covert actors who habitually engage in stealth and secrecy? The third question relates to Krebs' term 'plot,' which essentially alludes to a conspiracy by a community of covert actors. To rephrase the question, even if we manage to ascertain that some actors are covert or trying to hide their activities within a network, how do we determine if a group of such covert actors has the same aims?

These core questions raise some related issues which this dissertation has attempted to address. Firstly, can an investigator detect covert entities within an otherwise bright network without needing to intercept communications amongst the actors or without being intrusive? Second, how do we translate a more or less abstract sociological concept like covertness in a mathematically meaningful manner? A corollary to this question is whether covertness should be studied as a global attribute of an entire network or studied as a unit attribute that is manifest across all social networks? Is covertness an attribute akin to more popularly studied and invoked centrality measures like degree (or incidence), betweenness, closeness, and eigenvalue? If so, how do we breakdown the covert aspects of entire networks or parts of such networks into an indivisible atomic aspect?

The dissertation has chosen to define the concealment or confinement of information exchanged between entities or, more precisely, between a pair of nodes. To be more specific, the study has tried to translate the confinement of information exchanged between a pair of nodes into a variable that can be enumerated. In this research, the pathway along which such information is exchanged, i.e., the tie or the edge in SNA terminology, is used as the most basic unit of analysis rather than the more popular vertex node in traditional SNA literature. Since the dataset used here is the email corpus of the ENRON Corporation, a unit of covertness was sought that would reflect the transparency or mail exchange. The opacity in a dyad is defined as the proportion of emails that have been exchanged between two nodes and which have not been marked out as copies. The index was further modified to bring into play the volume of mail transactions between nodes, which was felt to have a bearing on the relative importance of the covertness metric of dyads known to have heavy exchanges.

This simple index of covertness of mail exchange of information confinement was computed for all edges within the network and used as a ranking mechanism to arrange the edges in descending order. A particular number of top-ranked covert edges was selected using a heuristic (minimum) threshold value of the Covertness Index. This select set was used as an input for developing a similarity coefficient, based on the Jaccard Distance formula to link pairs of top-ranked covert edges. These links are termed the Collusion Index, and this index is the first step towards identifying communities of covert edges having common intentions, otherwise known as conspiracy subgroups.

The Collusion Index represents the strength of the linkage between two covert edges, i.e., the higher the index value, the stronger the relationship. After calculating this index, further exercise was taken to arrange the edge pairs in descending order of their link strengths (Collusion Index values). A selection was made out of this set, giving a reduced set comprising edge-pairs with link strengths higher than a heuristic threshold of the Collusion Index value. This set represents those covert communities of actors within the most cohesive and strongly linked network. Hence, it represents the best pool of candidate structures for investigators to focus on in their search for conspiracy sub-networks. The

experiment results were encouraging. A large proportion of covert nodes (actors) and covert community structures were detected using both the Covertness Index and the Collusion Index in succession. Figure 7.1 gives an overview of the different stages in developing both indices and their net impact on the experiment's success.

Figure 7.1

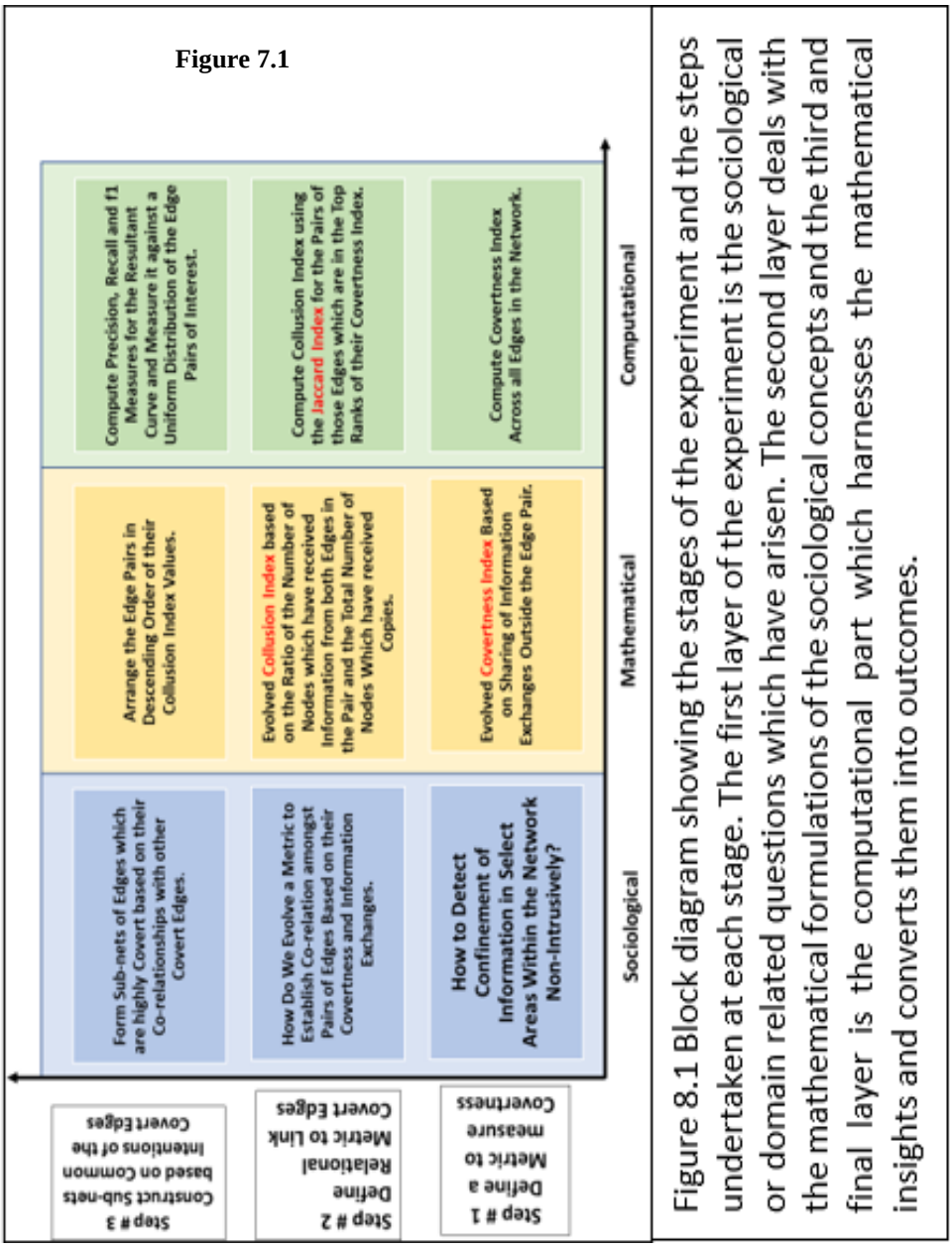


Figure 8.1 Block diagram showing the stages of the experiment and the steps undertaken at each stage. The first layer of the experiment is the sociological or domain related questions which have arisen. The second layer deals with the mathematical formulations of the sociological concepts and the third and final layer is the computational part which harnesses the mathematical insights and converts them into outcomes.

While the present research has concluded, several lingering issues merit attention. The first issue concerns the additional steps that need to be taken to identify larger communities within a network. So far, the results have provided us with communities of at most four nodes. In a comparatively small-sized network where there are not many nodes, 4-node communities of 4 nodes are significant. But, if the network sizes are bigger, we may need to look for larger sized communities with common (and covert) aspirations.

The traditional definition of communities in networks is that they are groups of nodes having greater ties internally than with the rest of the network. Traditionally a network community is defined as a group of nodes having greater ties among themselves than with the rest of the network. It bears repeating here that some community detection methods are more popular than others, but no single technique is applicable across all social networks. Since the study started by developing a Covertness Index metric defined on edges rather than on nodes, the broad approach was agglomerative, i.e., one based on detecting clusters of edges with high covertness values using a similarity measure (the Collusion Index). At this juncture, the choice available to us is to continue the aggregation or clustering of the communities that have already been identified through the second-ranking processes. We may continue to use the Jaccard Index as in the second step or choose another similarity measure that links the 4-node communities already in existence. As discussed earlier, several candidate methods exist for identifying these larger communities.

Another method is to cluster the available communities as neighbors of the nodes that have received copies. First, the nodes that have received copies of emails from different edge-pairs (obtained after applying the Collusion Index to link covert edges) can be ranked based on the number of links to different edge-pairs. Each of these nodes becomes a centroid for the cluster of edge-pairs that form its neighborhood. We may heuristically select a certain number of clusters from the ranking, which will account for most of the existing nodes in the Collusion Index select sets. This community detection mechanism can be further advanced by checking the number of edges between the clusters and merging clusters based on a certain minimum heuristic, i.e., merging clusters whose component nodes share more than a minimum threshold number of edges with each other.

The proposed additional steps for detecting (additional or larger) communities in the network are illustrated in Figure 7.2. The figure shows a fictitious email-based social network in which there are seven dyads: e_{ab}, e_{cd}, e_{ef}, e_{gh} e_{ij}, e_{mn}, and e_{kl}, and six other nodes that have received copies. Of these nodes, *p* receives copies of emails from 5 dyads, *r* receives copies from 4 dyads, *q* receives copies from 3, *s* from 2, *u* from 1, and *t* from none. The details of the dyads, contributing copies of emails to each of the above nodes are given in table 7.1. Based on the table, the detected communities are centered around, as shown in Figure 7.3.

