Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 10-018

Modeling Trust in Online Social Networks to Improve Adolescent Health Behavior

Young Ae Kim, Marla E. Eisenberg, Muhammad Aurangzeb Ahmad,
and Jaideep Srivastava

August 18, 2010

Modeling Trust in Online Social Networks to Improve Adolescent Health Behaviors

Young Ae Kim[a], **Marla E. Eisenberg**[b], **Muhammad A. Ahmad**[a], **Jaideep Srivastava**[a]

[a] Department of Computer Science and Engineering, [b] Division of Adolescent Health and Medicine, Department of Pediatrics

University of Minnesota

SUMMARY

A majority of adolescents use social networking sites, which have become a major communication channel for young people. While research has begun to examine representations of health risk behavior on-line and the efficacy of brief interventions to curb such behaviors, this work is in its infancy. The present report describes a new model for research using on-line social networks to improve adolescent health behaviors. After introducing key constructs (Section 1), we describe the theoretical framework of social influence, and the social network models on which this work is based, in Section 2. In Section 3, we propose the research framework for a computational trust model which covers data collection, computational web data analysis techniques and trust prediction models. Based on this framework, we describe applications of the trust network in order to maximize the impact of trust based social influence among adolescents on their health behavior in Section 4. Finally, in Section 5, we discuss several technical and ethics challenges inherent in conducting research using online social network sites.

Section 1: Introduction

- "Online social networks" refer to a web-based network of individuals who are brought together by common interests, common history or common background.
- Social networking sites allow users to create profiles of personal information, photographs and other digital media, and networks of other users who have access to their information ("friends" or "followers"). Posting brief updates and providing feedback on others' posts are common.
- Adolescents engage in a variety of health jeopardizing behaviors which contribute to morbidity and mortality. Many health-related behaviors are strongly influenced by others – including others in an on-line setting. Young people frequently post information in their profiles about risky behaviors.
- Trust is a subjective degree of belief about agents or objectives based on users' previous experiences or knowledge. The degree of trust in an actor usually builds on feedback from those who have direct interactions with the actor and the risk involved in trusting the actor.
- Successfully leveraging social interaction of users through trust in social networks has the potential to affect adolescent health by propagating correct healthy behavior information from highly trustworthy influential users to other users.
- With the advent of online social networks, it is possible to collect data about millions of people in a social network. This new direction offers exciting opportunities to positively influence adolescent health behaviors.

Section 2: Social Influence on Adolescent Health Behaviors

- A large body of theory and empirical research has demonstrated that individuals are influenced by forces in their social environment, operating at multiple levels.
- Ecological models are well-suited to health research, as a wide range of environmental or system-wide variables may influence health behaviors and outcomes and are, themselves, amenable to action. Influences operate at the intrapersonal, interpersonal, organizational/institutional and societal levels,

which can act directly on an individual's health behaviors, or indirectly through other factors.

- Social network analysis is a distinct approach to conceptualizing social groupings and examining social influence. A social network is made up of individuals (or "nodes") and connecting ties (i.e. relationships between nodes).
- Social influence occurs in social networks in both direct (e.g. communication between two connected individuals) and indirect ways. Indirect influence may come from non-connected individuals or from characteristics of the group as a whole, creating a social norm of expected behavior.
- Research has demonstrated the "spread" of health behaviors and conditions throughout networks of adolescents and adults. The same mechanisms may operate on-line, and these sites may offer new opportunities for intervening to promote healthy behaviors among youth.
- The on-line social networking context offers new challenges and advantages compared to "real world" networks, including network size and sources of data.
- Several types of adolescent health interventions can be adapted to an on-line context, including peer education, social norms interventions, and direct communication from an authority figure.

Section 3. Modeling Trust in Online Social Networks

- We suggest a research framework to model trust in online social networks, aimed at understanding the impact of social influence on adolescent health and improving health behavior through social influence.
- In a data collection, all publicly displayed sections of the web pages in MySpace, Facebook or Twitter can be collected and evaluated.
- One approach to obtaining data is by "crawling," i.e. navigating from one page to another on the website based on the link structure of the website and saving data from these pages. After collecting various type of data (image, text comments, links) from the social networking websites, image processing, human computing, information retrieval techniques, semantic analysis, social network analysis are applied to extract health behavior information from the data.
- A degree of trust among users can be measured by various trust prediction and propagation approaches. Basic underlying principles include: 1) Trust between people can be determined by the similarity between them which in turn depends upon their previous activities and interests; and 2) Direct experience with a connecting user is the most important factor. A history of interactions can be evaluated by the quality and quantity of interactions. Witness experiences like recommendation or evaluation of friends are also important factors.
- Using the examples of Facebook and Twitter, we describe the types of data and interactions in each model and how these interactions might influence adolescent health behaviors. By analyzing the information topics, we can predict a user's health behaviors. With analysis of the quality and quantity of comments between users, we can measure the strength of the relationship between them.

Section 4. Application of Trust Networks for Adolescent Health

- A trust model between two individuals can describe how strongly a person trusts another person on health related experience, opinion or knowledge and how actively a person is interested in another person with respect to health behaviors.

Given this information, the trust network can be used to identify key influential users or opinion leaders to increase positive influence on health behavior or to extract healthy or unhealthy adolescent communities.

- Influential users, or opinion leaders, are highly informed, highly connected and well respected individuals who exercise informal influence over others. Typically a small group of users influences other people's decision making.
- Several models can be applied to identify influential users who effectively propagate correct healthy behavior information to the network or significantly affect other users' health behavior in various ways, including PageRank, HITS and EigenTrustt.
- Identifying groups or communities of youth who positively reinforce healthy or risky behaviors in each other is a useful strategy, and is done by topic-based community detection. This approach examines the kind of discussions that people have with one another and simultaneously looks at their profile information (interests, images, icons, etc). Based on features this technique extracts communities of people who share certain common features.
- Predicting trust links between two users is an additional useful strategy. If influential users can be identified (as above) and links between them and other users can be predicted, interventions could then target specific individuals in a community to spread health related information, increasing the likelihood that others will adapt their behavior accordingly.
- We provide a case study of a specific public health problem, cigarette smoking among adolescents. We apply the social science frameworks described above and our computation model from data collection to modeling, to interventions, to validation.

Section 5: Conclusions

- Several technical challenges are relevant to this work: 1) Social network data often cannot be obtained directly, and must be gathered through indirect means with important limitations; 2) Individuals may provide information on multiple social networking sites, leading to duplication of data; 3) Automatic image annotation techniques require further development to be useful on a large scale; and 4) commonly used software for social network analysis is not able to handle extremely large datasets like those proposed here.
- This work also presents important ethical challenges. Data entered on social networking profiles are not intended to be public information; using it without specific permissions may violate the principle of research subjects' informed consent.
- On-line social networking can be considered within a policy context, and applied to policy change. Social networking sites themselves have policies regarding content, which could be expanded to promote health. Other options exist for supporting healthier on-line content for adolescents. Likewise, on-line activities can be parlayed into "real-world" action (e.g. voting), which may be particularly effective with young people.

TABLE OF CONTENTS

Section 1. Introduction

Today's adolescents were born into a world of computers. With computers and Internet connections in most homes and schools, teens report almost universal access, and they go online for social, informational and entertainment purposes as easily as their parents' generation picked up the phone. Social networking sites, such as MySpace, Facebook and Twitter, offer the opportunity for users to create a profile of their personal information, comments, interests, photos, blogs and other content, to be shared with other users online.

A majority of teens use social networking sites and have at least one profile there – 55% in 2007, and this proportion has steadily increased over time (Lenhart et al, 2007; Bausch and Han, 2009). Rather than making new friends online, most young people engage with friends they already know off-line, and use social networking sites as a mechanism to stay in touch, make plans, chat/instant message, and post pictures or share videos. In keeping with its interactive nature, most teens also report reading others' profiles and blogs, and posting comments in response to pictures or other content (Lenhart et al, 2007). Half of teens visit social networking sites at least once a day, and these sites are emerging as a major communication channel for young people (Lenhart et al, 2007; Lenhart and Madden, 2007).

1.1. Online social networks

An online social network refers to a web-based network of individuals who are brought together by common interests, common history or common background. According to a recent Nielsen report titled "Global Faces and Networked Places", 67% of the online global community are visiting and participating in activities in online social networks which is now the fourth most popular activity online (Nielsen, 2009). In addition to personal use of these websites, these have been employed for other applications as well. Online social networks have also been used by government agencies, such as the CDC, to spread awareness about diseases or infections (Seitz, 2009) and social networking websites such as PatientsLikeMe have been used to connect patients with similar ailments and symptoms to share common experiences (Goetz, 2009).

For purposes of this report, we will focus on three specific social networking sites (SNS) which are popular with today's young people, namely MySpace, Facebook and Twitter. We will briefly introduce these three sites here.

MySpace and Facebook are the most popular online social networking websites on the Internet, and both fall under the umbrella of social network services. While MySpace used to be the largest and the most visited social network, there is some evidence that it has now been surpassed by Facebook (Albanesius, 2009; Patriquin, 2007). MySpace was launched in August 2003 and by June 2006 it had become the largest online social networking website in America (Mashable, 2006). Common features in social network services include the option to create user profiles where users can give demographic information about themselves including age, gender, location, educational or work background, hobbies and interests, etc. Users can also upload pictures of themselves on their profile page. They can specify which of the other users are their "friends," and can also specify their privacy levels to manage which other users can view each part of their profile. These websites also have spaces where users can write "testimonials" or comments about each other. In short, there is a wealth of information about these users available online.

Facebook started out as a social networking website for college students. It was launched at Harvard University in 2004 and gradually other universities were added. In 2006, Facebook opened to non US college communities and experienced rapid growth thereafter. According to the information updated by Facebook on its website, Facebook

has more than 350 million active users. Facebook users spend more than 8 billion minutes on Facebook (Facebook Statistics). The structure of Facebook is similar to that of MySpace described above, in that users share information about themselves and link to other users. One difference between Facebook and MySpace is the overall demographics of these two websites. Thus Facebook users in general come from more affluent backgrounds while MySpace users come from blue collar background (Boyd, 2007).

Twitter is the most famous and representative form of online social networking known as "microblogging." Users of Twitter can send or post very brief messages, called "tweets." These messages are displayed on the page of the author but they can also be directly forwarded to other users who are "following" the user who posts a tweet. Users who are posting tweets also have the option of making the tweets open to everyone or to restrict them to users who are part of their social network or "friends." Twitter also allows users to send and receive the tweets as SMS messages. Twitter has permeated the mainstream media. For example, the 2008 Obama presidential campaign used tweets to inform its supporters of the up to date events on the campaign trail. Research has also shown that Twitter does a better job of informing people about disasters and enabling people to keep in touch with their loved ones in times of disasters as compared to other forms of media (Palmer, 2009).

In social networking websites, a significant component of the activity is writing and replying to wall posts, posting notes, etc. In brief, each of these activities involves writing some text or replying to text written by someone else. Users also have the option to upload pictures of themselves and of their friends. These pictures can be tagged by other users, where the tags usually identify the people in the pictures. In some cases users also provide a description of the image, or even objects in the image, and others can add comments on these images.

1.2. Adolescent health behaviors

According to the CDC's Youth Risk Behavior Surveillance System, young people in the United States engage in a variety of health jeopardizing behaviors (CDC, 2008). Substance use is high, with three-quarters of high school students reporting having drunk alcohol, and 20% having smoked cigarettes in the past month. Almost half of them were sexually experienced. A large majority fail to meet recommendations for fruit and vegetable consumption and physical activity. It has been well-established that these behaviors contribute to morbidity and mortality among young people, and to disease later in life. Many health-related behaviors include a social component, in that they occur with others or are strongly influenced by others. This includes others in an online setting: because for most adolescents their online world reflects their real lives, health behaviors are often represented, discussed and described on social networking sites. Using all the features of these sites, young people frequently post information in their profiles about risky behaviors (Moreno et al, 2009).

Approximately one-third of teens also use Internet (though perhaps not social networking sites specifically) to seek out health information on topics such as dieting and physical fitness (Lenhert et al, 2005). In addition, teens report using online sources to research more sensitive health topics, including drug use, sexual health or depression; 22% report searching for these topics (Lenhert et al, 2005). Sun and colleagues (2005) have found that adolescents with higher levels of psychosocial risk factors or high risk health behaviors were more likely to use the Internet, suggesting that an online approach might be a particularly effective strategy for reaching at-risk populations. Numerous health-related websites target young people with these types of concerns.

1.3. Modeling trust

The issue of trust arises wherever there is social interaction between individuals, groups or organizations. The success of social interactions for personal and professional information sharing among users highly depends on 'trust' between users, and thus trust is an important principle of society. Therefore, answering questions like "whom to trust, and to what extent?" for individual users in online social networks has been an issue of high interest; including application of trust in various domains. Various definitions of trust have been proposed which are applicable in different contexts. In general, trust is a subjective degree of belief about agents or objectives based on users' previous experiences or knowledge. The degree of trust in an actor usually builds on feedback from those who have direct interactions with the actor and the risk involved in trusting the actor. Thus, the degree of trust between two people or entities can easily translate into influence or facilitating flow of information between them.

Given some feedback, one's trustworthiness can be inferred through use of a trust function (model) (Golbeck, 2005). Many trust functions have been proposed, targeted at different application domains, and they differ in various aspects including trust inferences methodologies, necessary information to infer trust, complexity and accuracy.

Trust functions have a direct impact on one's decision regarding information collection and sharing, participation in social groups, establishment of actual transactions and other critical tasks. Thus, trust in online social networks has a significant impact on the way of understanding the pattern of social interactions, and the formation of groups by trust influence. In other words, successfully leveraging social interaction of users through trust in social networks has the potential to affect adolescent health by propagating correct healthy behavior information from highly trustworthy influential users to other users.

The potential power of interactive social networking sites as a means of collecting and distributing health information is, as yet, untapped. Traditionally psychologists have used surveys and field research to collect data about social networks in the real world. Due to the time consuming nature of collecting such data, the size of such datasets was also limited. With the advent of online social networks now it is possible to collect data about millions of people in a social network, which has also broadened the kind of questions that can be addressed.

In order to maximize the effectiveness of such approaches, researchers will need to identify key influential users as sources of information, ways in which information is transmitted within online social networks, extracting healthy or unhealthy youth communities for targeting health advertising/campaign and recommending healthy trustworthy friends and communities. This promising new direction offers exciting opportunities to positively influence adolescent health behaviors.

In this report, we describe the theoretical framework of social influence, and the social network models on which this work is based, in Section 2. In Section 3, we propose the research framework for a computational trust model which covers data collection, computational web data analysis techniques and trust prediction models. Based on this framework, we describe applications of the trust network in order to maximize the impact of trust based social influence among adolescents on their health behavior in Section 4. Finally, in Section 5, we discuss several technical and ethics challenges inherent in conducting research using online social network sites.

Section 2. Social Influence on Adolescent Health Behaviors

2.1. Social ecological models of influence on health behavior

A large body of theory and empirical research has demonstrated that individuals are influenced by forces in their social environment. One set of theoretical frameworks which have wide application to health behaviors are social ecological models (Bronfenbrenner, 1979; McElroy, 1988; Sallis and Owen, 2002). While they differ slightly in the particulars, these models emphasize multiple levels of influence, encompassing the intrapersonal, interpersonal, organizational/institutional and societal factors. Numerous theorists have noted the importance of using a multilevel framework for examining health behaviors (Wicker, 1979; Rose 1992; Rew, 2005), positing that individuals are only one part of a larger system which, by its design, promotes (or even demands) certain actions and discourages (or prohibits) others (Wicker, 1979). Ecological models are well-suited to health research, as a wide range of environmental or system-wide variables may influence health behaviors and outcomes and are, themselves, amenable to action (Stokols, 1992; Sallis et al, 2002). As we detail below, this framework can also be used to conceptualize health behaviors in the context of on-line social influences.

A general social ecological model is shown in Figure 2.1, and briefly described here. The centermost ring denotes the individual, and the *intrapersonal factors* that most immediately affect health behaviors. These include cognitive and psychological processes such as knowledge; attitudes, beliefs and meanings ascribed to the behavior; and expectancies regarding positive and negative consequences of the behavior. These intrapersonal factors are learned, observed, or gained through personal experience. *Interpersonal factors* are shown in the next ring. These influences are characterized by one-to-one or one-to-group interactions with other individuals such as friends, other peers, family members, teachers, mentors, neighbors, co-workers, etc. They can include a direct exchange of information (e.g. teaching, demonstrating) or communication (e.g. pressuring, teasing), or more indirect routes such as modeling the behavior and its expected outcomes (e.g. a smoking peer looking "cool," a smoking relative getting cancer). *Organizational/institutional factors* refer to characteristics of the overarching group or organization, such as the school, church, family or workplace. These can include policies (e.g. smoke-free workplace, insurance discounts for fitness center membership), programs (e.g. health education, condom distribution in high schools), or other characteristics (e.g. proximity to fast food outlet, family connectedness); and can influence health behaviors by improving or impeding access to the means of health behaviors, or setting clear guidelines regarding the desirability and acceptability of relevant behaviors. At the *societal* level, characteristics of the overarching social context can also exert a strong, if distant, influence on individuals' health behaviors. Public policy dictates access to health care, availability and pricing of the means of health behaviors (e.g. substances, weapons, healthy or unhealthy foods), and personal protection devices (e.g. seatbelts, helmets), to name a few. Advertising and entertainment media send powerful messages about health behaviors (e.g. unattainably thin or muscular bodies, sex without consequences), and create social norms which dictate appropriate behaviors to fit in with one's referent group. Cultural values regarding family, faith, gender roles, etc, also influence health behaviors.

In considering interventions based on a social-ecological framework, two features stand out. First, each level is theorized to influence and be influenced by each other level. For example, as generations of smokers (and those regularly exposed to second-hand smoke) have died prematurely, sociocultural values around smoking have shifted in the U.S., becoming less tolerant of this behavior. This has been borne out in landmark

lawsuits and resulting policy changes, such as smoking bans on airplanes, and subsequently in many other types of workplace (organizational/institutional changes). These changes have led to fewer smokers acting as *interpersonal* influences, and changed the *beliefs and attitudes* about smoking for many individuals. Individuals then act as influences on their own network of interpersonal contacts, who can then create, support or demand organizational change (e.g. an expansion of smoke-free spaces), which contributes to further changes in social norms, values, and public policy. The paths are fluid and dynamic, sometimes subtle and sometimes quite obvious. Second, intervening on the innermost circle is the most straightforward, and becomes increasingly difficult in each successive ring. However, the reach of the intervention is the smallest in the inner circle (i.e. one individual), and increases to a large population at the outermost level.

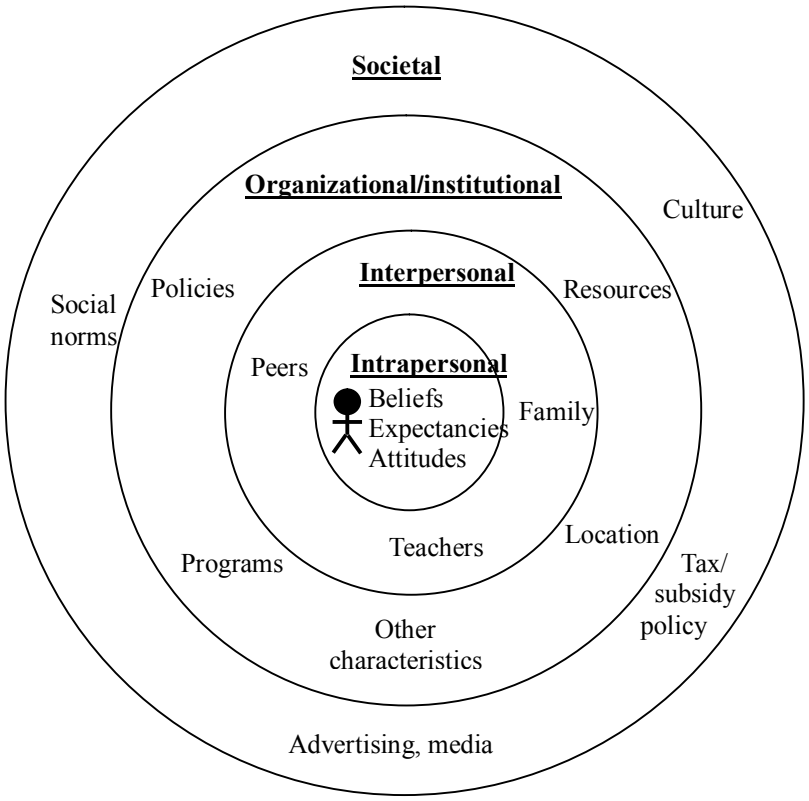


Figure 2.1. Multilevel influences on health behaviors

2.2. Social networks and its analysis

Social networks and social network analysis are a distinct approach to conceptualizing social groupings and examining social influence, which may be particularly salient to adolescents as they shift from home and family as the most significant influence towards increased attention to peers. Connections to others, both within and outside of an established group, serve an important function as adolescents go through the developmental process of establishing their own identity in a social context. A social network approach can be mapped onto the interpersonal, organizational and cultural levels of the social ecological model described above, and has direct application to on-line social networking.

A social network is made up of individuals (or "nodes") and connecting ties. Connections can be of many types – commonly friendship, kinship, romantic or sexual partnerships, trust or financial exchange. Networks are typically represented graphically, as a diagram or map of all nodes under study, and all their connecting ties. An example of a complex network is shown in Figure 2.2.

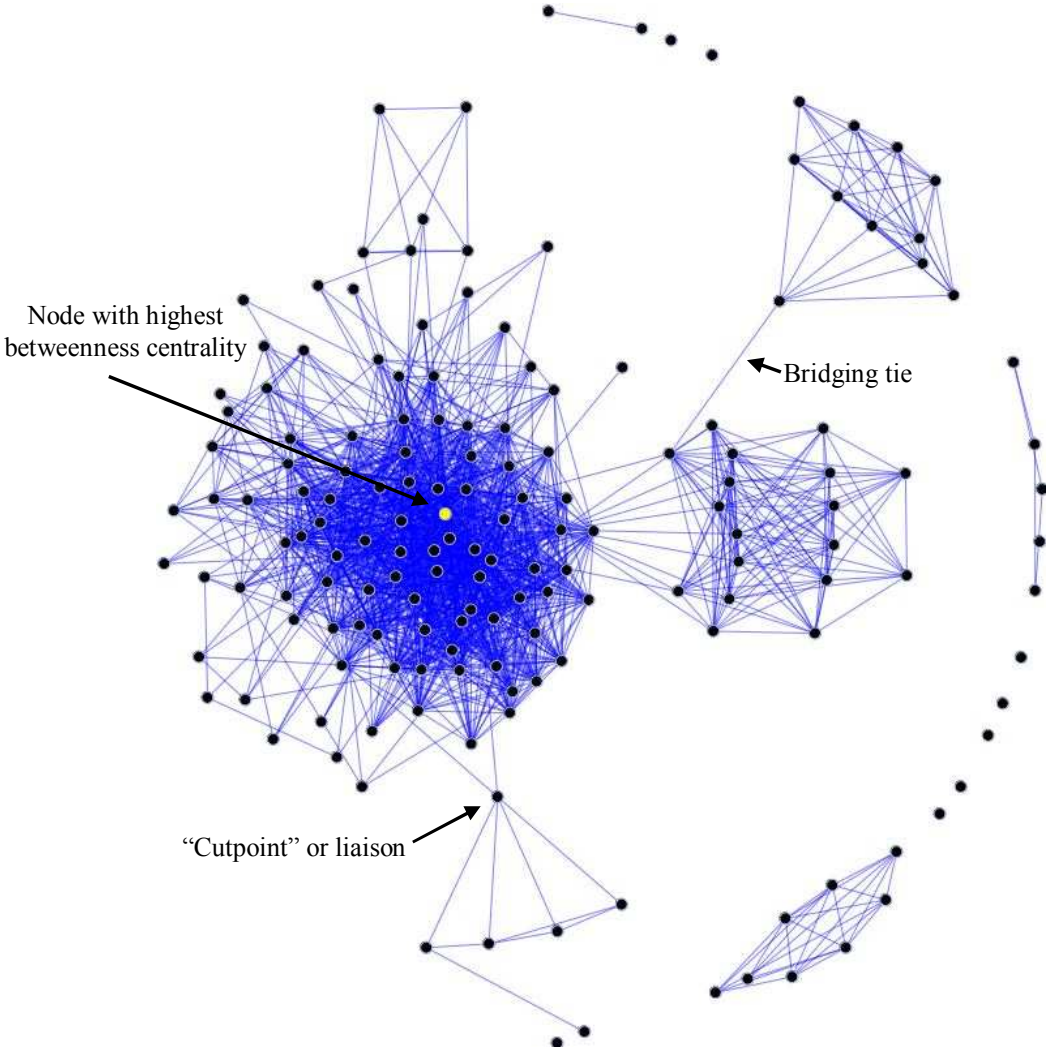


Figure 2.2. An example of a social network diagram

Several characteristics of social networks are relevant to research on the spread of information and influence (Cotterell, 2006). *Size* is a measure of the number of nodes in the network under study, and speaks to the potential reach of any new idea or intervention based on a social network model. *Composition* is a measure of the type of relations in the network, for example, friends, family members or business associates. *Density* is a measure of the number of ties within the social network; it is a fraction of all the possible ties among all nodes. Dense networks can spread information and influence more efficiently than sparsely connected networks, due to the higher number of links among individuals. *Reachability* is a measure of the number of steps required to connect any two individuals – a concept known in the popular culture as "degrees of separation." In order for individual A to influence individual B, they must be reachable through some pathway of connection through other individuals. *Betweenness* refers to

the extent to which an individual links other individuals or subnetworks in the system. Those with high "betweenness" are key players in transmitting information or influence across the network. *Centrality* can be measured in several different ways, including degree centrality (the number of direct ties with other nodes in the network), closeness centrality (a measure of closeness to all others in the network) and betweenness centrality (the extent to which individuals play a key role in linking portions of the network and function as a "middle man").

Two additional features of the social network model bear mentioning, due to their role in transmitting information. A liaison or "cutpoint" of a network graph represents an individual linking two distinct clusters of the network. While this individual does not have a particularly high number of direct ties or high betweenness, this node's position provides a unique and crucial link to a subgroup which would otherwise be isolated from the rest of the system. Likewise "bridging ties" between two players connect portions of the system. Dissolution of the relationship linking them would isolate the otherwise unconnected cluster in the upper right portion of Figure 2.2.