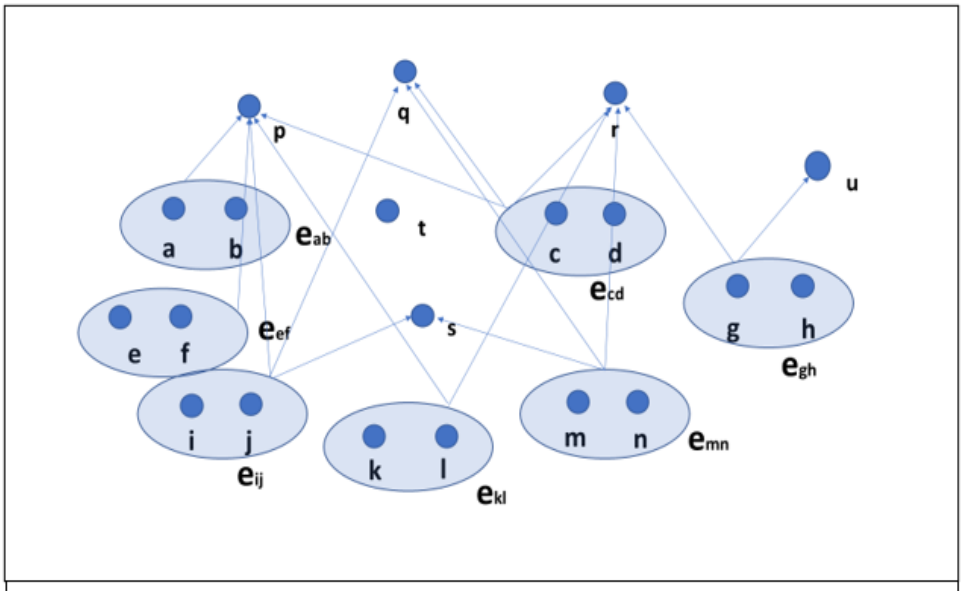


Figure 7.2 The figure shows an email based network with dyadic pairs e_{ab}, e_{cd}, e_{ef}, e_{gh} e_{ij}, e_{mn} and e_{kl} which are sending copies of mails exchanged among their constituent pairs of nodes to nodes $p, q, r, s, t,$ and u.

Node-id	Edge-Pairs from which copies received	Number of edge-pairs
p	e_{ab}, e_{cd}, e_{ef}, e_{ij}, e_{kl}	5
r	e_{kl}, e_{cd}, e_{mn}, e_{gh}	4
q	e_{cd}, e_{ij}, e_{mn}	3
s	e_{ij}, e_{mn}	2
u	e_{gh}	1
t		0

Table 7.1 The table above shows the same email based network of Figure 7.2 showing which of the dyadic pairs have sent copies of their mail exchanges to nodes $p, q, r, s, t,$ and u.

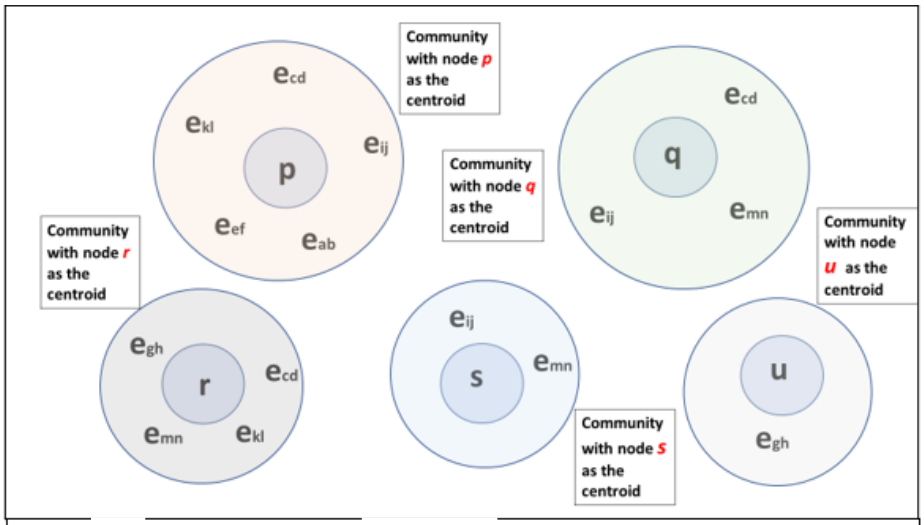


Figure 7.3 Based on the table 7.1 communities of dyad pairs with the nodes as centroids are shown. It may be seen that dyad pairs may recur across communities. That is, there is an overlapping community structure that has evolved.

Another important issue deserving attention concerns the element of time. This study's orientation has been non-longitudinal, i.e., the element of time as a variable is not considered. The same steps and experiments proposed in this dissertation can also be executed with a longitudinal slant. That is, a time element can be introduced into the equations to refine the results further. The discussion on *link prediction models* in Chapter 5 offered a glimpse into the problem of predicting a network's evolution, i.e., predicting what the structure of a network will look like after a short period Liben-Powell and Kleinberg (2007). The utility of time in unraveling covert communities' activities and eventual detection is a practice very prevalent in law enforcement. For example, it is common for offenders who plan a crime to stop conversing over their mobile phones after the planning is over and before the actual incident. More often than not, this is the behavior that the officers mounting surveillance seize upon before launching pre-emptive raids. The trends just opposite to this are also interesting for predictive policing. If a group of offenders is planning a major heist, the frequency of their calls to each other over their phones tends to increase for a while. There is a brief spike in their call frequencies that last until the planning phase is over, and the waiting phase begins. A sudden change (increase or decrease) in communications is a time-dependent phenomenon. For an automated investigation into such patterns, network models would need to be studied and stored for comparison at earlier points in time.

The above time-based approaches look at network evolution as a whole over time. But, recent studies have disengaged the problems of community evolution from that of overall network evolution. Spiliopoulou and Aggarwal (2011), in their survey on evolution in social networks, argue that an evolving community is not necessarily part of an evolving graph. They treat the problem as a subset of the larger problem of evolving networks or the study of volatility in networks. In a survey on community detection algorithms, Fortunato (2010) devoted an entire chapter to detecting *dynamic communities*. Sharma, Feng, Singhal, Kuang, and Srivastava (2015), in their paper on predicting small group accretion in social networks, also dealt with group evolution dynamics through increment accretion by process size increment the addition of more members. As a part of this

problem, they defined two subproblems: *incremental accretion* and *subgroup accretion*. In the first subproblem, given a certain size group, the paper predicts the likelihood of one more members being absorbed in it. The second subproblem, i.e., subgroup accretion, is the problem of incremental accretion on all possible subgroups of a given group to yield new prediction scores for the incremental groups. The intuition behind this idea is that many future groups are formed through these two processes, given a group collaboration history. The paper, fittingly enough, aims to build models that predict future groups likely to form through these two mechanisms. It notes that the work is an initial step towards a more general higher-order group prediction problem. The next steps proposed in this dissertation also point in this direction. A good third step in the experiment would be to work towards identifying communities or groups no larger in size than perhaps ten nodes while preserving their homogeneity to a large extent.

One constraint with longitudinal processing is the huge amounts of data to be crunched per snapshot for even a moderate-sized network. For instance, if we consider the solution given in this study as applying to only one point in time, we would need to repeat it; if we treat this as giving the solution at one particular point in time, we would need to repeat it for a sizeable number of timestamps for the results to be reliable on both the evolution of the overall network and evolution of communities within the network to be reliable. To cite an example, if we're going to predict what the 2500 potentially covert communities identified in this study are going to look like at a certain timestamp in the future, we need to consider each of the 2500 communities individually and also their constituent edges (4 per community) and use all of them in the algorithm that has been worked out for prediction purposes, the complexity will increase by leaps and bounds. The model used in this dissertation is not very large, and even so, the introduction of a time variable would pose formidable challenges, to say the least.

Most social networks today (especially those on the internet) can have up to millions of nodes, but the distribution of activity among the network members follows Zipf's law[69].

[69] Zipf's law is an empirical law which states that when events are ranked (denoted as R) from highest frequency to lowest frequency, the frequency (*f*) of the event is inversely proportional to the rank of the

For incremental timestamp-based methods, this implies that huge initial matrices must be built at each time point and that sophisticated heuristics are needed to fill the entries for new nodes with imputed average values of acceptable reliability. The sheer numbers of such nodes make it likely that the derived values may affect the model's viability. Even if we assume that there are no side effects on the resulting model's quality, there will surely be effects on the execution time and storage demand. As has been repeatedly stressed in this study, surveillance on covert communities within networks to pre-empt any adverse incident is a process that requires quick results with a minimum investment of resources. Hence, there is a need for economical or even frugal resource use when adapting models to such networks. How this can be achieved with incremental methods that process massive sized matrices (or tensors) is still an open issue.

Several studies have deployed machine learning techniques for predicting models of the evolution of social networks (Liben-Nowell and Kleinberg 2007; Krautz et al. 1997; Raghavan 2002). The idea of melding machine learning (ML) to link prediction using features like the Covertness Index and Jaccard Index as inputs are worth trying out in future research. If a partition-based approach were taken to the problem, these features could also be used to increase cluster modularity after ML-based classifiers had partitioned the network.

A third important topic that needs consideration in the future is to conclude that domain specificity remains in these kinds of problem-solving approaches to detecting covert edges. As discussed earlier, the main limitation of such research is that the devised solutions are tailored to particular contexts. What may be good enough for an email-based social network like the one proposed here may fail to do well in other scenarios. It is best to maintain a library of solutions and apply the best fit after comparing each solution's results.

One of the major problems with deciding which approach to adopt for community detection in networks (detection of covert communities or sub-groups is a specific sub-set

---

event. The result of Zipf's Law is equivalent to the Power Law. It is represented as $f \propto \frac{1}{R^{\alpha}}$ where α is some positive exponent.

of this) is that different approaches appear to do well in specific scenarios. In other words, there is no universal secret ingredient to success. In this context, it's worth revisiting what Fortunato states in his exhaustive survey paper (2010)

> "A newcomer, who wishes to find clusters in a given network and is not familiar with clustering techniques, would not know, off-hand, which method to use, and he/she would hardly find indications about good methods in any single paper on graph clustering, except perhaps on the method presented in the paper. So, people keep using algorithms because they have heard of them, or because they know that other people are using them, or because of the reputation of the scientists who designed them. Waiting for future reliable benchmarks that may give an objective assessment of the quality of the algorithms, there are at the moment hardly solid reasons to prefer an algorithm to another…However, we want to stress that there is no such thing as the perfect method, so it is pointless to look for it. Among other things, if one tries to look for a very general method that should give good results on any type of graphs, one is inevitably forced to make very general assumptions on the graph's structure and communities' properties. In this way, one neglects many of the system's specific features, leading to more accurate detection of the clusters. Informing a method with features characterizing some types of graphs makes it far more reliable to detect the community structure of those graphs than a general method, even if its applicability may be limited. Therefore in the future, we envision the development of domain-specific clustering techniques. The challenge here is to identify the peculiar features of classes of graphs, which are bound to become crucial ingredients in the design of suitable algorithms." (p.91)

Fortunato prescribes that for optimal results, the practitioner should leverage

> "specific information about a graph, whenever available… For instance, it may be that one has some information on a subset of vertices, like demographic data on people of a social network. Such data may highlight relationships between people that are not obvious from the network of social interactions. In this case, using only

> the social network may be reductive. Ideally, one should exploit both the structural and the non-structural information searching clusters, as the latter should be consistent with both inputs. How to do this is an open problem." (p.91)

The above remarks set the matter in perspective. When looking for communities in networks, domain specificity is unavoidable. The researcher must be aware of multiple aspects of the social network being studied, the nature of the entities that will likely populate the communities, how big (or how small) the community sizes should be. This dissertation leverages the availability of certain structural features of the ENRON mail corpus, notably the mail exchange between employees (to the extent these are present in the 151 in-boxes made public), links formed between mail-id pairs through marking of copies, etc. If one of the structural features said the copied mail information was unavailable, an entirely different approach to the problem of confinement of information exchange would have been required. Or, suppose information exchange in the network was in the form of telephone calls or social media messaging rather than email exchanges. Different approaches would have been called for since far more structural information would be available. In a social media platform like Facebook, sentiment-related data is easily observed without having to dig deep. In telephone call detail records, the voice-based exchanges are usually linear without the equivalent of the mail copies on which this study relies so heavily.

It is clear that social network research needs to be a close collaboration between practitioners in the computational field and domain experts from whom the network information is derived. The nature of the domain information should guide the selection of computational methodologies to be woven into the very fabric of the algorithmic solutions. There are calls for universal solutions across all domains of social networks, indeed of networks as a whole, but a "Grand Unified Theory" for detecting communities in networks remains distant, at least in the foreseeable future. One of the advantages of close domain-based collaboration would be the access that the researcher in the computational domain would have to the vast literature in the field where the domain expert works. Careful

choices of features in the network architecture can then be made with concomitant positive effects on the outcomes.

Things are no different in the study of covert networks (or covertness in networks). We saw in Chapter 2 how different such networks could be from each other: in their structure, resilience, secrecy-efficiency trade-offs, and the ways and means by which secrecy is maintained. So what may work for a terror-related covert network might not be optimal for a criminal network. There are divergences even within criminal networks, with different structures seen in networks related to different forms of crimes or criminal enterprises that underlie such networks' functioning. And even if a 'panacea' type solution were to be found for social networks which are wholly engaged in criminal activities, it still wouldn't work for bright or open networks like business or commercial enterprises where there may be embedded sub-structures that are covert and whose activities are harmful and even illegal to the interests of the organization (as in the case of ENRON where an overwhelming proportion of the employees were law-abiding and ignorant of the designs of the few bad actors). In other words, the research must be aware of what artifacts to look for, and this won't be possible if the researcher is not very familiar with the background and literature of the domain subject. More importantly, given the nature of social networks generally, computational research needs to accept the reality of specific best-fit models rather than a one-size-fits-all approach. There also needs to be an acceptance of approximate solutions whose accuracy may be enhanced incrementally by adding new features like variables and fine-tuning parameters.

Finally, despite the extensive research on covert networks, an overwhelming amount of it has focused on *post facto* analysis, i.e., the research is already aware of who the bad actors are, their activities, and how their removal from the network disrupts the network's activities as a whole. In a real-world scenario, such information is a luxury seldom available to the field investigators, who are often racing against unknown deadlines and competing against unseen foes. The task of predicting what payload a covert network might deliver and within what timelines are far more complicated than the study of past events. A related issue is the tendency to label entire networks as covert or dark. The

tendency needs to be resisted because most covert networks are incubated. Most of these covert networks are incubated within larger and essentially benign networks, where they seek shelter and try to camouflage their intentions. It is my considered opinion that for any predictive approach to succeed in the detection of covertness, the focus has to be broader. There needs to be an acceptance that malfeasant actors will also have more licit pursuits and may use such overt activities to mask nefarious actions. The scrutiny should be directed to the network's larger structure, first drilling down to the sub-structures of interest.

The field of covert network analysis is an intensively studied one, and interest in the predictive analysis is gaining traction. Powerful new algorithms are being developed. Their focus is on capturing the multifaceted nature of covertness, collusion, and dynamism in the structure of such networks (and communities within these networks) and on expressing these concepts in a computationally comprehensible way. Many open issues remain, however, while newer ones continue to emerge with further research results. This dissertation has attempted to bring to the fore some of these challenges in this field and present a somewhat unified underpinning to the whole set of dynamics involved.

REFERENCES

1. Acar, Evrim, and Dunlavy, Daniel M., Kolda, Tamara G. (2009). Link Prediction on Evolving Data Using Matrix and Tensor Factorizations. In Proceedings of the Workshop on Large Scale Data Mining Theory and Applications. ICDM Workshops:262-269.
2. Adafre, Sisay F., and Rijke, Maarten de. (2005). Discovering missing links in Wikipedia. LINK-KDD '05: Proceedings of the Third International Workshop on Link Discovery.
3. Adamic, Lada A., and Adar, Eytan. (2003). Friends and neighbors on the web. Social Networks, 25(3):211-230.
4. Ahmed, Elmagarmid, and Ipeirotis, Panagiotis G., and Verykios, Vassilios. (2007) Duplicate Record Detection: A Survey. In IEEE Transactions on Knowledge and Data Engineering 19 (1):1âAS16.
5. Aiello, L. M. et al. Friendship prediction and homophily in social media. ACM Transactions on the Web 6, 1–33 (2012).
6. Airodi, Edoardo M., and Blei, David M., and Xing, Eric P., and Fienberg, Stephen E. (2006). "Mixed Membership stochastic block models for relational data, with applications to protein-protein interactions." Proceedings of International Biometric Society-ENAR Annual Meeting.
7. Aldrich, Howard E., 1979. Organizations and Environments. Englewood Cliffs: Prentice-Hall.
8. Alkhereyf, S., & Rambow, O. (2017, August). Work hard, play hard: Email classification on the Avocado and Enron corpora. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing* (pp. 57-65).
9. Anthonisse, J.M., The rush in a directed graph, Technical Report BN 9/71, Stichting Mathematisch Centrum, Amsterdam (1971)
10. Apter, Michael J. (1970). The Computer Simulation of Behavior. London: Hutchinson & Co., p. 83.
11. Arce, D. G., Croson, R. T. A., & Eckel, C. C. (2011). Terrorism experiments. Journal of Peace Research, 48, 373–382. doi:10.1177/0022343310391502
12. Asal, V., Rethemeyer, RK. 2006. Researching terrorist networks. Journal of Security Education 1 (4), 65–74.

13. Asal, V., & Rethemeyer, R. (2008). The Nature of the Beast: Organizational Structures and the Lethality of Terrorist Attacks. *The Journal of Politics, 70*(2), 437-449. Ayling, J., 2009. Criminal organizations and resilience. J Int Crim Justice 37(4):182–196.
14. Backstrom, L., Huttenlocher, D., Kleinberg, J., & La, X. 2006. Group formation in large social networks: Membership, growth, and evolution. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 44–54). New York: ACM.
15. Baker, W., Faulkner, R., 1993. The social organization of conspiracy: illegal networks in the heavy electric equipment industry. American Sociological Review 58 (6), 837–860.
16. Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and Vicsek, T.: "Evolution of the social network of scientific collaboration." Physics A, 311(3-4):590-614, 2002.
17. Barabasi, Albert-Laszlo, and Albert, Reka. (1999) The emergence of Scaling in Random Networks, Science, 286(5439):509.
18. Bayse, W.A., and Morris.C.G., 1987. "FBI automation strategy: Development of Al applications for national investigative programs." Signal Magazine May.
19. Beránek L Novák V., 2006. Use of Graph Theory and Networks in Biology: Konference Technical Computing, Prague, dsp.vscht.cz.
20. Berlusconi, G., Calderoni, F., Parolini, N., Verani M.; Piccardi C. –"Link prediction in criminal networks: A tool for criminal intelligence analysis."; PLoS ONE, 01 January 2016, Vol.11(4), p.e0154244.
21. Bilgic, Mustafa, and Namata, Galileo M., and Getoor, Lise. (2007). Combining collective classification and link prediction. In Proceedings of the Workshop on Mining Graphs and Complex Structures at ICDM Conference.
22. Black, D. (1976). The behavior of law. New York: Academic Press.
23. Black, D. (1998). The social structure of right and wrong (Rev. ed.). San Diego: Academic Press.
24. Black, D., 2000. "Dreams of Pure Sociology." Sociological Theory 18:343-67.
25. Black, D. (2004). The Geometry of Terrorism. Sociological Theory, 22(1), 14-25.

26. Borgatti S.P. (2003). The Key Player Problem. In: Carley KM, Pattison P (eds) Brieger R.Workshop Summary and Papers, National Academy of Sciences Press, Dynamic Social Network Modeling and Analysis.
27. Borgatti S.P. (2006a) Identifying sets of key players in a social network. Compute Math Organ Theory 12 (1):21–34.
28. Borgatti S.P. and Everett M. G., (2006b) "A graph-theoretic perspective on centrality," Social Networks, vol. 28, no. 4, pp. 466–484.
29. Boches, D. J. (2020). Social geometry and the 'terrorism' label. Dilemmas, 13(1), 147-168.
30. Bouchard, M., 2007. On the Resilience of Illegal Drug Markets. *Global Crime* 8 (4): 325-344.
31. Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, *25*(2), 163-177.
32. Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. Social Networks, 30(2), 136-145.
33. Brass D, Butterfield K, Skaggs B. 1998. Relationships and unethical behavior: a social network perspective. Academic Management Review 23(1):14–31.\
34. Brantingham, P. J., & Brantingham, P. L. (Eds.). (1981). *Environmental criminology* (pp. 27-54). Beverly Hills, CA: Sage Publications.
35. Brantingham, P. L., & Brantingham, P. J. (2004). Computer simulation as a tool for environmental criminologists. *Security Journal*, *17*(1), 21-30.
36. Brafman, Ori. and Rod A. Beckstrom. 2006. *The Starfish and the Spider: The Unstoppable Power of Leaderless Organizations*. New York: Portfolio.
37. Brin, Sergey, and Page, Lawrence. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):107-117.
38. Broder A.Z., GlassmanS.C., Manasse M.S.: and Zweig G.: Syntactic clustering of the web. Computer Networks and ISDN Systems, 29(8-13):1157–1166, 1997.
39. Bu, D. B., Zhao, Y. &Cai, L. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Research 31, 2443–2450 (2003).
40. Burstein, P. (1991). Policy Domains: Organization, Culture, and Policy Outcomes. Annual Review of Sociology, 17(1), 327-350.