Social influence occurs in social networks in both direct and indirect ways. Direct communication between two connected individuals is perhaps the most obvious form of influence, and an example of an interpersonal factor in the social ecological framework. In the realm of adolescent health behavior, these overt messages can be in the form of encouragement, teasing, or providing means (e.g. sharing cigarettes) – all of which might fall under the familiar heading of "peer pressure." Such exchanges are concrete and observable; they are therefore relatively easy to assess and target with intervention activities. Indirect influences are far more subtle. At the interpersonal level, one network member can influence an adjacent member by espousing particular beliefs or modeling new behaviors in an observable setting. Even without any "peer pressure," per se, one friend hearing of or observing a behavior in a close peer can be a powerful motivator; the exchange can serve to increase familiarity with new views or behaviors, revise expectancies of outcomes associated with the behavior, or provide a clear guideline for what is needed to fit in or maintain closeness with this peer. At the broader societal levels, indirect influence can occur when the shared attitudes and behaviors of network members create a social norm or an accepted standard of behavior. Simply by perceiving what "everyone" does or what key opinion leaders do, individuals' own judgments of the acceptability or desirability of a behavior are shaped and re-shaped. Indirect influences are much more difficult to capture – they are not directly observable and individuals typically do not perceive they are influenced in this way at all. Creating interventions to target these influences is therefore much more challenging. Successful interventions, however, have the power to be quite far-reaching, as they can apply to the entire network system without relying on transmission across each specific network tie.

One of the clearest applications of the social network analysis and model to public health is the spread of infectious disease, which becomes a useful metaphor for conceptualizing the spread of social influence. An illustrative example comes from the work of Klovdahl and colleagues (1994), who mapped a large network of over 600 adults to detect the spread of HIV in the Colorado Springs area. In analyzing the core of the network, the researchers determined that one of the three HIV positive individuals was in this highly connected cluster, and he was able to connect to 563 other individuals in seven or fewer steps (and many within three steps), demonstrating the reach of a single infection and the pathways on which it could move in the absence of preventive measures (such as consistent condom use) which could break the "tie" between individuals. They further identified a key associate of the infected individual who was both uninfected and had a particularly large number of network ties. If the virus were to spread to this individual, it would have the potential to dramatically increase the spread

throughout the network; this individual, the researchers argued, should therefore be a focal point of intervention in order to prevent his infection and further transmission of the virus.

While infectious diseases take advantage of biological processes to move across networks, other health conditions and behaviors can be similarly spread through social processes. Epidemiologists have recently applied social network theory and analysis to the spread of obesity, cigarette smoking, and even happiness in a large network of participants in the Framingham Heart Study (Christakis and Fowler, 2007; Christakis and Fowler, 2008; Fowler and Christakis, 2008a). They found that obesity spreads over time throughout social networks out to third degree contacts, and make a compelling case for the influence of friends on each other, rather than similar individuals simply becoming friends (Fowler and Christakis, 2008b). Likewise, they found that network phenomena appear to be at play in smoking cessation, as they observed smoking behavior spreading through close and distant social ties, smokers quitting in clustered groups, and smokers becoming increasingly socially marginalized (Christakis and Fowler, 2008).

Although social network research has captured national attention in the past few years, similar research with adolescents has been conducted for considerably longer, and has a fair presence in both the health and economics literatures. Researchers long ago identified friends' behaviors as a key influence on certain high-risk or delinquent behaviors such as substance use or vandalism, which typically take place in groups. However, dieting, unhealthy weight control and binge eating (i.e., more private behaviors) have also been observed to spread through social ties. Crandall, for example (1988), studied college sororities and found that members' binge eating behaviors grew increasingly similar to their friends' behaviors over time. Paxton and colleagues (1999) used social network analysis to identify 79 friendship cliques among high school students, and found that body image concerns, dietary restraint and disordered eating behaviors were more similar within than between friendship groups.

In recent years, comprehensive friendship databases and sophisticated analytic techniques have become available to address the problems of directionality (i.e. friends influencing one another vs. similar individuals becoming friends) and other contextual confounding (i.e. friends being similarly affected by an external, environmental characteristic) that plague much of this body of work. As with adults, Christakis and Fowler (2008) and Trogdon and colleagues (2008) found that weight status is correlated and spreads in networks, even after appropriately controlling for endogenous effects. Substance use research in peer networks has also been replicated in adolescents. Researchers have found behaviors clustered within well-defined social boundaries, resulting in characteristics of "smoking" and "drinking" schools, and have found significant peer group effects for alcohol use in particular, such that students appear to respond to the behaviors of their peers (Clark and Loheac, 2007). The influence of peers is complex, however, and appears to go beyond simply the peers' own behavior. Characteristics such as friendship quality, peer social status and embeddedness (the degree to which individuals are involved in a social network) of friendships have also been associated with adolescent smoking involvement (Ennett et al, 2008), suggesting that many other elements of peer influence have yet to be fully explored.

2.3. Application to online social networking sites

Strong evidence is mounting regarding "real-life" social networks and social influences on the health status and behaviors of young people. However, in less than the past decade, adolescents have begun using on-line social networking sites to communicate with others and share information. This new technology represents a

brand new arena in which to study social networks and social influences on health behavior. The theoretical frameworks described above can be directly applied to the on-line experience, and these sites may offer new insights and opportunities for intervening to promote healthy behaviors among youth.

As in real life, researchers have noted in recent years that adolescent users of on-line social networks often share quite detailed and personal information, including about health behaviors (Moreno et al, 2007; Williams and Merten, 2008; Moreno et al, 2009). Using all the features of these sites, young people frequently post information to their profiles about risky behaviors: for example, a recent review of 18-year-olds' MySpace profiles found that 24% referenced sexual behaviors, 41% referenced substance use, and 14% referenced violence; over half included one or more of these references (Moreno et al, 2009).

In considering new social influences – or new venues for social influence – on adolescents' health behaviors, the social ecological framework can readily apply to social networking sites. Each aspect can speak to trust relationships in the on-line world, can influence adolescents' health behaviors and can serve as a potential point of intervention. The interpersonal level of influence is the most obvious, as individual users of SNS link to others in the network by identifying them as "friends," "followers," etc. Typically it is these identified users who have access to an adolescent's profile or page, to read material or post messages, comments, photo tags, and other input that could be relevant to health or behaviors. Such feedback can take the place of real-life verbal commentary such as sharing information, encouraging or teasing, to act directly on an adolescent's beliefs and expectations about risk behaviors.

The social network service (SNS) itself could be construed as the organizational level of influence, as its policies and resources can act on individuals in that space, similar to school or other institutional policies. Prohibitions against sexually explicit photos, for example, maintain an environment where certain behaviors are clearly unacceptable. Likewise, virtual bumper stickers, gifts or other downloaded icons and slogans, which can be exchanged among users and used to decorate profile pages, often include health-related content, such as a marijuana leaf, playboy bunny, or beer brand logo. The availability – or not – of various images also confines to some extent the types of behaviors which can be displayed and thereby indirectly promoted. Another organizational feature is partnerships with external agencies to create or promote outside applications, such as online social games available on Facebook or MySpace like Mafia Wars (See appendix), a popular feature of some sites. All of these examples demonstrate ways in which the policies of the sites themselves can promote or restrict imagery or messages regarding health behaviors. At this time, restrictions on content are few; this represents a potential future direction for intervening on health issues via SNS.

Perhaps the most significant type of social influence in the on-line context is occurs at the cultural level, in the creation of social norms of behavior. This is particularly germane in this case, as individuals have the capacity to directly observe the postings of all other individuals within their network, as well as the exchanges about these materials made by other users, including those *outside* their own personal network. For example, a posted photo of Adolescent *A* drinking beer at a party may be visible to Adolescent *B*, along with comments by other friends of *A* which might convey the fun of the party, popularity, coolness or "in-group" status of Adolescent *A*. In this way, it is not only the posted photo of a single friend that can be influential at an interpersonal level, but also the feedback collected from linked networks which create a powerful image of youth life. Ubiquitous images of peers engaging in any behavior create a social expectation around

that behavior – the idea that "everyone's doing it" – which then contributes to uptake of that same behavior (Berkowitz, 2005).

On-line social network research is a natural extension of existing social network modeling, but with a higher degree of complexity. First, in real-world social network analysis with young people, friends are typically nominated from a universe of possibilities (names on a school roster, for example). Data are collected from as many network members as are available and links within this universe are mapped. Published studies of hundreds or perhaps thousands of networked individuals are typical, and establishing discrete networks (e.g. in different school districts) is possible. In on-line social networks, however, the number of friends or followers can be much higher, and they are not bounded by school, age group, geography or other clustering factor. While the average Facebook user (across age groups) has 130 networked friends, this number is much higher for adolescents who may claim hundreds of friends, as a large proportion of their social universe is tech-savvy and present on-line, and this type of popularity can confer status. Contrasted with real-world network research in which individuals are asked to name, typically, 1-10 friends, online social network mapping grows exponentially more complicated and challenging. But this domain offers advantages as well. Where previous research may rely simply on friendship nominations or perhaps more nuanced characteristics such as friends' status, using on-line data mining techniques provides insight into which connections are most salient to an individual (for example, by frequency of and time spent viewing others' content, or exchanging feedback) and may therefore be most influential. Direct analysis of the content of comments, feedback or photos (as detailed in Section 3) provides an additional opportunity to gain insight into the mechanisms of influence which are not usually available in real-world network research.

Possibilities for using on-line social networking to influence adolescent health behaviors

In this section we summarize three types of intervention strategies which rely on social influence and have been implemented with young people. For each, we briefly describe the approach, evidence of its effectiveness, and ways in which it could be adapted for use in the SNS setting.

Peer education is a theoretically grounded approach to reaching adolescents with health information, which employs adolescents themselves as sources of health information and models of healthy behavior (Ochieng, 2001). It is often used as a component of a more comprehensive educational program, and has most commonly been applied to sexuality education and HIV prevention (Sawyer et al, 1997; Ochieng et al, 2001; Ebreo et al, 2002). Peer educators are, ideally, well-respected and central figures in the social network, enabling them to connect directly with a large number of peers and use their status to enhance the influence of their message. Unfortunately, peer education programs are difficult to evaluate in isolation from other educational activities. Research has shown, however, that the greatest effect of peer education efforts appears to be on the educators themselves (Komro and Perry, 1996; Shiner, 1999; Harden et al, 2001; Karcher et al, 2004). In an on-line context, then, we might expect a similar process. After identifying key individuals who are widely followed, friended, or otherwise appear to hold the trust of the peer network, intervention researchers can employ them – as they do in real-world settings – to share health information, and promote and model healthy behaviors. For example, on-line peer educators might post health-related information on their own profile, status update or blog; or comment on the images and messages on others' pages. This approach uses the social network to

affect both the interpersonal level of influence as well as the social norms around health behaviors.

Social norms interventions (Berkowitz, 2005) seek to correct misperceptions about the prevalence of risk behaviors in a social setting, particularly behaviors that are commonly overestimated, such as binge drinking and high risk sexual practices (Perkins and Wechsler, 1996; Licciardone, 2003; Scholly et al, 2005). These interventions are based on the premise that young people adopt behaviors they believe will help them fit into a social group or setting (e.g. college campus), and that a more realistic perception of behaviors (i.e. *not* everyone's doing it) will reduce their uptake. Social norms interventions typically provide feedback about risk behavior using a web-based format (including e-mail), in-person or group sessions, or a community-wide campaign. Effectiveness in real-life settings has been mixed, with greater effects seen for specific types of normative feedback and some behaviors but not others (Werch et al, 2000; Scholly et al, 2005; Perkins and Craig, 2006; Moriera et al, 2009). These methods could easily be applied in a SNS setting, building on research demonstrating that web-based personalized normative feedback is an effective strategy with young people (Moreiera et al, 2009). This feedback can come from health professionals or other network members who are linked to adolescent users, with attention to careful design and tailoring of messages in order to be credible and trustworthy in this population. Additionally, sidebar advertising could include normative information about health behaviors, and be targeted to those profiles where high risk health behavior content is detected – or to those who spend time viewing profiles where such content is detected.

Direct communication from an authority figure, health professional or teacher is a staple of health education, based on the premise that having knowledge about health issues is necessary (but not sufficient) to establish healthy behaviors. Classroom-based health education is routinely used to promote healthy dietary intake, cardiovascular health, personal safety, and prevention of substance use, teen pregnancy and sexually transmitted infections. In addition to classroom teachers, the well-known Drug Abuse Resistance Education (D.A.R.E.) program, for example, uses police officers as instructors to warn students of the dangers of substance use (Hammond et al, 2008). This approach has been tried in an on-line context, with promising results. Moreno and colleagues (2009b) identified MySpace users with public profiles which included references to sex and/or substance use, with low-security settings. They sent a brief e-mail message noting the public nature of these references and suggesting that the user revise his/her page to better protect privacy; the researchers found significant protective changes at 3 month follow-up. Expanding this type of application to include other salient authority figures, opinion leaders or SNS webmasters, as well as capitalizing on the linked network structure (i.e. intervening with several key actors in the network to maximize reach and change norms) could prove even more effective at changing adolescent behaviors.