41. Calderoni, F., 2011. Strategic positioning in Mafia networks.
42. Campana P, Varese F (2013) Cooperation in criminal organizations: kinship and violence as credible commitments. Ration Soc 25(3):263–289.
43. Campbell, Bradley. The Geometry of Genocide: A Study in Pure Sociology. Charlottesville: University of Virginia Press, 2015.
44. Campbell, Bradley; Manning, Jason"Social Geometry and Social Control." In: DEFLEM,Mathieu(org.). The Handbook of Social Control. Hoboken: Wiley-Blackwell, 2019, pp. 50-62.
45. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Sci. Rep. 3, 1613 (2013).
46. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. Minimum curvilinearity enhance topological prediction of protein interactions by network embedding. Bioinformatics 29, 199–209 (2013).
47. Cao, Z., Zhang, Y., Guan, J., & Zhou, S. (2018). Link Prediction based on Quantum-Inspired Ant Colony Optimization. *Scientific Reports*, *8*(1), 1-11.
48. Carley K., Lee J.S., and Krackhardt D. 2002. Destabilizing Networks1. *Connections* 24 (3): 79-92.
49. Carley, K., Dombroski, M., Tsvetovat, Reminga, J., Kamneva, N. 2003. Destabilizing Dynamic Covert Networks. In: Proceedings of the 8th International Command and Control Research and Technology Conference, Washington.
50. Carley, K. M. (2003). Dynamic network analysis. In R. Breiger, K. Carley &P. Pattison (Eds.), Dynamic social network modeling and analysis: Workshop summary and papers (pp. 133–145).Washington, DC: Committee on Human Factors, National Research Council.
51. Carley, K. M. (2006). A dynamic network approach to the assessment of terrorist groups and the impact of alternative courses of action. In Visualising Network Information. Meeting Proceedings RTO-MP-IST-063, Keynote 1 pp.KN1-1–KN1-10. Neuilly-sur-Seine, France: RTO. Retrieved on 20.06.2020 from https://apps.dtic.mil/dtic/tr/fulltext/u2/a477116.pdf
52. Carley, K., 2006. Destabilization of covert networks. Compute Math Organ Theory 12:51–66.

53. Carrington PJ, and van Mastrigt SB. 2013. Co-offending in Canada, England, and the United States: a cross-national comparison. *Global Crime* 14 (2-3): 123-140.
54. Chambers D, Wilson P, Thompson C, Harden M.2012. Social Network Analysis in Healthcare Settings: A Systematic Scoping Review.EDGES OF INTEREST: 10.1371/journal.pone.0041911
55. Chapanond, A., Krishnamoorthy, M., and Yener, B. (2005), "Graph-Theoretic and Spectral Analysis of Enron Email Data," Computational & Mathematical Organization Theory, 11, 265–281.
56. Chaurasia, N., Dhakar, M., Tiwari, A., & Gupta, R. K. (2012). A survey on terrorist network mining: Current trends and opportunities. International Journal of Computer Science & Engineering Survey, 3(4) Retrieved from http://www.airccse. org/journal/ijcses/papers/3412ijcses05.pdf. doi:10.5121/ipcses.2012.3405
57. Chung, Fan, and Zhao, Wenbo (2010). PageRank and random walks on graphs. Proceedings of the "Fete of Combinatorics" conference in honor of Lovasz.
58. Clauset, Aaron, and Moore, Christopher, and Newman, M. E. J. (2008).Hierarchical structure and the prediction of missing links in the network. Nature453:98-101.
59. Clutterbuck, L., 2008. Rethinking Al Qaeda: Leaderless Jihad: terror networks in the twenty-first century.
60. Coady, W.F., 1985. "Automated link analysis - artificial intelligence-based tool for investigators." Police Chief 52(9): 22-23.
61. Cockbain E, Brayley H, and Laycock G. 2011. Exploring Internal Child Sex Trafficking Networks Using Social Network Analysis. *Policing* 5 (2): 144-157.
62. Cohen P. R. and Morrison C. T., "The HATS simulator," Proc. of the 2004 Winter Simulation Conference, 2004.
63. Coleman JS. 1990. Rational action, social networks, and the emergence of norms. *Structures of power and constraint*: 91-112.
64. Coles, N. (2001). It's not what you know – it's who you know that counts: Analyzing serious crime groups as social networks. British Journal of Criminology, 41, 580 – 594.
65. Cooney, Mark. Warriors & Peacemakers: How Third Parties Shape Violence. New York: New York University Press, 1998.

66. Cooney, Mark. Is Killing Wrong? A Study in Pure Sociology. Charlottesville: University of Virginia Press, 2009.
67. Copeland, M., Reynolds, K., & Burton, J. (2008) Social identity, status characteristics, and social identity: predictors of advice-seeking in a manufacturing facility. Asian Journal of Social Psychology, 11, 75-87.
68. Crenshaw, M., 2002. The Logic of Terrorism: Terrorist Behavior as a Product of Strategic Choice in Terrorism and Counterterrorism: Understanding the New Security Environment, Reading &Interpretations. Connecticut: Mc-Graw-Hill Companies.
69. Crossley N, Edwards G, Harries E, and Stevenson R. 2012. Covert social movement networks and the secrecy-efficiency trade-off: The UK suffragettes (1906–1914). *social networks* 34 (4): 634-644.
70. Crossley N, Stevenson R, Edwards G, and Harries E. 2010. Covert Social Movement Networks: A Report for the British Home Office.
71. Cunningham, D., Everton, S., & Murphy, P. (2016). *Understanding dark networks: A strategic framework for the use of social network analysis*. Rowman & Littlefield.
72. Davidsen, J., Ebel, H., and Bornholdt, S., "Emergence of a small world from local interactions: Modeling acquaintance networks." Physical Review Letters, 88(128701), 2002.
73. Davis R.H. 1981 "Social network analysis - an aid in conspiracy investigations." FBI Law Enforcement Bulletin 50(12): 11-19.
74. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., and Hu, W. (2012), "Author Gender Prediction in an Email Stream Using Neural Networks," Journal of Intelligent Learning Systems and Applications, 4, 169–175
75. Demiroz, Fatih.and Kapucu, Naim. 2012. Anatomy of a dark network: the case of the Turkish Ergenekon terrorist organization.
76. Desmarais, B. A., & Cranmer, S. J. (2013). Forecasting the locational dynamics of transnational terrorism: A network analytic approach. Security Informatics, 2, 1–13. Retrieved on 20.06.2020 from http://www.security-informatics.com/content/pdf/2190-8532-2-8.pdf. doi:10.1186/2190-8532-2-8
77. Dhillon I.S., Guan Y. and Kulis B. Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. IEEE

Trans. Pattern Anal. Mach. Intell.,29(11):1944–1957, 2007.
78. Diani, M., 1997. Social movements and social capital: A network perspective on movement outcomes. *Mobilization: An International Quarterly* 2 (2): 129-147.
79. Diesner, J., & Carley, K. M. (2004). Using network text analysis to detect the organizational structure of covert networks. Retrieved on 20.06.2020 from http://www.casos.cs.cmu.edu/publications/protected/2000-2004/2003-2004/diesner_2004_usingnetwork.pdf
80. Diesner, J., Frantz, T. L., and Carley, K. M. (2005), "Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different," Computational & Mathematical Organization Theory, 11, 201–228.
81. Diviák T., Dijkstra J.K. & Snijders T.A.B. Structure, multiplexity, and centrality in a corruption network: the Czech Rath affair. *Trends Organ Crim* **22,** 274–297 (2019). https://doi.org/10.1007/s12117-018-9334-y.
82. Djikstra E.W. A note on two problems in connexion with graphs. Numer. Math. **1**(1), 269-271 (1959).
83. Dong Yuxiao; Ke Qing; Rao Jun; Wu Bin; Predicting missing links via the local feature of common neighbors: 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), July 2011, Vol.2, pp.1038-1042.
84. Van DongenS.: Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht, 2000.
85. Doppa, Janardhan R., and Yu, Jun, Tadepalli, Prasad, and Getoor, Lise. (2009). Chance-Constrained Programs for Link Prediction. In Proceedings of Workshop on Analyzing Networks and Learning with Graphs at NIPS Conference.
86. Doreian, P., "On the evolution of the group and network structure." Social Networks 2: 235-252.1988.
87. Doreian, P., "Borgatti toppings on Doreian splits Reflections on regular equivalence." Social Networks 10(3): 273-287.1988.
88. Duijn, Paul A. C; Kashirin, Victor; Sloot, Peter A. 2014.The Relative Ineffectiveness of Criminal Network Disruption Scientific Reports, Vol.4.
89. Elias, P., A. Feinstein, and C. E. Shannon, 1956, IRE Trans.Inf. Theory IT-2, 117
90. Emirbayer M, Goodwin J (1994) Network Analysis, Culture, and the Problem of Agency. American Journal

of Sociology 99:1411–1454. Edges of Interest:10.1086/230450.
91. Enders, W., Su, X., 2007. Rational terrorists and optimal network structure. The Journal of Conflict Resolution 51 (1), 33–57.
92. Erickson BH. 1981. Secret Societies and Social Structure. *Social Forces* 60 (1): 188-210.
93. Everton SF. 2011. Network Topography, Key Players, and Terrorist Networks. *Connections* 32 (1):12.
94. Everton, S. F. (2012). *Disrupting dark networks* (Vol. 34). Cambridge University Press.
95. Everton, S. F., & Cunningham, D. (2013). Detecting significant changes in dark networks. Behavioral Sciences of Terrorism and Political Aggression, 5(2), 94-114.
96. Farley, J. (2003). Breaking Al Qaeda Cells: A Mathematical Analysis of Counterterrorism Operations (A Guide for Risk Assessment and Decision Making). Studies in Conflict & Terrorism, 26(6), 399-411.
97. Farley, J.D. (2007). Toward a mathematical theory of counterterrorism. The Proteus Monograph Series, 1, 1–72.
98. Farley, J. D. (2009). Two theoretical research questions concerning the structure of the perfect terrorist cell. In N. Memon, J. D. Farley, D. L. Hicks & T. Rosenorn (Eds.), Mathematical Methods in Counterterrorism (pp. 91–103). New York, NY: Springer-Verlag/Wien.
99. Faust, K. 1988. "Comparison of methods for positional analysis: Structural and general equivalences." Social Networks IO: 313-341.
100. Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications, and techniques: A survey. *Knowledge-based systems*, *70*, 301-323.
101. Feld SL. 1981. The focused organization of social ties. *American Journal of Sociology*: 1015-1035.
102. Feld SL. 1982. Social structural determinants of similarity among associates. *American sociological review*: 797-801.
103. Fellman, Philip Vos, and Wright, Roxana. 2006. "Modeling Terrorist Networks-Complex Systems at the Mid-Range." Working Paper. Available at: http://arxiv.org/ftp/arxiv/papers/1405/1405.6989.pdf.

104. FERC (2003), “Order Directing the Release of Information,” Available at http://www. mresearch.com/pdfs/139.pdf accessed: 2020-06-23.
105. FERC (2013), “Information Released in Enron Investigation,” Available at http://www.ferc. gov/industries/electric/indus-act/wec/enron/info-release.asp, accessed: 2020-06-23.
106. Fiedler, M., Algebraic connectivity of graphs, Czech. Math. J. 23, 298-305 (1973).
107. Flake G.W., Lawrence S., and Giles C.L., Efficient identification of web communities. In KDD ’00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, page 160. ACM, 2000.
108. Flick, U. (1992) ‘Triangulation revisited. The strategy of or alternative to the validation of qualitative data', *Journal for the Theory of Social Behavior*, 22 (2): 175–97.
109. Flick, U. (2018). Triangulation in data collection. In U. Flick *The sage handbook of qualitative data collection* (pp. 527-544). 55 City Road, London: SAGE Publications Ltd doi: 10.4135/9781526416070.n34
110. Fortunato, S., Community detection in graphs, Phys. Rep. 486, 75-174 (2010).
111. Freeman, L. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, 40, 35-41.
112. Freeman, L. (1979). Centrality in social networks: I.Conceptual clarification; *Social networks 1*, 215–239.
113. Freeman, L. C. (1996). Some antecedents of social network analysis. Connections, 19(1), 39-42.
114. Freeman Linton C., 2004. The development of social network analysis—with an emphasis on recent events. University of California, Irvine.
115. Freeman Linton C., 2011. The Development of Social Network Analysis—with an emphasis on recent events” Available at http://moreno.ss.uci.edu/91.pdf.
116. Freschi, Valerio. (2009). A Graph-based Semi-Supervised Algorithm forProtein Function Prediction from Interaction Maps. In Learning and IntelligentOptimization, Lecture Notes in Computer Science, 5851:249-258.
117. Friedkin, N.E., 1983. Horizons of observability and limits of informal control in organizations. Social Forces 62: 54-77.
118. Fu, Wenjie, and Song, Le, and Xing, Eric P. (2009). In Proc. of the 26th International Conference on Machine

Learning. Gerdes LM (2015a) Dark dimensions: classifying relationships among clandestine actors. In: Illuminating dark networks: the study of clandestine groups and organizations. Cambridge University Press, Cambridge, pp 19–38.
119. Geffre, J.L. (2007). A layered social and operational network analysis. MS thesis, AFIT/GOR/ ENS/07 – 07, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.
120. Getoor, Lise, and Friedman, Nir, and Koller, Dephne, and Taskar, Benjamin. (2002) Learning Probabilistic Models of Link structure. Journal of Machine Learning Research, 3:679-707.
121. Gill J, and Freeman JR. 2013. Dynamic elicited priors for updating covert networks. *Network Science* 1 (01): 68-94.
122. Gimenez-Salinas Framis A. 2011. Illegal networks or criminal organizations: Power, roles, and facilitators in four cocaine trafficking structures. Universidad Autónoma de Madrid.
123. D. G´omez, E. Gonz´alez-Arangena, C. Manuel, G. Owen, M. del Pozo, and J. Tejada, ″Centrality and power in social networks: a game-theoretic approach,″ Mathematical Social Sciences, vol. 46, no. 1, pp. 27−54, 2003.
124. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. (2014). Generative Adversarial Networks.
125. Granovetter, Mark. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 73 (6): 1360-1380.
126. Hamill, J.T. (2006). Analysis of layered social networks (Ph.D. dissertation, AFIT/DS/ENS/06– 03, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH).
127. Hanneman RA, and Riddle M. 2005. Introduction to Social Network Methods. Riverside, CA: University of California, Riverside.
128. Hanneke S, Fu W, Xing EP: Discrete Temporal Models of Social Networks. Electron. J. Stat, 2010, 4: 585–605. 10.1214/09-EJS548
129. Hardin, J. S., Sarkis, G., & Urc, P. C. (2015). Network analysis with the Enron email corpus. *Journal of Statistics Education*, *23*(2).
130. Harris-Hogan, S., 2012. Anatomy of a terrorist cell: a study of the network uncovered in Sydney in 2005.

Behavioral Sciences of Terrorism and Political Aggression,5 (2): 137-154.
131. Hasan, Mohammad Al; Zaki, Mohammed J; Aggarwal, Charu C (Editor), A Survey of Link Prediction in Social Networks, 1 Boston, MA: Springer US 2011.
132. Hasan, Mohammad A., and Chaoji, Vineet, and Salem, Saeed, and Zaki, Mohammed. (2006) Link Prediction using Supervised Learning. In Proceedings of SDM Workshop of Link Analysis, Counterterrorism, and Security.
133. Hauck, R.V; Atabakhsb, H; Ongvasith, P; Gupta, H; Hsinchun Chen, Using Coplink to analyze criminal justice data. Computer, March 2002, Vol.35(3), pp.30-37.
134. Helfstein S, and Wright D. 2011. Covert or Convenient? Evolution of Terror Attack Networks. *Journal of Conflict Resolution* 55 (5): 785-813.
135. Herbranson, T.J. (2007). Isolating key players in clandestine networks (MS thesis, AFIT/GOR/ ENS/07 – 11, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH).
136. Howlett, J.B.1980 "Analytical, investigative techniques: Tools for complex criminal investigations." Police Chief 47(12): 42-45.
137. Huang, Zan, and Li, Xin, and Chen Hsinchun. (2005) Link Prediction approach to collaborative filtering. Proceedings of the fifth ACM/IEEE Joint Conference on Digital Libraries.
138. Van der Hulst RC (2009) Introduction to social network analysis (SNA) as an investigative tool. Trends Organized Crime 12(2):101–121.
139. Jeong, H., Mason, S. P. & Barabási, A. L. Lethality and centrality in protein networks. Nature 411, 41–42 (2001).
140. Jones, N.P., Dittmann, W.L., Wu, J. *et al.* A mixed-methods social network analysis of a cross-border drug network: the Fernando Sanchez organization (FSO). *Trends Organ Crim* **23,** 154–182 (2020). https://doi-org.ezp3.lib.umn.edu/10.1007/s12117-018-9352-9.
141. Jin, E. M., Girvan, M., and Newman, M. E. J. "The structure of growing social networks."Physical Review Letters E, 64(046132), 2001.
142. Kashima, Hisashi, and Abe, Naoke. (2006) A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction. ICDM '06:

Proceedings of the Sixth IEEE International Conference on Data Mining. 340-349.
143. Katryn Oliver. 2014. Covert networks: structures, processes and types. www.socialsciences.manchester.ac.uk/.../covertnetworks/.../working_paper.
144. Katz, Leo. (1953) A new status index derived from sociometric analysis. Psychometrika, 18(1):39-43.
145. Keegan B, Ahmed MA, Williams D, Srivastava J, and Contractor N. 2010. Dark gold: Statistical properties of clandestine networks in massively multiplayer online games.
146. Keller, J., Desouza, K., & Lin, Y. (2010). Dismantling terrorist networks: Evaluating strategic options using agent-based modeling. Technological Forecasting & Social Change, 77(7), 1014-1036.
147. Kenney, M., 2007. From Pablo to Osama: Trafficking and Terrorist Networks, Government Bureaucracies, and Competitive Adaptation. University Park, PA. The Pennsylvania State University Press
148. Klerks, P., 2001. The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. Connections24: 53–65. CiteSeerX: 10.1.1.129.4720.
149. Koschade, S., 2006. A social network analysis of Jemaah Islamiyah. Studies in Conflict and Terrorism 29, 559–575.
150. Kossinets, G., & Watts, D.J., 2006. The empirical analysis of an evolving social network. Science, 311, 88–90.
151. Kossinets, G. (2006). Effects of missing data in social networks. Social Networks, 28, 247 –268.
152. Kossinets, G. (2008, February 2). Effects of missing data in social networks. E-Print, arXiv: cond-mat/0306335v2, 1– 31.
153. Karypis G. and Kumar V. A fast and high-quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing, 20, 1999.
154. Krautz, H., B. Selman, and M. Shah. ReferralWeb: Combining social networks and collaborative filtering. Communications of the ACM, 40(3):63{65, March 1997.
155. Kennedy, K.T. (2009). Synthesis, interdiction, and protection of layered networks (Ph.D. dissertation, AFIT/DS/ENS/09– 01, Air Force Institute of

Technology, Department of Operational Sciences, Wright-Patterson AFB, OH).
156. Kernighan, B. W., and Lin, S., An efficient heuristic procedure for partitioning graphs, Bell System Technical Journal 49, 291-307 (1970).
157. Kirby, A., 2007. The London bombers as self-starters: A case study in indigenous radicalization and the emergence of autonomous cliques. Studies in Conflict & Terrorism, 30(5): 415-428.
158. Kleinberg, Jon M. (2000). Navigation in a small world. Nature 406, (845).
159. Knoke, D., and Burleigh, F., 1989. "Collective Action in National Policy Domains: Constraints, Cleavages, and Policy Outcomes." Research in Political Sociology 4:187-208.
160. Knoke, D., and Laumann E, O., 1982. "The Social Organization of National Policy Domains: An Exploration of Some Structural Hypotheses." Pp. 255-70 in Social Structure and NetworkAnalysis, edited by Peter V. Marsden and Nan Lin. Beverly Hills: Sage.
161. Knoke, D., PappiF U., BroadbentJ., Kaufman NJ., and TsujinakaY. 1991. "Policy Networks and Influence Reputations in the U.S., German, and Japanese National Labor Policy Domains." Paper presented to Second European Social Network Conference, 20-22 June, Paris.
162. Knoke, D. (1993a). Networks of Elite Structure and Decision Making. Sociological Methods & Research, 22(1), 23–45.
163. Knoke, D. (1993b). Networks as political glue: explaining public policy-making. In W. J. Wilson (Ed.), *American Sociological Association Presidential Series: Sociology and the public agenda* (pp. 164-184). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781483325484.n9.
164. Knoke, Yang, Yang, Song, & Knoke, David. (2008). *Social network analysis* (2nd ed., Quantitative applications in the social sciences; 154). Los Angeles: Sage Publications.
165. Knoke, D. Emerging Trends in Social Network Analysis of Terrorism and Counterterrorism. John Wiley & Sons (2015).
166. Krebs, V., 2002. Uncloaking terrorist networks. First Monday, http://pear.accc.uic.edu/ojs/index.php/fm/article/view/941/863 (accessed06.20.20).