Section 3. Modeling Trust in Online Social Networks for Adolescent Health

3.1. Overall framework

In this section we describe our research framework to model trust in online social networks, aimed at understanding adolescent health behavior. The overall framework is described in Figure 3.1, which shows some of the online social networking sites that are popular with adolescents in general. These can be used to collect self-reported user profile information on age, social interactions, status updates, posts - communication/text exchanges, list of friends and list of affiliated communities or bloggers. In section 3.2, we briefly describe a general data collection process from online

social networks. All publicly displayed sections of the web pages can be collected and evaluated; e.g. sexual behavior display (e.g., personal sexual preference, self-disclosures of sexual experiences, pictures of profile owner in undergarments and downloaded sexually suggestive icons such as Playboy bunnies and so on), substance use display (e.g., pictures, blogs, lists of favorites, and downloaded icons for display alcohol use, tobacco use, or drug use) and healthy behavior display (e.g., church/religion involvement or sports/hobby related pictures, blogs and links).

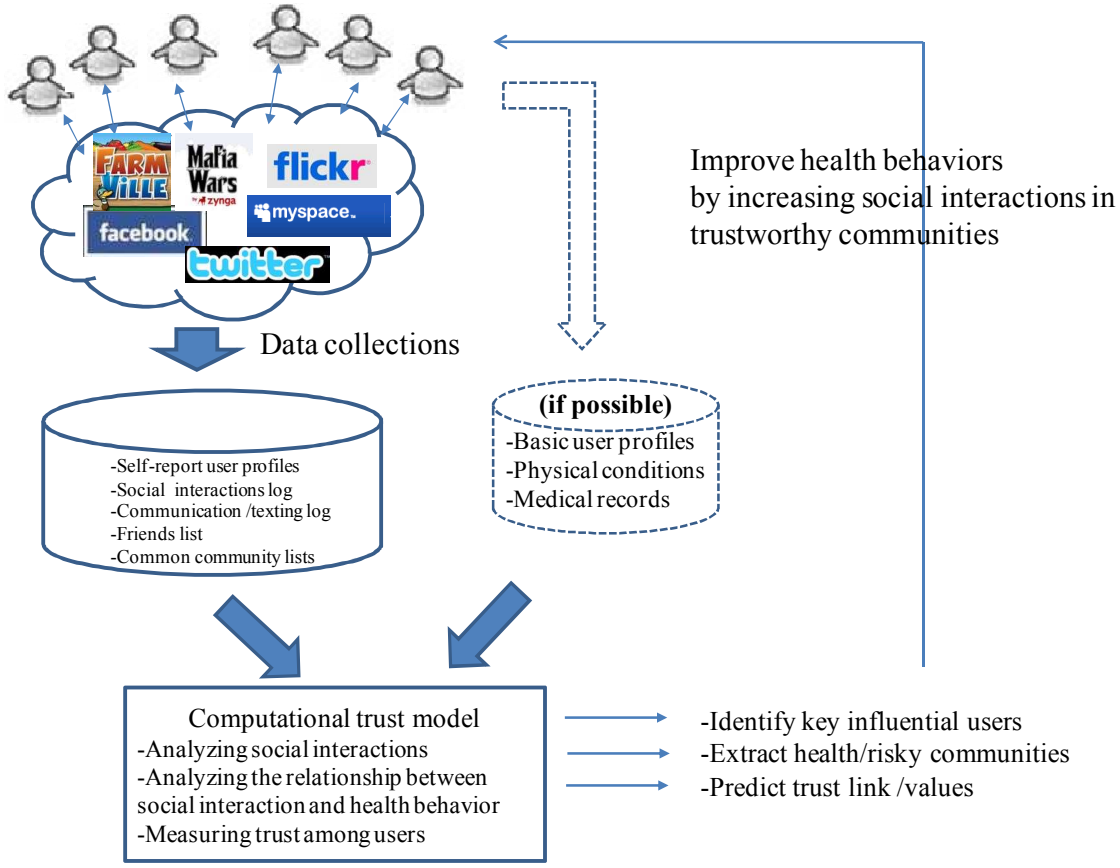


Figure 3.1. A research framework

If it is possible to collect information regarding personal physical condition or medical records about users from online social networks, we can do a deeper analysis of how social interactions from one's peer group or trustworthy acquaintances or friends affect adolescent health behavior and physical health condition. In section 3.3, we describe different computational techniques which can be used on different types of data (text, photos, links etc) from various social network services and the insights that can be gleaned from them.

The first task in our framework is data collection, which is then used to develop a computational trust model. Since we utilize these social networks as a new way to intervene and improve adolescent health, a trust model in online social networks should be constructed by considering social interaction around health-related experiences, knowledge and affiliation with health communities. Thus, we consider relevant data available to us like public profiles and social interactions in online social networks and

discover the positive and negative impact of the possible social interaction on health behavior. Given social interaction that can implicitly or explicitly show a strong relationship with adolescent health behavior, a degree of trust among users can be measured by the following approaches: game-theoretical approaches, referral network approaches, belief-oriented approaches, agent-based approaches, propagation approaches etc. In section 3.4, we explain various trust prediction and propagation approach. In section 3.5, Facebook and Twitter are used as case studies and we explain what kind of social interactions can be used as good indicators of healthy behavior.

A trust model between two individuals can describe how strongly a person trusts another person on health related experience, opinion or knowledge and how actively a person is interested in another person with respect to shared experiences regarding health behavior. Given this information, the trust network can be used to identify key influential users or opinion leaders to increase positive influence on health behavior or to extract healthy or unhealthy adolescent communities. This is explained in more detail as application of trust network in section 4.

3.2. Collecting data from online social networks

Online social networks can be of various types. They can range from social networking websites like Facebook or MySpace which are explicitly designed for this purpose to microblogging website like Twitter where the social networks can be constructed by how users respond to each other and how they "follow" each other. Additionally some blogging websites like LiveJournal also have a social networking component. Yet another category of social networks are social networks formed by online messaging services like AOL Messenger, MSN Live Messenger, etc. All these different online social networks represents different *types* of social networks with respect to the kind of interactions that people can have e.g., in case of instant messaging networks the interaction between the users is mostly synchronous, mostly simultaneous and instantaneous. On the other hand in case of Online Social Network Service websites like MySpace and Facebook the interaction between the users is asynchronous. Some web services like Facebook have tried to bring various aspects of communication together by introducing live chat as a feature on their website in additional to other types of messages. Here we give an overview of two different types of online social networks for which it is possible to collect data off the internet and then analyze it to make useful inferences about the userbase.

While there is potentially a very rich reservoir of information present about users of online social networks and their friends, the availability of the data for the purpose of analysis is limited for a number of reasons. Thus one of the foremost issues with analyzing data from online social networks is accessing the data itself; companies which are running these websites do not want to share all of their data with outsiders due to users' privacy concerns. One approach to obtaining data is by "crawling" these websites. Crawling refers to navigating from one page to another on the website based on the link structure of the website and saving data from these pages (Kobayashi, 2000). While crawling does not guarantee that all of the relevant data will be extracted, it can still extract a fair chunk of the data. One limitation with crawling is that some websites explicitly disallow crawlers to extract more than a certain amount of data from their websites. However crawling is the one of the best methods to extract data from some social networking websites like MySpace where most of the profiles are publically available. Similarly most of the Twitter data is also publically viewable unless the user explicitly changes the privacy settings to restrict the viewer to only his or her "friends." Data from Twitter can also be extracted by crawling in a manner similar to crawling a website. For a social networking application, one would follow the "friends" that a user

has and extract the data accordingly, rather than simply following hyperlinks as is done in standard web crawling.

On many other social networking sites e.g., Facebook most profiles are not publicly available and in fact require registration to even view the profiles of most users. However, in the case of Facebook, data can still be extracted by developing an application for the Facebook platform and asking Facebook users to install that application. The application can then extract the relevant information for these users. Common examples of widely used applications are quizzes which are available on sites and advertised explicitly or by "word of mouth" from one user to another. These applications allow one to extract not only information collected through the application e.g., responses to quizzes, but also profile, friend, and other information about the individual user.

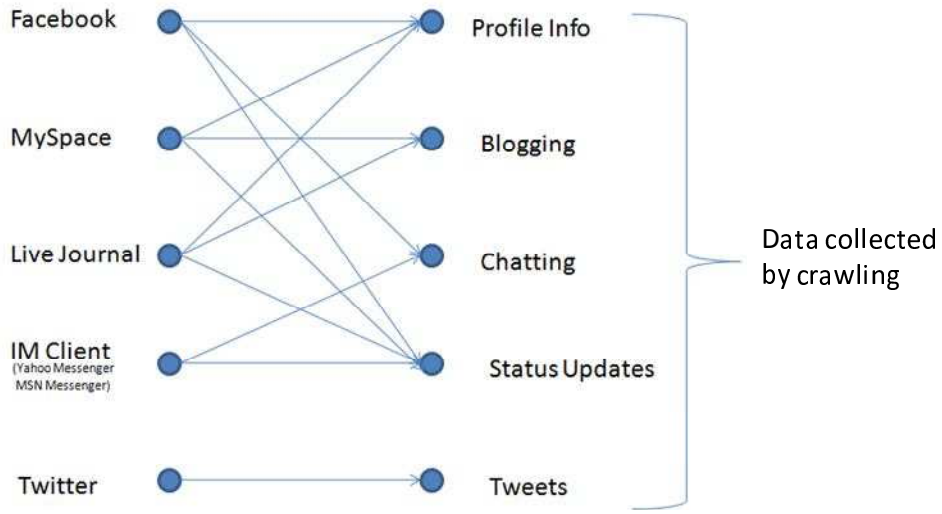


Figure 3.2. Types of datasets available from various sources

Different sources provide different types of data which have to be extracted by crawling these websites as described previously. Many of the websites have profile information available about the users which describes their background, demographic information, interests, hobbies etc; this is common to social networking websites like Facebook and MySpace and to blogging websites like LiveJournal. We also note that some of the social networking websites like MySpace also have blogging capabilities which provide a rich data source for data mining. People in many of these websites also have the option to update the rest of the people in their social networks about their status which can be another source of information about their behavior. Tweets are somewhat analogous to status messages but with the difference that tweets are generally more frequent.

3.3. Mapping web data to health knowledge

In section 3.2 we described various sources which can be used to collect data about individuals for the purpose of positively affecting healthy habits in adolescents. However just collecting the data is insufficient; one also has to process the data in order to understand what it represents and how extracted knowledge from the data can be used into actionable recommendations. In this section we describe different computational techniques which can be used on different types of data from various social network services and the insights that can be gleaned from them. Figure 3.3 gives the summary of

various types of data from the social networking websites, what kind of techniques can be applied to each and what health information can be extracted.

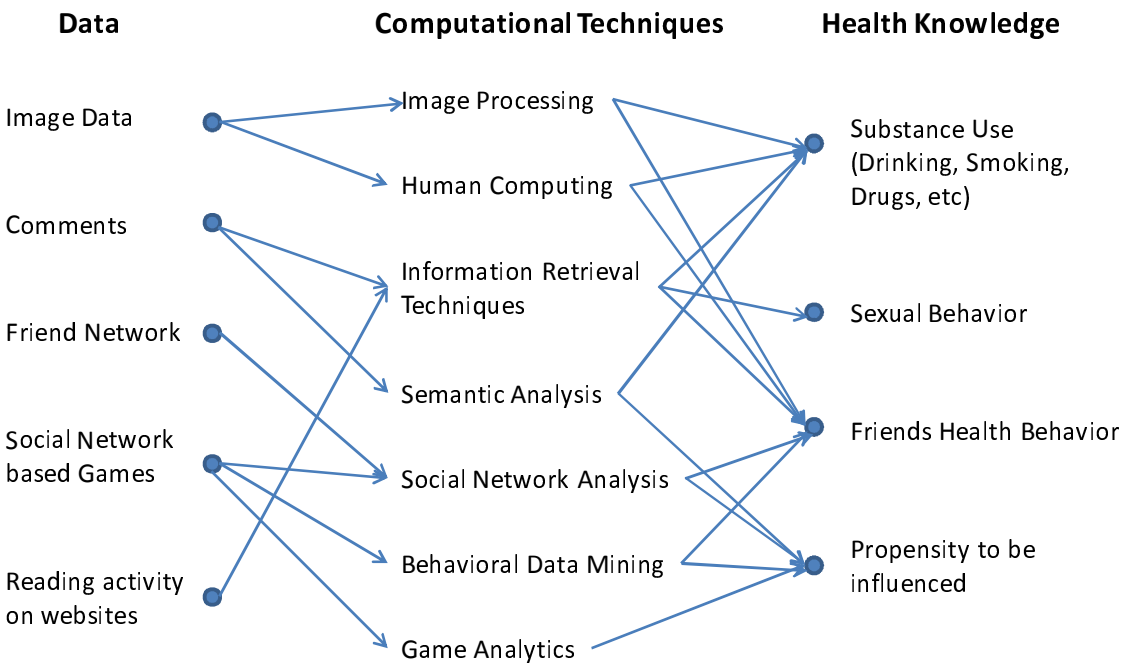


Figure 3.3. Mapping from data to knowledge in social networks.

The content of images posted on SNS can be used as markers of risky behaviors e.g., consumption of alcohol can be gleaned by the presence of alcoholic beverages in images and the frequency of such images. Identifying such markers like alcohol or alcoholic beverages can be done by two different techniques: *Image processing* and *Human Computing*. Image processing can be used for object identification in images to determine the types of objects present in image (Jain, 1989). However object identification is still considered to be a hard problem in computing and the performance of the algorithms varies in different domain. The other set of techniques that can be applied fall under the umbrella of human computing (von Ahn, 2006) where the idea is to seek the help of humans in a distributed manner to solve problems which are considered hard by computer. This strategy has been successfully employed by Luis von Ahn to get people to correctly annotate hundreds and thousands of images via the ESP (Extra Sensory Perception) game (von Ahn and Dabbish, 2004).

Techniques from *information retrieval* (IR) can be particularly useful for analyzing text data on SNS pages, such as wall posts, comments on posts or pictures, etc. Thus content analysis (Ricardo, 1999) can be employed to determine the kinds of subjects that people are discussing and sentiment analysis (Turney, 2002) can be used to determine the mood of the people in various conversations. Since it is known (Moreno et al. 2009) that adolescents disclose information about risky health behaviors like smoking, alcohol use and unsafe sexual behaviors on these social networking websites, IR and related techniques can be used to automate the processing of discovering such of markers from these websites.

As described in the previous section, social network analysis can be used for various purposes like determining central and potentially influential people in the network. These people can then later be targeted to disseminate valuable information to other users. *Social network analysis* can be combined with other techniques like IR

based techniques and *behavioral mining* to determine not only the prominent individuals in the network but also groupings of individuals and the kinds of topics that they discuss (Blei et al 2003). Additionally these techniques can be used to determine individuals as well as groups who may be engaging in risky behaviors. Another feature in social networks that has grown in prominence in recent months is *analysis of social network based games* (e.g. Zynga). Unlike other types of on-line interactions, these games denote a stronger type of relationship (See Appendix). Behavioral mining can thus be used to link characteristics of users within the game to real world characteristics.

3.4. Trust prediction and propagation model for adolescent health

Trust between individuals can be defined in different ways e.g., direct trust between two parties (i.e. trust toward a directly connected user), indirect trust in the form of reputation, and mediated trust (i.e. trust toward an indirectly connected user). In some cases information for predicting direct trust may be available for only a subset of actors who are directly connected in a social network. For the rest of the people in the network, indirect trust can be propagated in the network and has to be inferred indirectly through a chain of directly trusted user. In addition to the mechanisms and algorithms for inferring trust, researchers have made other observations that can be used to establish trust. Thus it has also been observed that trust between people can be determined by the similarity between them which in turn depends upon their previous activities and interests (Ziegler and Golbeck, 2004). This implies that homophily is at play in establishing trust i.e., similar people are likely to establish trust with one another.

In this section, we briefly introduce the trust model classification scheme that taking into account the characteristics of the current computational trust prediction models (Zhang et al, 2004).

- *Subjective Trust vs. Objective Trust*: If an entity/user's trust value is related to the quality of service (or interaction) it provides to others and the quality of a service can be objectively measured, the entity's trust value for the service is objective trust. For example, if a blogger provides specific health information in his website, the quality (or accuracy) of the information can be evaluated by outside professionals. In some cases, the quality of service (information) cannot be objectively measured. The property of trust personalization states that trust is inherently a personal opinion. Since trust is established based on history of personal direct interaction with a target user and two people can have very different opinions about the same target, it is meaningful to predict personalized trust for each user in many cases.

- *Transaction-Based vs. Opinion-Based*: Some trust models rely on the information of individual transactions to evaluate an entity's trust value, while others use opinion information. Here, transaction information from A to B means various types of interaction information like transaction type (e.g., downloading file/image, writing comments), the time of the transaction, the quantity of the transactions (e.g., total size of download files, the number of images downloaded, the number of written comments). In addition, it may capture the feedback on the transaction; feedback is a statement by a user about the quality of a service in a single transaction. Opinion information is a user's general impression about the target entity. The opinion is generally derived from feedback on all the transactions. In our target social network services such as Facebook, MySpace and Twitter, transaction information is more common. Zhang et al (2004) argue that while a transaction-based trust model does not always need detailed information of every transaction (especially feedback on the transaction) and does rely on static information (e.g. total number of positive/negative interaction, the number of transactions during a certain period), it needs more information compared to opinion-

based trust model in order to be effective and thus incurs high computational cost.

- *Global trust vs. Local trust*: Global trust is computed from all transactions or opinions from all the members of the community and thus it represents an aggregated opinion of an individual user *A*. Local trust value on the other hand is a personalized value for each entity/member of the community with respect to more personalized information such as direct experiences, information from directly connected neighbors etc. The local trust value is established on an individual user *A* from the perspective of another individual user *B*. Although a global trust model lacks personalization, it is also useful to globally identify the most influential users as well as local trusted users.

Many trust models assume the property of trust transitivity (Golbeck and Hendler, 2006); if *A* trusts *B* and *B* trusts *C*, then we may be able to infer that *A* trusts *C*. Referral models or trust propagation models usually adopt the trust transitivity property to infer indirect trust value among users. However, trust is not perfectly transitive; the level to which *A* trust *C* would be less than the level at which *B* trusts C. The level of trust is expected to degrade along a chain of networks. TidalTrust is one of the popular trust propagation method proposed by Golbeck (2005). Thus given a social network where a subset of the actors has given trust ratings, TidalTrust propagates trust based on the shortest path between the pair of people for whom the trust has to be determined. The predicted trust is a weighted combination of trust and ratings along the trusted path. In other words if there is a social network where people have specified trust and rating information for products or any transaction, then this algorithm predicts trust by finding the shortest paths between the people for whom the trust has to be predicted and then for each person in the path determines how much they trust the next person in the path and then computes a weighted average which becomes the predicted trust.

Although assessing trust for various online social networks is not a simple task and different methodologies and strategies need to be developed, the basic underlying principles are similar. First of all, direct experience with a connecting user is the most important factor for trust evaluation. A history of interactions can be evaluated in terms of the quantity and quality of interactions to access trustworthiness. Therefore, the lack of necessary direct interaction will negatively affect trust prediction. In order to resolve this issue, many researchers approach this problem by combining direct experiences and witness experiences such as recommendation or evaluation of friends. Aside from direct experiences and witness experiences, social connections and interaction through a trust network via indirect connections are also important factors. This is so because there is a higher chance of trusting the word of someone who is a friend or at least socially connected to one. On the other hand this factor may not be the best way to predict trust to a user but make it does make it possible for a user to connect a lot of trustworthy users without previously having any direct interaction. Another important thing that we may consider is that time can play an important role in trust evaluation. When historical information is used for evaluation, recent events or interactions could be more relevant (Sai et al. 2008). Therefore, the recent history and evaluation have to be considered more than the overall history and detecting the change of interactions and the evolution of trust might be useful.

In the appendix 1, we provide more techniques and models which are used to predict trust, propagate trust in social networks and also infer trust.

3.5. Analyzing the relationship between social interaction and personal health behaviors

In this section we discuss what kind of interactions can be studied in social networks like Facebook, MySpace and Twitter from the perspective of positively

impacting personal health decisions.

Case Study: Facebook

As described above (Section 2), the manner in which people live their lives and their daily habits are greatly affected by other people that they know and the activities that they perform. Thus, in order to study the behavior of people in an online setting and address questions like how does this setting impacts their habits and health related decisions, it is important to look at not only their activities but also the activities of their "friends" on the social network, in this case Facebook. Figure 3.5 shows an Entity Relationship Diagram of the salient features of Facebook. The figure shows that a user provides various types of information about himself or herself (e.g. name, location, contact information, etc in top row). This information can only be edited by the user and cannot be commented about by others. Any other type of information that the user provides in the form of pictures links, wallposts or notes about physical activities, smoking, drinking, etc can be commented upon by others. Additionally the users have the option to chat with others in real time.

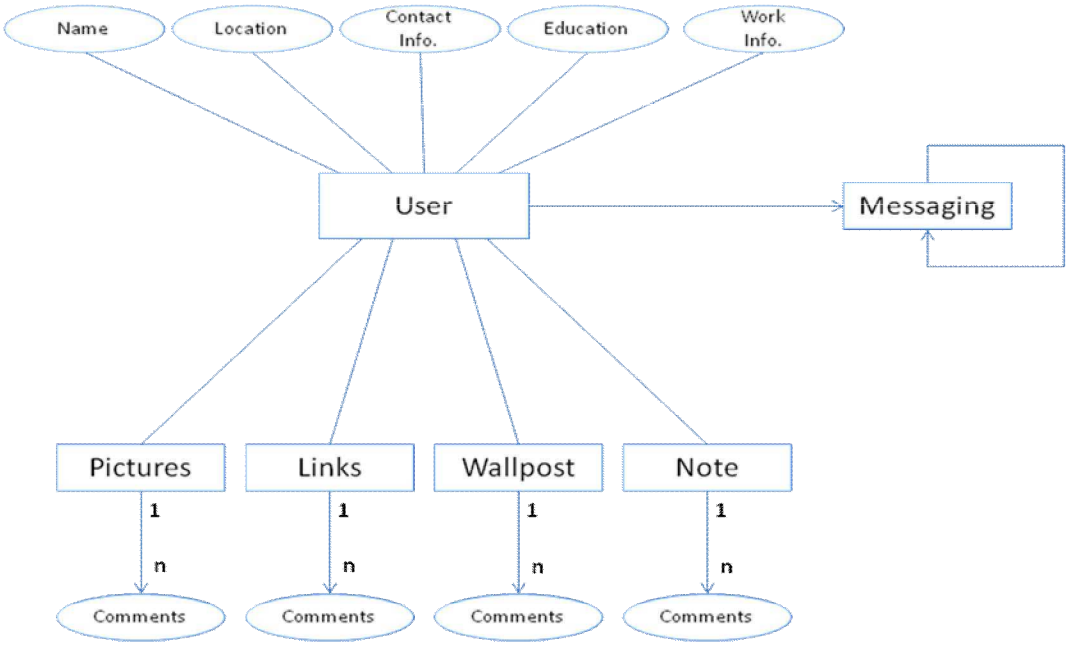


Figure 3.5. Entity Relationship Diagram showing the major salient features of Facebook

Information from these sources can be used to determine various types of health markers e.g., images and accompanying text in the profile can be used to determine if the users engage in risky health behaviors. Since information is also available about other people that a person may be "friends" with, a simultaneous analysis of different people can be done to determine if a person is influenced by other people in engaging in any of these activities. Alternatively one could also determine if there is unhealthy behavior involved i.e., is it the case that people engage in a certain type of risky and non-risky behavior group (smoking group, healthy dieting group, regular workout group) when they are friends.

Case Study: Twitter

In Twitter anyone can say almost anything to anybody in brief status updates. Like most social media services, Twitter allows users to post (tweet), subscribe (follow), share (retweet) or reply to as many twitters as they like. When users start to use the twitter service they create an account and a home page is created where a simple profile is provided which contains information like name in twitter, physical location, web blog that they own, biography and self-describing profile in text (hobby, jobs, interests and so on). Since Twitter asks users to fill out very basic information in self-report profile, to determine the correct age of a twitter user from the profile would be one of the challenges in our research. Twitter users may leave indirect evidence of their age in other sections of Twitter including free text (i.e. "I'm a 16", "I'm a sophomore in X high school, or "I want to go to college next year"), or original tweet or replied tweet from friends that shows younger age (i.e. message of a 14-year-old birthday party).

After account creation, a user gets a twitter handle '@user_account' and then they can post own their own messages which are less than 140 characters long, can include useful links and hash tags like key topics of the message. When a user follows other users, every tweet created by a user is shown in the follower's twitter home page along with any replies (See Figure 3.6).

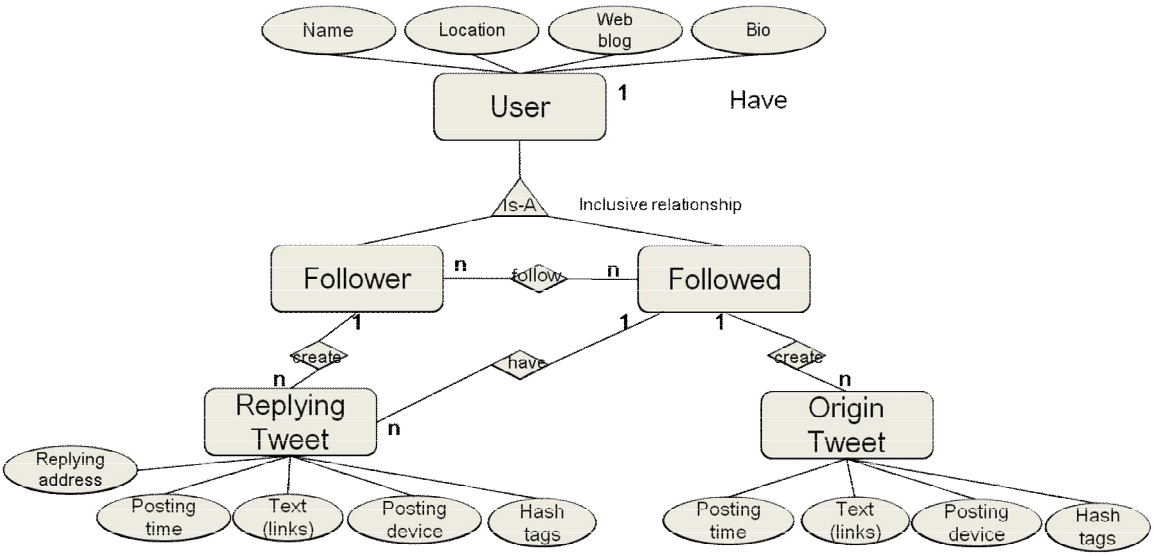


Figure 3.6. Entity Relationship Diagram showing the major salient features of Twitter

Twitter service broadcasts each tweet to all followers every time a tweet is created, but a replying tweet is sent to a specific user. Thus the number of replying tweets toward the twitter that is being followed can be an index of how actively a user follows another user. Since users talk about a wide array of topics ranging from professional knowledge to personal daily event or life, content analysis techniques are useful in discovering topics or main interests in an origin tweet or a replying tweet.