167. Krebs, V. Mapping networks of terrorist cells. Connections, 24(3):43-52, Winter 2002.
168. Kunegis, Jerome, and Lommatzsch, Andreas. (2009) Learning Spectral Graph Transformations for Link Prediction. In Proceedings of the International Conference on Machine Learning, pp 561-568.
169. Von Lampe K, Ole Johansen P (2004) Organized crime and trust: on the conceptualization and empirical relevance of trust in the context of criminal networks. Global Crime 6(2):159–184.
170. Von Lampe, K., 2009. The study of organized crime: an assessment of the state of affairs. Organized crime: norms, markets, regulation, and research. Oslo, UNIPUB.
171. Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. Physical Review E, 78(4), 046110.
172. Lang K. and Rao S., A flow-based method for improving the expansion or conductance of graph cuts. Lecture notes in computer science, pages 325–337, 2004.
173. Lauchs M, Keast R, and Yousefpour N. 2011. Corrupt police networks: uncovering hidden relationship patterns, functions, and roles. *Policing & Society* 21 (1): 110-127.
174. Laumann, Edward O., and Peter V. Marsden. 1979. "The Analysis of Oppositional Structures in Political Elites: Identifying Collective Actors." American Sociological Review 44:713-32
175. Laumann, E.O., Marsden, P.V., & Prensky, D. (1983). The boundary specification problem in network analysis. In R.A. Burt (Ed.), Applied network analysis: A methodological introduction (pp. 18 –34). London: Sage
176. Laumann, Edward O., David Knoke, and Yong-Hak Kim. 1985. "An Organizational Approach to State Policy Formation: A Comparative Study of Energy and Health Domains." American Sociological Review 50:1-19.
177. Laumann Edward O., and David Knoke, 1987. *The organizational state: social choice in national policy domains*. Madison: University of Wisconsin Press.
178. Laumann, Edward O., and David Knoke. 1987. The Organizational State: A Perspective on National Energy and Health Domains. Madison: University of Wisconsin Press.
179. Laumann, Edward O., Peter V. Marsden, and David Prensky. 1989. The boundary specification problem in

network analysis. Research Methods In Social Network Analysis 61:87
180. Leinart, J.A. (2008). Characterizing and detecting unrevealed elements of network systems (Ph.D. dissertation, AFIT/DS/ENS/08– 01W, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH).
181. Leskovec, Jure, and Kleinberg, Jon M, and Faloutsos, Christos. (2005). Graphs over time: densification laws, shrinking diameters, and possible explanations. KDD '05: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.
182. Li, Xin, Chen Hsinchun. (2009). Recommendation as link prediction: a graph kernel-based machine learning approach. Proceedings of the ninth ACM/IEEE Joint Conference on Digital Libraries.
183. Li, F. et al. A Clustering-based Link Prediction Method in Social Networks. Procedia Computer Science 29, 432–442 (2014).
184. Liu, Yan, and Kou, Zhenzhen. (2007). Predicting who rated what in large-scale datasets. SIGKDD Exploration Newsletter, 9 (2).
185. Liben-Nowell, David, and Kleinberg, Jon. (2007). The Link Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology, 58(7):1019-1031.
186. Lindelof, R., Borm, P., Hamers, H., 2009a. The influence of secrecy on the communication structure of covert networks. Social Networks 31,126–137.
187. Lindelauf R, Borm P, Hamers H 2009b. Understanding Terrorist Network Topologies and Their Resilience against Disruption. Discussion Paper. No. 2009-85. Tilburg University, Center for Economic Research.
188. Lindelauf, R., Hamers, Herbert, Borm, Peter, & Econometrics Operations Research. (2011). Design and analysis of covert networks, affiliations, and projects. 2011.
189. Luce, R. D., and A. D. Perry, 1949, Psychometrika 14(2), 95.
190. Magalingam, P., Davis, S., & Rao, A. (2015). Using the shortest path to discover the criminal community. 15(C), 1-17.
191. Mahesh, S., Mahesh, T. R., & Vinayababu, M. (2010). Using data mining techniques for detecting terror-related

activities on the Web. Journal of Theoretical & Applied Information Technology, 16, 99–104.
192. Malm, A., Kinney, J., & Pollard, N. (2008). Social Network and Distance Correlates of Criminal Associates Involved in Illicit Drug Production. Security Journal, 21(1-2), 77-94.
193. Malm A, and Bichler G. 2011. Networks of collaborating criminals: Assessing the structural vulnerability of drug markets. *Journal of Research in Crime and Delinquency* 48 (2): 271-297.
194. Malin, Bradley, and Airoldi, Edoardo, and Carley, Kathleen M. (2005). A Network Analysis Model for Disambiguation of Names in Lists. In Journal of Computational and Mathematical Organization Theory, 11(2):119-139.
195. Mandell, Myrna P., and Robyn Keast. 2008. Evaluating the effectiveness of inter-organizational relations through networks: Developing a framework for revised performance measures. Public Management Review 10:715–31.
196. Marsden, P.V. (1990). Network data and measurement. Annual Review of Sociology, 16, 435– 463.
197. Marsden, Peter V., 2005. Recent developments in network measurement. Models and Methods in Social Network Analysis 8:30.
198. Martin, S., Sewani, A., Nelson, B., Chen, K., and Joseph, A. D. (2005), "Analyzing Behavioral Features for Email Classification," in Berkeley, CA: University of California at Berkeley.
199. Maslov, S. & Sneppen, K. Specificity, and stability in topology of protein networks. Science 296, 910–913 (2002).
200. Mathison, S. (1988) 'Why triangulate?', *Educational Researcher*, 17(2): 13–17.
201. McGloin, Jean Marie; Piquero, Alex R. On the relationship between co-offending network redundancy and offending versatility. Journal of Research in Crime and Delinquency, February 2010, Vol.47(1), pp.63-90.
202. McLean, B. and Elkind, P. (2013), The Smartest Guys in the Room, Portfolio Trade.
203. McPherson, M., Smith-Lovin, L., & Cook, J.M., 2001. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27, 415–444.
204. Mehra A, Kilduff M, Brass DJ (2001) The social networks of high and low self-monitors: implications for

workplace performance. Adm Sci Q 46:121–146. Edges of Interest:10.2307/2667127.
205. Meila, M., and Shi, J. A random walks view of spectral segmentation. AI and Statistics (AISTATS), 2001, 2001.
206. Memon, B. (2012). Identifying Important Nodes in Weighted Covert Networks Using Generalized Centrality Measures. *2012 European Intelligence and Security Informatics Conference,* 131-140.
207. Memon N, Hicks DL, Larsen HL. 2007. Understanding the structure of terrorist networks. Int J Bus Intell Data Min 2(4):401–425.
208. Memon N, Hicks DL, Harkiolakis N, and Rajput AQK. 2008. Small world terrorist networks: a preliminary investigation. In *Applications and Innovations in Intelligent Systems XV*, 339-344. Springer.
209. Mickolus EF, Sandler T, Murdock JM, Flemming PA: International Terrorism: Attributes of Terrorist Events (ITERATE), 1968–2007. Vinyard Software (Dunn Loring, 2008) (Dunn Loring, 2008)
210. Milward HB, and Raab J. 2006. Dark networks as organizational problems: Elements of a theory. *International Public Management Journal* 9 (3): 333-360.
211. Moore, C. and Mertens, S., *The nature of computation*, Oxford University Press, Oxford (2011).
212. Morris, J. F., & Deckro, R. F. (2013). SNA data difficulties with dark networks. *Behavioral Sciences of Terrorism and Political Aggression*, *5*(2), 70-93.
213. Morselli, C., Giguère, C., Petit, K., 2007. The efficiency/security trade-off in criminal networks. Social Networks 29, 143–153.
214. Morselli, C., 2009. Inside Criminal Networks. Springer, New York.
215. Morselli, C.,2014.Crime and networks: Criminology and justice studies. New York: Routledge, 2014, 288.
216. Nallapati, Ramesh, and Ahmed, Amr, and Xing, Eric P., and Cohen, William W. (2008). Joint Latent Topic Models for Text and Citations. In Proc. of The Fourteen ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining.
217. Natarajan, M., 2000. Understanding the structure of a drug trafficking organization: a conversational analysis. *Crime Prevention Studies* 11: 273-298.
218. Natarajan, M., 2006. Understanding the structure of a large heroin distribution network: a quantitative analysis

of qualitative data. *Journal of Quantitative Criminology* 22 (2): 171-192.
219. Nefedov, N. (2011). Multiple-membership communities' detection in mobile networks. Proceedings of the International Conference on Web Intelligence, Mining and Semantics, 1-6.
220. Newman, M. E. J. (2001a) Clustering and preferential attachment in growing networks. Physical Review Letters E, 64(025102),
221. Newman, M. E. J. (2001b)The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98:404-409.
222. Newman, M. (2001c). Scientific collaboration networks. II.Shortest paths, weighted networks, and centrality. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *64*(1 Pt 2), 016132.
223. Newman, M. E., J. The structure and function of networks. Computer Physics Communications, 147:40-45, 2002.
224. Newman, M. E., J. The structure and function of complex networks. SIAM Review, 45:167-256, 2003.
225. Newman, M. (2004). Analysis of weighted networks. Physical Review E, 70(5), 056131.
226. Newman, M. E. J. and Girvan, M., Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113 (2004).
227. Newman, M. E. J., Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74, 036104 (2006).
228. Newman, M. E. J., Modularity, and community structure in networks, Proc. Natl. Acad. Sci. USA 103, 8577-8582 (2006a).
229. Newman, M. E. J., Random graphs with clustering, Phys. Rev. Lett. 103, 058701, (2009).
230. Newman, M. E. J. Forrest, S., and Balthrop, J., Email networks, and the spread of computer viruses, Phys. Rev. E 66, 035101 (2002).
231. Newman, M. E., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. Proceedings of the National Academy of Sciences, 99(suppl 1), 2566-2572.
232. Networks: An Introduction by MEJ Newman: Oxford, UK: Oxford University Press. 720 pp., (2011).
233. Nowell, B., T. Steelman, A. Velez, and S. Godette. 2016. Operationalizing performance measures in networked

settings: lessons from large-scale wildfires in the United States. In Social Networks and Disasters, ed. E. C. Jones and A. J. Faas. Atlanta, GA: Elsevier