To make twitter more useful for a specific objective like finding a leader in a certain topic or searching and keeping up with various topics, there are already several useful tools that are available for free (Jantsch, 2009):

- Twellow: http://www.twellow.com is like a twitter phone directory that sorts people by industry. This is a great way to find people in professions of interest. A profile in Twellow shows simple information like twitters and includes the number of followers that they have. From this information we can infer industry

leaders by first finding out their followers who are interested in the industry as well as leaders who have high influence.

- Twitter Search: http://search.twitter.com is a simple tool provide by Twitter. Advanced search allows a person to filter all content by keywords. Thus one can find users who have interest in the topic and monitor what people are saying about the topic or issue. Using the RSS function, one can subscribe to all the searches one has set up and determine a useful topic.
- OneRiot: @oneriot is a bookmarking site for twitter. Users share tweets that contain URLs to web pages and this site keep track and returns search results based on topic. Using this tool, one can discover which topic a tweet or replying tweet containing URLs is talking about. This tool can supplement the results of a content analysis of the tweet.
- Crowdeye: @crowdeye is the only twitter search engine that returns the results from tweets and retweets including graphs and charts.
- Topsy: @topsy shows statistics including a collection of twitters by volume for each topic when we search a specific topic. It is useful to fine people who are very active around the topic or who are the best retweeters.

Our main objective for extracting data from Twitter is to discover conversation topics, who's influencing what topics, who's saying what and how often, and trends around a topic, where the main topic being discussed is health behavior of adolescents. It should also be noted that adolescent may not directly express keywords that we are interested in such as drugs, smoke, alcohol, diet, workout, sex. Researchers will therefore need to use a wide range of search terms and keep abreast of colloquial language popular with youth.

Here, we mainly consider the most popular social networking sites, "Facebook" and "Twitter" for data collection. But, other social networking sites such as Flickr and socialization game Mafia Wars and Farmville can supplement the information about social interactions among adolescents and their health behaviors collected from Facebook and Twitter. In the appendix 2, we introduce Flickr and social interaction game sites.

Section 4. Application of Trust Networks for Adolescent Health

As online social networks have created new and effective means of socialization among adolescents and the usage of online social networks has grown exponentially, its impact on adolescent health behavior is also expected to be higher. This new technology especially augments the interpersonal level of influence in the social ecological framework described in Figure 2.1. Influence in this case can be measured with respect to the concept of "trust" between users. Moreover, the tremendous data collection available from online social networks offers new opportunities to study socialization among adolescents in terms of health behavior with a more sophisticated and quantitative model which was previously unimaginable. We posit that use of sophisticated and advanced computational models to analyze the detailed behavioral data is going to drastically change the way of studying social science questions, and enable exploration of new questions such as:

(i) *How does social influence in online social networks impact adolescent health, and how can this impact be assessed?*

Once the source of influence and risky behavior factors in online social networks have been identified, we can address multiple questions in order to understand adolescent health behavior and the impact of social influence on it through trust

networks; How much is an individual user influenced from the influential and trusted neighbors?, How can we classify users into healthy groups and risky groups?, How much and how fast is the risky information transferred to one's friends?, and How does the level of trust among users affect to improve health behavior?

(ii) How can social networks be used to proactively improve adolescent health by increasing positive social influence or reducing negative social influence on adolescents?

By classifying healthy behavior groups or risky behavior groups with community detection models and selecting the most influential users in each group, we can increase positive influence or decrease negative influence through the trust network (e.g., by encouraging positive key influential users to adopt healthy behaviors, or by cutting the links from the negative key influential users to the rest of the networks).

(iii) How can the impact of social influence for adolescent health be validated?

In order to validate the effectiveness of intervention or trust assessment, we can follow up the change of user behavior such as reducing risky behavior or increasing the interaction with healthy behavior groups. The causal relationship between a level of trust and change of health behavior can be also evaluated with a type of longitudinal analysis.

For understanding the influence of social networks on adolescent health and maximizing the impact of social influence through trust networks, we briefly introduce terms such as "Key influential user identification", "Community detection", and "Link prediction" in this section.

4.1. Identifying key influential users in an online social network

In this section, we introduce some popular approaches to determining which users have significant effects on the activities of others in a network. The importance of identifying influential users on a social networking site has been recently highlighted by Google's ad targeting strategy at MySpace (Green, 2008). The main idea is that targeting ad at some members who get more attention from many users, is much more effective compared to targeting at individuals with certain characteristics.

Influential users, or opinion leaders, are highly informed, highly connected and well respected individuals who exercise informal influence over others (though not necessarily leaders in any formal sense of the term). Katz and Lazarsfeld (Katz 1955, Lazarsfeld 1968) did pioneering work in studies that showed that people are influenced more by each other than by the media with respect to decision making. Their theory stated that there is usually a small group of influential users that influence other people's with respect to other people's decision making. Thus in terms of information flow, information flows from the media to these influential users and then from these influential users to the rest of the population. There has been extensive work on the role of influential users in the diffusion of information in a variety of fields (Coleman, 1966, Rogers 1995, Valente 1995, Van den Bulte 2007).

With the degree of trust or the level of strength in relationship among users, the following approaches can be applied with little modification to identify influential users who effectively propagate correct healthy behavior information to the network or significantly affect other users' health behavior in various ways.

(1) PageRank

PageRank (Brin and Page, 1998) is a link analysis algorithm used by Google Internet search engine to measure each web page's relative importance within a hyperlinked set of documents. In essence, Google interprets a link from a page A to a page B as a vote, by page A, for page B, and then gives a more highly voted page higher value. Simply the PageRank of each page is based on the quality of inbound link (vote) as well as the PageRank (i.e., the importance level) of the page providing the links as follows:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in I(p_i)} \frac{PR(p_j)}{\left|O(p_j)\right|}$$

where N is the number of pages

$I(p_i)$ is the set of page that link to p_i

$O(p_j)$ is the set of outbound links on page p_j

$\left|O(p_j)\right|$ is the number of outbound links on page p_j

Given a trust function with our dataset which will be a transaction-based trust model, an individual user p is considered as a webpage p and a trust link whose estimated link strength is higher than certain amount from user p_i to user p_j is used as inbound link to p_j in the PageRank model. Then a user who is globally trusted by many highly trustworthy users eventually receives high PageRank value through multiple iterations. The final PageRank values are used to rank all users based on trust based influence in a social network. Afterwards we can select the top n high influential users with high PageRank values to provide health information to the network. The activated health information will be propagated from top n high influential users to their directly connected users and from the influenced users to the rest of whole networks as time goes by.

(2) HITS

The main idea of HITS (Hyperlink-Induced Topic Search) is also a link analysis algorithm that rates Web pages developed by Kleinberg (1999). It measures Hub and Authority of a page. The main idea of HITS is to regarded the presence of a hyperlink from page i to page j as granting page i's authority to page j. It ranks web pages by exploiting the mutually reinforcing relationship of hubs and authorities by hyperlinks as follows:

$$a(i) = \sum_{j \to i} h(j)$$

$$h(i) = \sum_{i \to j} a(j)$$

$a(i)$ is authority of page i

$h(i)$ is hub of page i

The relationship is exploited by iterative algorithm that updates numerical weights for the page. A page with high *h(i)* score is considered to be a page that connects to many good pages with high authority. A page with high *a(i)* score is considered to a page that is linked by many good hub pages.

Basically the weight of web pages depends on the structure of hyperlinks, but the influence in a social network heavily depends on a level of trust ("the strength of links between users"). Therefore, given a trust network with our dataset, we can use a level of trust *t(i,j)* as a probability of that a user *i* will be influenced by a user *j* at a certain context as follows:

$$a(i) = \sum_{j \to i} t(j,i)h(j)$$

$$h(i) = \sum_{i \to j} t(i,j)a(j)$$

$a(i)$ is authority of page *i*

$h(i)$ is hub of page *i*

$t(i,j)$ is a degree of trust from user *i* to user *j*

To illustrate the application of this method in our context consider the example of Twitter where users with high hub value would be those who know many globally trusted users who send helpful health information. Even though they may not provide health information to other by themselves, they follow many global trustworthy users with high authority values. Users with high authority value definitely broadcast many high quality health information tweets and would be followed by many hub users.

(3) EigenTrust

Eigentrust algorithm (Kammar et al. 2003) is based on the notion of transitive trust. If a user *i* trusts a user *j*, then the user *i* would also trust user *k* trusted by *j*. Based on the transitive property of trust, a user *i* will ask its trustworthy users *j* if they know about user *k* and weigh responses from the trustworthy users by a level of trust a user *i* places in them as follows:

$$t_{il} = \sum_j c_{ij} c_{jl} \quad \text{(for each user i, } \sum_j c_{ij} = 1\text{)}$$

t_{il} is a predicted trust value from user *i* to *l*

c_{ij} is a local trust value from user *i* to *j*

Here, if a user has a previous direct interaction (i.e., any kind of transaction such as downloading photos, writing comments, following someone) with any user *j*, a user evaluates a local trust value on user *j* based on its experience. When we assume that we have all c_{ij} values for the whole network in the form of a matrix *C,* the final trust matrix *T* for the whole network can be obtained after *k* iterations (propagation) where a large value of *k* make the trust matrix *T* converge (i.e., $T = C^k$)

Guha et al. (2004) developed a framework for trust propagation for datasets where both trust and distrust information are available in a social network. They show that a small number of expressed trust/distrust per users can increase the prediction accuracy of trust between any two users compared to only using trust information. They compare three different propagation models to see how trust and distrust propagate together: (a) Trust Only Propagation model: In this case, it ignores distrust completely, and propagates trust scores. (b) One-Step Distrust Propagation model: Assume that when a user distrusts someone, they also discount all opinions from the user; thus distrust propagate only a single step at the last time after propagating repeatedly. (c) Propagated Distrust model: trust and distrust both propagate together k-times. In their experiments,

One-Step Distrust Propagation model, $F = T^k \cdot (T - D)$ (where F is a final trust matrix, T is a trust matrix, D is a distrust matrix) shows the best performance in trust and distrust prediction. In our dataset, trust is interpreted positive influence on health behavior and distrust is interpreted as negative influence on health behavior. The positive and negative influence is measured with both transaction-based information (including feedback of the transaction) and opinion-based information that is evaluated from feedbacks (See in section 3.4). The positive words and negative words in comments can be used to identify if the feedback is mostly positive or negative using content analysis techniques. With trust propagation, it is possible to predict the most positive influential users and the most negative influential users in a social network.

(4) Influence maximization model

In certain domains like marketing, the literature has traditionally focused on finding the most influential users. This translates into the problem of finding the most influential customers. The main question is that if we can try to convince a subset of individuals to adopt new product or innovation, and the goal is to trigger a large cascade of further adoptions, which set of individuals should we target? Thus, Domingos (Domingos 2001) presented a network based model of customer value where the value of the customer is determined in terms of not just their activity with respect to their previous interactions with the service or company but also with respect to how much influence they exert on other people in the subscription network or the shopping network. Kimura et al. (2009) address the problem of ranking influential nodes in social networks by estimating diffusion probabilities from observed diffusion data using the popular independent cascade model. They formulate the likelihood for information diffusion data (i.e. a set of time sequence data of active nodes) and propose an iterative method to search for the probabilities that maximizes the likelihood. Their model can predict the high ranked influential nodes much more accurately than conventional heuristic methods.

When we select a set of target group for health information advertising or campaigns, finding optimal number of high influential users is critical in maximizing the effectiveness of the target advertising given cost, time or other constraints. Thus, along with Domingos' study (2007) and Kimura's study (2009) we should consider the probability of that each user accepts the advertisement with interest, the diffusion rate of transferring observed data, as well as the amount of influence to the directly connected friends and the whole network.

4.2. Extracting healthy or unhealthy youth communities for health advertising target

Community extraction refers to identifying groups of similar people in a social network. There is a massive body of literature on community extraction in Computer Science, Sociology (Wasserman, 1994) and Physics literature (Newman, 2004). Although many techniques have been developed for community extraction, there is no universal criterion for evaluating the results from community detection because of lack of ground truth in most domains. The techniques which are chosen are thus based on what makes the most sense in a certain domain i.e., the domain experts define what constitutes a "good" community of people in a social network. In the context of the current problem of adolescent health, the problem of community detection would involve determining groups or communities of youth which exhibit similar type of health or unhealthy behavior e.g., being engaged in drug abuse, unsafe sex, smoking or a combination of these. Once such groups have been identified they could be targeted for advertisement campaigns or given certain incentives to adopt healthy behaviors.

One of the oldest techniques for graph partitioning (mathematics consists of dividing a graph into subgroup) is called the Min-Cut method. It was invented to solve the problem of load balancing in parallel computing which is the problem of roughly equally dividing tasks in parallel computing (Karapis, 1999). The main idea behind this technique is that the nodes within a partition (a grouping of nodes) should be strongly connected to one another as compared to nodes across partitions. Another well known method is the Garvin-Newman method (Newman, 2004) which determines communities by removing edges between nodes until one gets disconnected components and these are determined to be communities. Newman's modularity (Newman, 2006) is one of the most widely used methods for detecting communities. It detects communities by looking into structures amongst nodes in a network and comparing that with the structures that would form if the network was just a random network. One weakness of this method is that since it is a global optimization method, it fails to capture small communities in global networks. It should noted that all the methods described so far take into account only the graph structure of the nodes but not the type of interaction that may be going on in a network.

Another area of community extraction which is relevant to the current application is topic based modeling combined with community extraction. Topic models are a class of generative models which extract communities from corpus of documents based on co-occurrence of words, authorship and other criteria within the corpus. Thus the Group-Topic model (GT) discovers groups of entities such that within group entities show similarities with one other as compared to entities which are outside the group. The author-recipient-topic (ART) model (McCallum, 2005) extracts topics based on communication between people but it cannot be used to extract communities. In the community-user-topic (CUT) models (Zhou, 2006), a community is modeled as a joint distribution of topic distributions and user distributions and can thus extract communities based on latent topics that may be discussed in the community. The CART (Community-Author-Recipient-Topic) model for community extraction (Pathak, 2008) can be used to extract communities based on the type of communication and the topics being discussed in a group. It also allows people to belong to multiple groups. Such models can be applied to social network data where communication information, e-mail exchanges, message exchanges and even image information is available. This information based on information about the activities of users and not just the links between the users can then be used to extract communities of people and then analyzed based on what kind of information transactions are happening there e.g., are patterns of risky behavior recurring repeatedly e.g., discussion about drinking, images of alcohol etc.

4.3. Predicting links and trust values to recommend adolescent trustworthy friends or a healthy community

Link prediction (Liben-Nowell, 2003) refers to the problem of predicting links that may form between entities in a network. The network in question could be a social network between different people or it could be a network of people and items. In our context, link prediction can be applied to predicting the links that can be formed between two people in a network, the strength of interaction between them and how that information can be used later on to select people who can influence others. Hasan et al (2006) posed the link prediction problem as a supervised learning problem and identified a set of features that can be used to predict future links between nodes in a graph. Out of all the features they discovered that topological features of the graph like shortest distance and clustering index were the most helpful in determining future links between the nodes. Adafre et al (2005) addressed the problem of finding "missing" links in Wikipedia by first narrowing down the set of links that have to be analyzed for

predicting the formation of links based on graph clustering and then using the characteristics of subset obtained from graph clustering to make predictions. Huang generalizes the clustering coefficient to describe a graph topology based method to predict links (Huang, 2006). Many other methods have also been proposed which use Markov Networks (Taskar, 2003), Statistical learning (Popsecul, 2003), probabilistic learning models based on Markov Random Fields (Wang, 2007) and so on.

Xiang (2009) divides the techniques for link prediction in three broad categories: node-wise similarity based methods, topological similarity based methods and probabilistic methods. He also notes that link prediction has been applied in a number of domains like social networks, recommendation systems, citation networks, protein interaction networks etc. While there is a large body of literature on the problem of link prediction (Xiang, 2009), predicting the strength of interaction between two nodes in a network or the change in the interaction strength has not really been studied in detail. This can be used to predict not just the likelihood of people connecting in the future but how strong the link is going to be and thus how influential one person can be with respect to another person. There are some exceptions where the link prediction task incorporates the history of previous interactions between the nodes in order to make predictions about the future links.

The problem of trust link prediction in the context of health applications can be considered complimentary to the problem of determining influential users in the network as above. Suppose that we could predict the likelihood of a link between two users in the future and it was also possible to predict the influence of the same user with respect to others, then we could target specific individuals in a community to spread health related information to others so that the others are likely to listen to them. This problem is different from just finding the most popular person in a group because we are interested in finding not only the person from whom the others are likely to take advice but also the person to whom they are likely to connect and listen in the future.

4.4. Case study

In this section, we provide a case study of a specific public health problem, cigarette smoking among adolescents and describe in some detail how we can apply our proposed approach to improve adolescent health behavior. First, we explain cigarette smoking behavior within a social ecological framework and then flesh out our computation model from data collection to modeling, to interventions, to validation.

The Social Ecological Model and cigarette smoking

The social ecological model can be readily applied to specific adolescent health behaviors, such as cigarette smoking. At the intrapersonal level, individual beliefs, expectancies and attitudes are most directly linked to that individual's smoking behavior. For example, the belief that cigarette smoking is addictive, cool or bad-smelling; or the expectation that smoking will make one more popular or mature, or cause one to be punished if caught, contribute directly to the uptake of this behavior. These intrapersonal factors are often influenced substantially by factors at the *interpersonal level*. Interactions with peers can include talking about positive or negative attributes of cigarette smoking; modeling the behavior (and its consequences); providing cigarettes to one another; teaching how to inhale, obtain cigarettes, or hide evidence of smoking; or pressuring or teasing one another to try cigarette smoking. These real-life interactions can be mirrored in the on-line world as well, through posts, comments, photos and videos on individual profile pages. Family members are another well-established

interpersonal influence on smoking behavior among youth. Having parents or siblings who smoke also contributes to beliefs about this behavior and its likely outcomes, and greater exposure to smoking engenders more favorable attitudes. Even more distant family members, such as an aging grandparent who has smoked for years and maintains good health, can send powerful messages. Likewise, teachers and others can provide more didactic information aimed at increasing knowledge about the health risks of cigarette smoking, and health education further aims to change attitudes towards smoking (focusing more on immediate social consequences). This, too, has application in the on-line setting, where information about tobacco can be provided by a health care provider, educator or authority figure within the social network. *Organizational/institutional factors* regarding cigarette smoking can include policies such as smoke-free public places and convenience-store policy on the placement of cigarettes, and statewide programs like Minnesota's erstwhile Target Market campaign (Target Market, 2010). *Societal factors* are also at play for adolescent cigarette smoking. Large scale public policies such as legal age to purchase tobacco products, cigarette advertising, or cigarette pricing and taxation affect access across a large swath of society. Entertainment media plays a critical role in shaping social norms around cigarette use, portraying this behavior on-screen as cool, rebellious, dangerous or sexy; and celebrities do their part to re-enforce or counter these messages in their own off-screen behaviors, which are captured and delivered to the public via other media (e.g. gossip magazines). Certain cultural values can also speak to the desirability of cigarette smoking. For example, the American ideals of individual rights and personal freedoms are easily applied to the adoption of high risk health behaviors like tobacco use.

As with other health behaviors, each level of influence dynamically affects other levels of influence, as well as affecting the behavior directly. For example, smoke-free public places not only reduce opportunities for individual young people to smoke cigarettes, they also reduce the visibility or frequency of smoking among peers and family members at the interpersonal level. Such a policy may also force smokers to remove themselves from a setting in order to smoke, which changes the expectancies of the behavior at the intrapersonal level (i.e. instead of fitting in, they are ostracized), as well as the social norm of acceptability at the broader societal level. Likewise, a family member dying of lung cancer can change intrapersonal attitudes substantially enough that an individual works towards policies and programs which restrict tobacco use at an organizational level; writ large, such changes become part of our cultural values regarding smoking.

Modeling "Trust" and its application

Step 1. Data collection

MySpace was a pioneer in social networking websites and is currently the second largest social networking website. While MySpace's various features for privacy control and restriction of information are set by users, most information on MySpace is publically viewable and publically available. Thus it is possible to collect profile information about users of MySpace by building programs called "crawlers" that can identify and download information from MySpace. Such information can be used for adolescent health improvement. Here we consider the case of smoking among adolescents and how information MySpace can be used to potentially identify such behavior.

A wide variety of information about users in the form of text, images, tags, links, icons etc is available on MySpace. We note that the data should be segmented by

demographics i.e., gender, race etc because it could very likely be the case that different smoking related behaviors are exhibited by different demographics and may even require different interventions. This kind of data is readily available in MySpace in the main profile information section. Information can either be extracted directly from the age field in the profile or indirectly computed based on other information on the website e.g., users may indicate their date of graduation or the high school that they are currently attending. This information can be used to determine the age of the user.

As described previously there are types of markers which can be used for inferring smoking behavior. Here we describe the type of data that can be used and techniques that can be applied to extract the relevant information from the data.

Content Analysis of Messages: These are the messages that the users write on each other walls. By analyzing the content of the messages one can determine if the users are discussing smoking related topics, the frequency of discussion point to its relative importance, the recency of the messages can tell us information about its importance over time etc.

Semantic Analysis of messages: In some contexts just looking at the content of the data is not sufficient because the users may be talking in slang, codewords or brand names. In such cases a semantic analysis of the message can reveal relevant information from the messages.

Image analysis: Another obvious marker of smoking behavior is the presence of cigarettes or cigars in images that the users have posted or even tags that may hint at usage of cigarettes.

Affiliation/Group(s) Membership: Users may have joined groups or communities which may imply preference for certain brands of cigarettes or a tobacco company.

Step 2. Modeling the degree of influence among users

Given the popularity of MySpace among adolescents, the social networking site will provide a new opportunity to identify, screen and ultimately intervene in teenagers who are considering or engaging in smoking with high probability (Moreno et al. 2009). Using the displayed information on smoking and its related topics in MySpace, we try to measure a level of engaging in smoking topics for each user and the degree of influence between connected friends in terms of smoking behavior. The basic assumption of this modeling is that users who frequently display smoking related blogposts or photos are more likely to be associated with smoking in the real world and at least represent their attitudes toward smoking. In addition, a user who is exposed to the posts and photos by reading and commenting on the posts of a friend can be influenced, and the influence may spread to the reader's friends.

2.1. A function for measuring a level of affinity for a smoking topic

In MySpace, adolescents communicate about multiple topics, some of which are commonly popular and some are prevalent in certain groups. To measure the prevalence on smoking behavior topics in a user's MySpace webpage compared to other common topics for teenagers, first we find major topics which are popular and discussed a lot among adolescents in MySpace, and discover possible sub-topics and terms for each topic. Here, we consider "smoking" as one of the major topic and identify some key terms and sub topics which represent smoking behavior involvement like tobacco, cigarettes, cigs, lights, menthols, puff, Marlboro, etc.

We briefly introduce four important factors that a function for measuring a level of affinity for a smoking topic considers:

- Frequency of writing posts and uploading photos in each topic in user i's blog :

$F(u_i,t_1), F(u_i,t_2), \ldots, F(u_i,t_n)$

where $F(u_i,t_n)$ is the number of posts and photos of user *i* in topic *n*

- The distribution of affinity levels over all major topics in user i's blog:
  $P(u_i,t_1), P(u_i,t_2), \ldots, P(u_i,t_n)$

  where $P(u_i,t_n) = \frac{F(u_i,t_n)}{\sum_{1}^{n} F(u_i,t_n)}$ is the probability of user *i* engaging in topic *n*
- Global frequency of writing posts and uploading photos in each topic:
  $F(t_1), F(t_2), \ldots, F(t_n)$

  where $F(t_n)$ is the average number of posts and photos in topic *n* over the adolescent population
- The global distribution of affinity levels over all major topics:
  $P(t_1), P(t_2), \ldots, P(t_n)$

  where $P(t_n) = \frac{F(t_n)}{\sum_{1}^{n} F(t_n)}$ is the average probability of adolescent engaging in topic *n*

The frequency of writing posts and uploading photos in a smoking topic is the basis for assessing how much user i engages in the smoking related topic. In addition, the ratio of the frequency in a smoking topic to total frequency over all topics and to the average frequency from all adolescent are more important factors to represent how risky user i is in terms of engaging in smoking behavior. Therefore, the level of affinity for a smoking topic of user i (i.e. the risk level in terms of engaging in smoking behavior) is a function of the above four factors as follows:

$$S(u_i,t_{smoking}) = f\left(F(u_i,t_{smoking}), P(u_i,t_{smoking}), F\left(t_{smoking}\right), P\left(t_{smoking}\right)\right)$$

where $S(u_i,t_{smoking})$ is the level of engaging in smoking topic of user *i*

2.2. A degree of influence between users

We derive a general model to measure a degree of influence between two connected users in terms of smoking topics. When user i writes posts or uploads photos about a smoking topic in his/her blog, some users that are friends of user i frequently read and write comments on the posts or photos and other neighbors infrequently. If user j frequently visits user i's blog and writes comments which support smoking behavior like "You look cool and hot!" on posts and photos when user i writes posts and uploads photos on smoking topics, the probability that user j responds user i's posts with smoking supportive comments on any given time is much higher. Then, a degree of influence on user j' s attitude toward smoking by user i also increases .

$$p(u_j, u_i, t_{smoking}) = \frac{C(u_j, u_i, t_{smoking})}{F(u_i, t_{smoking})}$$

where $p(u_j, u_i, t_{smoking})$ is the probability of user *j* responding user *i*'s posts or photos in topic *smoking* with smoking supportive comments

$F(u_i, t_{smoking})$ is the number of posts and photos of user *i* in topic *smoking*

$C(u_j, u_i, t_{smoking})$ is the number of writing smoking supportive comments on posts or photos in topic *smoking*

In addition to the probability, user i's level of affinity for a smoking topic (i.e. the risky level in terms of smoking behavior) also affects a degree of influence in terms of smoking topics. As the more actively user i writes posts on smoking topics, the more influenced user j is when $p(u_j, u_i, t_{smoking})$ is high. Therefore, the degree on influence between connected users is a function of $p(u_j, u_i, t_{smoking})$ and $S(u_i, t_{smoking})$ as follows:

$$I(u_j, u_i, t_{smoking}) = f\left(S(u_i, t_{smoking}), p(u_j, u_i, t_{smoking})\right)$$

$$e.g.\ I(u_j, u_i, t_{smoking}) = S(u_i, t_{smoking}) \cdot p(u_j, u_i, t_{smoking})$$

where $I(u_j, u_i, t_{smoking})$ is a degree of influence on user *j*'s attitude on *smoking* topics by user *i*

Step 3. Identifying high influential users

Using one of the popular algorithms to identify key influential users, HITS (See section 4.1.(2)), we can identify a small group of influential users that are actively engaged in smoking topics (high authority users) or connect many high influential and active users in the topic (high hub users). When the HITS algorithm runs, the adjacency matrix which represents the link connection from user i to user j with 1 (trust) or 0 (non-trust) is necessary. Since we have a degree of influence on user j by user i , $I(u_j, u_i, t_{smoking})$, we replace the adjacency matrix into an influence matrix filled with a degree of influence. After the convergence of authority and hub values for all users, we select the top n users based on authority and hub values.