234. Nowell, B., Velez, A., Hano, M., Sudweeks, J., Albrecht, K., & Steelman, T. (2018). Studying Networks in Complex Problem Domains: Advancing Methods in Boundary Specification. *Perspectives on Public Management and Governance, 1*(4), 273-282.
235. O'Toole, Laurence J. 1997, Treating networks seriously: Practical and research-based agendas in public administration. *Public Administration Review* 57: 45-52.
236. Oliver K, Crossley N, Everett MG, Edwards G, Koskinen J (2014) Covert networks: structures, processes, and types. The Mitchell Center for Social Network Analysis working paper.
237. Ovelgonne, M., Chanhyun Kang, Sawant, & Subrahmanian. (2012). Covertness Centrality in Networks. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 863-870.
238. P. Raghavan. Social networks: From the web to the enterprise. IEEE Internet Computing, 6(1):91{94, January/February 2002.
239. Papachristos AV, Smith C (2014) The embedded and multiplex nature of al Capone. In: Crime and networks. Routledge, New York, pp 97–115.
240. Parthasarathy, S; Ruan, Y; Satuluri, V; Aggarwal, Charu C (Editor): Community Discovery in Social Networks: Applications, Methods and Emerging Trends 1 Boston, MA: Springer US 2011.
241. Pathak N. Srivastava J. 2006, Automatic extraction of concealed relations from email logs. (CSE Dept., University of Minnesota, Twin. Cities).
242. PattipatiK. R. WillettP.K., Allanach, J., Tu H., and Singh S., "Hidden Markov Models and Bayesian Networks for Counter-terrorism," R. Popp and J. Yen (editors) Emergent Information Technologies and Enabling Policies for Counter-Terrorism, Wiley-IEEE Press, May 2006, pp. 27-50.
243. Peterson, K., Hohensee, M., and Xia, F. (2011), "Email Formality in the Workplace: A Case Study on the Enron Corpus," in Proceedings of the Workshop on Languages in Social Media, Stroudsburg, PA, USA: Association for Computational Linguistics, LSM '11, pp. 86–95.

244. Petraeus, General David. 2007. *U.S. Army/ Marine Counterinsurgency Field Manual*. Old Saybrook, CT: Konecky & Konecky.
245. Petroff, V. B., Bond, J. H., & Bond, D. H. (2013). Using hidden Markov models to predict terror before it hits (again). In V. S. Subrahmanian (Ed.), Handbook of computational approaches to counterterrorism (pp. 163–180). New York, NY: Springer-Verlag/Wien.
246. Pedazhur, Ami. And Arie Prelinger. 2006. "The Changing Nature of Suicide Attacks: A Social Network Perspective." *Social Forces* 84 (4): 1987-2008.
247. Popescul, Alexandrin, and Ungar, Lyle H. (2003). Statistical Relational Learning for Link Prediction. In Proceedings of Workshop on LearningStatistical Models from Relational Data at IJCAI Conference.
248. Popescul, Alexandrin, and Ungar, Lyle H. (2003). Structural LogisticRegression for Link Analysis. In Proceedings of Workshop on Multi-Relational Data Mining at KDD Conference.
249. Pothen, A., Simon, H., and Liou; K-P Partitioning sparse matrices with eigenvectors of graphs, SIAM J. Matrix Anal. Appl. 11, 430-452 (1990).
250. Raab, J., Milward, H., 2003. Dark networks as problems. Journal of Public Administration Research and Theory 13 (4), 413–439.
251. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D., Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101, 2658-2663 (2004).
252. Rapoport, D. C. (2001). The fourth wave: September 11 in the history of world terrorism.Current History, 100, 419–424. Retrieved on 06.22.2020 fromhttps://doi-org.ezp3.lib.umn.edu/10.1525/curh.2001.100.650.419
253. Rapoport, D. (2015). It Is Waves, Not Strains. Terrorism and Political Violence, 28(2), 1-8.
254. Reed, B.J. (2006). Formalizing the informal: A network analysis of an insurgency (Ph.D. dissertation, University of Maryland, Department of Sociology, College Park, MD).
255. Reed, B.J., 2007. A social network approach to understanding an insurgency. DTIC Document.
256. Reiss, Albert J., 1986. Co-offender influences on criminal careers. In A. Blumstein, J. Cohen, J. Roth, & C.A. Visher (Eds.), Criminal Careers and "Career

Criminals" (Vol. 2, pp. 121–160). Washington, DC: National Academy Press.
257. Reiss, Albert J. Co-offending and criminal careers. Crime and Justice, 1 January 1988, Vol.10, pp.117-170.
258. Reiss, Albert J.; Farrington, David P. Advancing knowledge about co-offending: Results from a prospective longitudinal survey of London males. The Journal of Criminal Law and Criminology (1973-), 1 July 1991, Vol.82(2), pp.360-39.
259. Renfro, I.R. (2001). Modeling and analysis of social networks (Ph.D. dissertation, AFIT/GOR/ ENS/07 – 11, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH).
260. Ressler, S. (2006). Social network analysis as an approach to combat terrorism: Past, present, and future research. Homeland Security Affairs, II, 1 – 10.
261. Roberts, N., & Everton, S. F. (2011). Strategies for combating dark networks.
262. Roberts, N., & Everton, S. (2016). Monitoring and disrupting dark networks: A bias toward the center and what it cost us. In *Eradicating Terrorism from the Middle East* (pp. 29-42). Springer, Cham.
263. Robins, Gary, "Understanding individual behaviors within covert networks: the interplay of individual qualities, psychological predispositions, and network effects." Trends in Organized Crime, June 2009, Vol.12(2), p.166(22).
264. Robins G, Kashima Y (2008) Social psychology and social networks. Asian J Soc Psychol 11:1–12. Edges of Interest:10.1111/j.1467-839X.2007.00240. x.
265. Rodriguez, J., 2009. Weakness and strengths of terrorist networks: The Madrid March 11 Attacks. Paper presented at the Annual Meeting of the American Sociological Association. Available at http://www.allacademic.com/meta/p243052 index.html.
266. Rombach, P., Porter, M. A., Fowler, J. H., & Mucha, P. J. (2017). Core-periphery structure in networks (revisited). *SIAM Review*, *59*(3), 619-646.
267. Ross, J. (1993). Structural Causes of Oppositional Political Terrorism: Towards a Causal Model. Journal of Peace Research, 30(3), 317-329.
268. Russell, G., 2007. Hacking social networks: Examining the viability of using computer network attack against

social networks; Naval Postgraduate School Monterey, California.

269. Sade D. S., ″Sociometrics of Macaca mulatta iii: n-path centrality in grooming networks,″ Social Networks, vol. 11, no. 3, pp. 273–292, 1989.
270. Sageman, M., 2004. *Understanding Terror Networks*. University of Pennsylvania Press, Philadelphia.
271. Sageman, M., 2008. *Leaderless Jihad. Terror Networks in the Twenty-First Century*. University of Pennsylvania Press, Philadelphia.
272. Santos, R. B. (2016). *Crime analysis with crime mapping*. Sage publications.
273. Sandler T., and Siqueira K., ‘‘Games and Terrorism: Recent Developments,’’ Simulation & Gaming, Sep 2003; vol. 34: pp. 319 - 337.
274. Sarna, G., & Bhatia, M. (2018). Identification of Suspicious Patterns in Social Network using Zipf's Law. 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 957-962
275. Sarukkai, Ramesh R. (2000). Link Prediction and Path Analysis using Markov Chain. WWW ’2000: Proceedings of the Ninth World Wide Web Conference, 377-386.
276. Satuluri V. and Parthasarathy S. Scalable graph clustering using stochastic flows: Applications to community discovery. In KDD’09, pages 737–746, New York, NY, USA, 2009. ACM.
277. Satuluri, V. Parthasarathy, S., and Ucar.D. Markov Clustering of Protein Interaction Networks with improved Balance and Scalability. In Proceedings of the ACM Conference on Bioinformatics and Computational Biology, 2010.
278. Schaeffer, S. E., Graph clustering, Comp. Sci. Rev.1, 27-64 (2007).
279. Schneider, Volker, and Raymond Werle, 1991. Policy networks in the German telecommunications domain. In *Policy networks: Empirical evidence and theoretical considerations*, edited by B.Marin and R. Mayntz, 97-137. Boulder/Frankfurt: Westview Press/ Campus Verlag.
280. Schwartz DM, Rouselle T. 2009. Using social network analysis to target criminal networks. Trends Organized Crime 12(2):188–207.

281. Seder, J.S. (2007). Examining clandestine social networks for the presence of non-random structure (MS thesis, AFIT/GOR/ENS/07 – 24, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH).
282. Senechal de la Roche, R. 1996. "Collective Violence as Social Control." Sociological Forum.
283. Sharma, A., Feng, X., Singhal, K., Kuang, R., & Srivastava, J. (2015). Predicting Small Group Accretion in Social Networks: A topology-based incremental approach.
284. Shetty, J. and Adibi, J. (2004), "The Enron email dataset database schema and brief statistical report," Tech. rep., University of Southern California—Information Sciences Institute
285. Shi, J., and Maliki. Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8):888–905, 2000.
286. Sliva, A., Subrahmanian, V., Martinez, G., & Simari. (2008). The soma terror organization portal (Stop): Social network and analytic tools for real-time terror group analysis. Social Computing, Behavioral Modeling, and Prediction, 2008, 9-18. Smith CM, Papachristos AV (2016) Trust thy crooked neighbor Multiplexity in Chicago organized crime networks. American Sociological Rev 81(4):617–643.
287. Sliva, A., Subrahmanian, J., Mannes, V., & Shakarian. (2011). A computationally-enabled analysis of Lashkar-e-Taiba attacks in Jammu & Kashmir. Proceedings - 2011 European Intelligence and Security Informatics Conference, EISIC 2011, 224-229.
288. Smith, Justin J, Santos, Rachel B, & Santos, Roberto G. (2018). Evidence-Based Policing and the Stratified Integration of Crime Analysis in Police Agencies: National Survey Results. *Policing: A Journal of Policy and Practice, 12*(3), 303-315
289. Song, Han H., and Cho Tae W., and Dave, Vacha, and Zhang, Yin, and Qiu, Lili. (2009). Scalable proximity Estimation and Link Prediction in Online Social Networks, IMC '09: In Proceedings of the Internet Measurement Conference.
290. Sparrow, M., 1991. The application of network analysis to criminal intelligence. Social Networks 13, 251–274.
291. Spielman D.A. and Srivastava N. Graph sparsification by effective resistances. In STOC '08: Proceedings of the

40th annual ACM symposium on Theory of computing, pages 563–568, New York, NY, USA, 2008. ACM.
292. Spielman D.A. and Teng S.H. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In Proceedings of the thirty-sixth annual ACM symposium on Theory of computing pages 81–90. ACM New York, NY, USA, 2004.
293. Spiliopoulou, M., & Aggarwal, C. (2011). Evolution in Social Networks: A Survey. In Social Network Data Analytics (1st ed., pp. 149-175). Boston, MA: Springer US.
294. Sterling, S.E. (2004). Aggregation techniques to characterize social networks (MS thesis, AFIT/ GOR/ENS/04 – 12, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH).
295. Stohl, M., 2008. Networks, terrorists, and criminals. Crime, Law, and Social Change 50, 59–72.
296. Stumpf, M. P. H. et al. Estimating the size of the human interactome. Proceedings of the National Academy of Sciences of the United States of America 105, 6959–6964 (2008).
297. Sun Duo-Yong Guo, Liu, Shu-Quan Li, Xiao-Peng Jiang Multi-relational Network Analysis for Covert Organizations.
298. Tasker, Benjamin, and Wong, Ming F., and Abbeel, Pieter, and Koller, Daphne. (2003). Link Prediction in Relational Data. NIPS ’03: In Proceedings of Neural Information Processing Systems.
299. Tayebi, M., Jamali, M., Ester, M., Glässer, U., & Frank, R. (2011). CrimeWalker: A recommendation model for suspect investigation. Proceedings of the Fifth ACM Conference on Recommender Systems, 173-180.
300. Tayebi, M. A., Ester, M., Glässer, U., & Brantingham, P. L. (2014, August). Crimetracer: Activity space-based crime location prediction. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)* (pp. 472-480). IEEE.
301. Tayebi, Mohammad A; Glässer, Uwe; Alhajj, Reda (Editor); Glässer, Uwe (Editor): Social network analysis in predictive policing: Concepts, Models, and Methods. Cham: Springer International Publishing 2016, Lecture Notes in Social Networks.

302. Teng S.H. Coarsening, sampling, and smoothing: Elements of the multilevel method. Algorithms for Parallel Processing, 105:247–276, 1999.
303. Thelwall, M., 2008. Social networks, gender, and friending: An analysis of MySpace member profiles. Journal of the American Society for Information Science and Technology, 59, 1321–1330.
304. Toth N, Guly L., Legendi RO, Duijn P, Sloot PM, and Kampis G. 2013. The importance of centralities in dark network value chains. *The European Physical Journal Special Topics* 222 (6): 1413-1439.
305. Tsvetovat, M. & Carley, 2003. K. M. Bouncing back: Recovery mechanisms of covert networks.NAACSOS Conference 2003.
306. Tsvetovat, M., Carley, K., 2005. Structural Knowledge and Success of Anti-Terrorist Activity. Journal of Social Structure 6 (2).
307. Tucker, J. The Therapeutic Corporation. New York: Oxford University Press, 1999.
308. Tucker, James. "The Geometry of Suicide Law." International Journal of Law, Crime, and Justice, vol. 43, n. 3, pp. 342-365, 2015.
309. Turrini, Alex, Daniela Cristofoli, Francesca Frosini, and Greta Nasi. 2010. Networking literature about determinants of network effectiveness. Public Administration 88:528–50
310. Tylenda, Tomasz, and Angelova, Ralitsa, and Bahadur, Srikanta. (2009). Towards time-aware link prediction in evolving social networks. SNA-KDD '09: Proceedings of the Third Workshop on Social Network Mining and Analysis.
311. Vinayagam, A., Stelzl, U. & Foulle, R. A directed protein interaction network for investigating intracellular signal transduction. Science Signaling 4, rs8 (2011).
312. Wang, Chao, and Satuluri, Venu, and Parthasarathy, Srinivasan. (2007). Local Probabilistic Models for Link Prediction. ICDM'07: In Proceedings of International Conference on Data Mining.
313. Wasserman, S., & Faust, K. (1994), *Social network analysis: Methods and applications.* New York: Cambridge University Press.
314. Wasserman, S., & Pattison, P. (1996), Logit models and logistic regressions for social networks: An introduction to Markov graphs and p*. *Psychometrika, 61(3),* 401-425.

315. Watts, D, and Stogatz, S. (1998). Small world. Nature, 393:440-442.
316. Watts, D. J., 1999. Networks, dynamics, and the small-world phenomenon. American Journal of Sociology 13(2): 493-527.
317. Weber, Edward P., and Anne M. Khademian. 2008. Wicked problems, knowledge challenges, and collaborative capacity builders in network settings. Public Administration Review 68:334–49.
318. Weinstein, C., Campbell, W., Delaney, B., & O'Leary, G. (2009). Modeling and detection techniques for counter-terror social network analysis and intent recognition. Institute of Electrical and Electronics Engineers. Retrieved from dspace.mit.edu/openaccess-disseminate/1721.1/71803.
319. Weiss, Gary M. (2004) Mining with rarity: a unifying framework, In SIGKDD Explorations Newsletter, 6(1):7-19.
320. Williams P., 2001 Transnational Crime Networks. In: Arquilla J, Ronfeldt D (eds) Networks and NETWARS: the future of terror, crime, and militancy. RAND, Santa Monica.
321. Wu M., Knoke D. (2017) Dark Networks: The Terror–Crime Nexus. In: Koops J., Biermann R. (eds) Palgrave Handbook of Inter-Organizational Relations in World Politics. Palgrave Macmillan, London.
322. Varese, Federico. 2006. The structure of a criminal network examined: The Russian Mafia in Rome. *Oxford Legal Studies Research Paper* No.21.
323. Wang, Chao, and Satuluri, Venu, and Parthasarathy, Srinivasan. (2007). Local Probabilistic Models for Link Prediction. ICDM'07: In Proceedings of International Conference on Data Mining.
324. Wang, H., Tang, M., Park, Y., and Priebe, C. E. (2014), "Locality statistics for anomaly detection in time series of graphs," IEEE Trans. Signal Process., 62, 703–7
325. www://https://en.wikipedia.org/wiki/Heuristic_(computer_science) accessed on 6 June 2020.
326. Xu, Zhao, and Tresp, Volker, and Yu, Shipeng, and Yu, Kai. (2005). Nonparametric Relational Learning for Social Network Analysis. SNA-KDD '08: In proceedings of the Second Workshop on Social Network Mining and Analysis.
327. Xu JJ, Chen H. 2005 CrimeNet explorer: A framework for criminal network knowledge discovery. ACM Trans Inf Syst 23(2):201–226.

328. Yu, H. Y. et al. High-quality binary protein interaction map of the yeast interactome network. Science 322, 104–110 (2008).
329. Zachary, W. W. (1977). "An Information Flow Model for Conflict and Fission in Small Groups." Journal of Anthropological Research. 33 (4): 452–473. JSTOR 3629752.
330. Zhang, Q. M., Shang, M. S. & Lü, L. Y. Similarity-based classification in partially labeled networks. International Journal of Modern Physics C 21, 813–824 (2010).
331. Zhang, Y., Zheng, Z. & Lyu, M. R. An online performance prediction framework for service-oriented systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems 44, 1169–1181 (2014).
332. Zhang, Z. K., Zhou, T. & Zhang, Y. C. Tag-aware recommender systems: a state-of-the-art survey. Journal of Computer Science and Technology 26, 767–777 (2011).
333. Zhou, D., Manavoglu, E., Li, J., Giles, C. L., and Zha, H. (2006), "Probabilistic Models for Discovering e-Communities," in Proceedings of the 15th International Conference on World Wide Web, New York, NY, USA: ACM, WWW '06, pp. 173–182.
334. Zhou, Y., Goldberg, M., Magdon-Ismail, M., and Wallace, W. A. (2007), "Social Communication Networks for Early Warning in Disasters. Strategies for Cleaning Organizational Emails with an Application to Enron Email Dataset," 5th Conf. of North American Association for Computational Social and Organizational Science (NAACSOS 07), Emory - Atlanta, Georgia.
335. Zhou, Tao, Lü, Linyuan, Zhang, Yi-Cheng: Predicting missing links via local information. The European Physical Journal B, 2009, Vol.71(4), pp.623-630.
336. Zhu, Jianhan, and Hong, Jun, and Hughes G. (2002). Using Markov models for web site link prediction. HYPERTEXT'02: Proceedings of the Thirteenth ACM Conference on Hypertext and Hypermedia.
337. Zimmerman, K., 2013. The Al Qaeda Network: A New Framework for Defining the Enemy. A Report by AEI's Critical Threats Project.