- The users with high authority values: The user writes a large number of posts and uploads photos about smoking and many of posts and photos are very popular and commented by neighbors
- The users with high hub values: The user may not actively write posts or uploads photos about smoking in their blogs, but knows many active friends in terms of smoking topics and writes comments on their friends' posts and photos with high interest

Step 4. Identifying smoking behavior community detection

As described previously, smoking is not just an individual activity but rather a social activity where many people smoke because of social reasons i.e., peer pressure, social norms, acceptance of smoking as a positive trait in a social group, etc. Thus complementary to the task of identifying people who smoke or who can influence others to smoke is the task of identifying groups or communities of people who smoke and thus positively reinforce such behaviors in each other. We can pose this problem in the form of topic based community detection. The main idea behind topic-based community

detection is to look at the kind of discussions that people are having with one another and also simultaneously look at other profile information like interests, images, icons etc and then based on features extract communities of people who have certain common features, in the current case these would be text related to smoking, smoking indicators in images, friendship information between people who are smokers etc. Notice that this technique is different from extracting communities of people just based on the "friend" information available about them in MySpace i.e., a person may have many friends but only a subset of these friends actually smoke and thus a "community" of such smokers should be extracted based only on the smoking markers. After extracting such communities one could also identify individuals who may be susceptible to smoking in the future e.g., individuals who are embedded in a community where a large number of people smoke but they themselves do not smoke but are known to be susceptible to be influenced by others.

Step 5. Intervention strategies and their evaluation

The main objective of modeling the social influences on smoking behaviors and identifying high negative (bad) influential users and smoking groups is to intervene in adolescent attitude on smoking in online and ultimately improve health behavior in real world through leveraging the impact of social influence.

Suppose that we identify top n high authority users or top n high hub users using HITS algorithm in step 3 (Some of users are in both the high authority user list and the high hub user list). One of the recommended interventions is to send the selected users (i.e. high influential and active users) the email that point out that they are highly involved in displaying smoking related topics in their content or are influenced from high risk friends in terms of smoking. The email can also suggest changes such as stopping writing smoking-related posts and commenting their friends' posts. The intervention can also provide links to online or offline smoking cessation programs.

In order to validate how effective the suggested intervention strategy is, we propose to measure the change of user behavior in displaying smoking related topic information in their blogs before and after the intervention.

Validation 1)

$$average\left(\frac{S(u_i,t_{smoking})_{\text{after intervention}} - S(u_i,t_{smoking})_{\text{before intervention}}}{S(u_i,t_{smoking})_{\text{before intervention}}}\right)$$

where $S(u_i,t_{smoking})_{\text{before intervention}}$ is the level of affinity for a smoking topic of user i before an intervention

Validation 2)

$$\sum_{all\ user\ i} w_i \cdot \frac{S(u_i,t_{smoking})_{\text{after intervention}} - S(u_i,t_{smoking})_{\text{before intervention}}}{S(u_i,t_{smoking})_{\text{before intervention}}}$$

where w_i is the authority/hub value for user i $(\sum_i w_i = 1)$

Simply, validation 1 measures the average change ratio of the level of affinity for a smoking topic from all users. The more effective an intervention is, the more negative the average change ratio is which means reducing the engagement in smoking topics. The second validation proposes a weighted average change ratio which places more weights to high influential users' change.

Section 5. Conclusions

We conclude this report by describing some of the technical and ethical challenges inherent in collecting and analyzing health related data and suggesting the role of online social networks in influencing real-world policy.

5.1. Technical challenges

In the previous sections we have described various social networking websites,. Mainly due to privacy concerns described elsewhere in the paper, data from these service providers cannot be directly obtained. Thus, data has to be collected through indirect sources; all the websites that we have described here have their API (Application Programming Interface)s available. These APIs can be used to collect data about the users by building crawlers which can recursively extract data from these websites. Additionally there are open source and commercially available crawlers available which can be used to extract data from these websites. There are however some limitations in this approach e.g., websites may limit the amount of data that can be collected, and the subset of the network that is extracted may not be representative of the universe of data. These limitations can be overcome by using a larger time window for extracting the data and thus incrementally building the dataset. The latter issue can be resolved by concentrating on a smaller subset of the global network e.g., only extracting data from a particular middle school or high school.

Social Network Aggregation refers to the process of gathering information from various social networks and combining them at a single location. This has become an issue of interest because of overlap of membership between different social networks because users who are part of multiple social networks may feel the need to replicate their information at multiple website. Single sign-on systems like OpenID allow the users of social networks to use a single ID to log into multiple social networks.

Since data are being extracted from multiple sources, in this case from MySpace, Twitter and Facebook, there is likely to be an overlap between the various social networks i.e., the same person may have an account on MySpace, Facebook and Twitter and in each case they are likely to provide different types of information about themselves so information from these various sources has to be combined in order to identify that various pieces of information refer to the same person. While it may be relatively easy to do this in MySpace or Facebook since there is usually detailed information available about the person, it would be more difficult to do so in Twitter because of the limited profile information available. In this case other clues like e-mail id may to be used to link the person to a profile on other social networks like Facebook or MySpace.

Additionally on the issue of de-duplication, information about the same person from multiple sources which can be achieved by using similarity functions for profile information that be common to various sources and using social network information to match profiles of people who cannot be matched up by using profile information alone.

As described previously automatic image annotation techniques have not advanced enough to be used in a generalized setting. Machine learning techniques which are described previously may work in limited settings where the pool of images which have to annotated have certain common characteristics. A more scalable approach would be the human computing approach described above. The main obstacle in that case would be building the infrastructure necessary to get people to tag the images without invading their privacy. The main problem here would be that if people are asked to annotate images with an application like Amazon's Mechanical Turk, then having other people annotate the images would violate the privacy of the people whose images are being annotated. The alternative would be have people within the research team annotate a set

of images and then apply a machine learning approach on the annotated images by training a classifier on these images and the labeling the rest. This would thus be a semi-supervised learning problem (Zhu, 2009).

Social Network Analysis (SNA) has been applied on network data for more than 60 years. With the availability of large scale datasets, however much of the software which is widely used for SNA is no longer usable because of scalability issues. In order to conduct network analysis for large scale datasets, many of the SNA algorithms have to be implemented so that they are scalable e.g., in parallel forms that divide a large computation into smaller ones and then solve them concurrently. Porting these algorithms to these environments is a challenge in its own right. Additionally there are other techniques like link prediction, etc, which are not widely available in SNA software but are greatly relevant to our task.

5.2. Ethical challenges

As described before, many users create their own Web pages and post details about themselves: school, age, gender, hobbies, religion and their friendship list in MySpace, Facebook and Twitter. They also link to friends by inviting them and accepting the invitation on the same online social network websites. Then, they are easily able to see what their friends are doing, what are the major interests and topics and how they are thinking and influencing each other. While these online social networking websites become useful tools for exchanging information and for retaining active friendship, there has been growing concerns over breach of privacy in these social networking services.

Many users concern that their personal profiles and posts that they intend to share with their friends can be circulated far more widely than they intend to. So, many sites restrict who can join a site and access a user's information. Who can access user's information, what kind of information and how easily are affected by the search tools that each social networking site offers and the privacy levels set by individual users.

MySpace allows the general public to search its database of members, using search terms such as name, e-mail address, or school. This search system can filter out users based on a particular country or a postal code. If users have not intentionally changed their privacy settings from the default level, the search system can collect their full profiles including personal information such as sexual orientation, occupation, address as well as photos of users and their family and friends list.

However, Facebook allows a more limited search features. Users must register as a member with the site in order to conduct any search, and see the profiles of their friends' networks: the friendship is constructed based on invitation and allowance.

As the privacy issue in social networking service websites becomes more important, the social networking sites receive more pressure to set default privacy setting at a higher level. Then, a user who doesn't want a great deal of detail personal information to be displayed with anyone would have immediately more control. Unfortunately, collecting data only from profiles which are left publicly accessible is likely to bias the sample on known and unknown characteristics.

Although individual users share their information with the public and the search tools can collect the information, there is still possibility to breach a privacy policy that social networking service site affirms. Even data entered in a publicly availably setting were not volunteered in a research context, and using it without specific permissions may violate the principle of research subjects' informed consent. Therefore, the best of way to respect individual users' privacy is to acquire their permission to use data for research purposes, in accordance with appropriate consent procedures.

5.3. The role of online social networks and policy changes

On-line social networking, and all its human and technical features, can be considered within a policy context, and applied to policy change. Social networking sites are unique as the organization, institution, or environment in which social exchanges occur, but, like other organizations, have certain policies in place to govern these exchanges. For example, in order to " [protect] our user experience by keeping the site clean, consistent, and free from intrusive advertising," Facebook.com restricts advertising promoting a variety of health-related products including tobacco, weapons, gambling and pornography. Other restrictions on content more broadly are also in place, but are more limited, such as prohibitions on graphically sexual images. One could imagine additional restrictions which would similarly limit imagery or other content of cigarettes or smoking behavior or linkable and downloadable content such as smoking-related icons, games or quizzes.

Efforts to tighten restrictions on SNS content are likely to be met with significant resistance, as are efforts to censor personal expression and communication in other domains. Other approaches include working directly with the developers of games, icons and other linked content to promote healthier images, as has been done (with only modest success) in other segments of the entertainment industry. Likewise, employing ratings based on content (as with movies, video games) or invoking parental controls which are often used to restrict internet content more broadly could limit young people's exposure to unhealthy content – with all the caveats and shortcomings of this approach in other areas.

Group profiles and fan pages reflecting shared appreciation of everything from musical groups to banal life experiences like passing notes in class are additional features of some SNS, and are popular with adolescents. These have evolved spontaneously, generated by users, and some boast very large numbers of members, fans or followers. In addition, many are health related. A Facebook search of "smoking," for example, yields hundreds of groups and fan pages with this key word. The several largest among them (with hundreds of thousands of fans each) have anti-smoking titles such as "smoking doesn't make you cool, sorry." (Many groups do promote smoking behavior, but interestingly, these are much smaller, with only hundreds of followers each.) These on-line groups reveal views, behaviors, and importantly, social norms, of a large body of the on-line social network in a way that might not be evident on individual pages and profiles. Health promotion efforts which capitalize on this by adding and promoting positive health messages throughout a social network may offer a promising – and less controversial – avenue for changing the environment around on-line exchanges. Influencing social norms in this way may yield the political will necessary to force change in the on-line community (as has been seen recently with increased privacy protections, in response to user demand).

Social norms and political will established in an on-line context can also be brought to bear in the real world. The 2008 presidential race offers an excellent example of how a grassroots movement can use on-line social networking to effect behavior (i.e. voting) off-line: it provides a fast and efficient means of sharing information, with the added ability to easily embed photos, videos, and external links, making it a particularly powerful tool. This approach proved especially effective at reaching young people, and content was tailored to this demographic. The same mechanisms of could be employed on-line to inspire users to exercise their political will to influence real-world health policy. Users can unite, for example, to lobby for smoke free spaces or bans on tobacco sponsorship of youth-focused events in their own communities.

Appendix 1. Trust prediction and propagation models

In this appendix, we describe more trust prediction and propagation models that are not covered in the main body.

Golbeck and Hendler (2006) proposed a trust metric to determine binary trust i.e., trust and no trust (which is not the same as distrust) in a social network. They integrated the algorithm in an e-mail filtering system called TrustMail. Levin (2002) proposed a trust mechanism, Advogato, which uses the trust flow model and three levels of certification to determine the level of trust between users within a group. Advogato uses the same set of nodes to compute trust for all the nodes in the network. Unreliable nodes are excluded from computing trust by determining the nodes which have rated the "bad" nodes positively in the network.

Appleseed is another group trust metric proposed by Ziegler and Lausen (2004) that uses spreading activation strategies to determine trust. One weakness of this method is that the final value of trust is normalized. Thus for example if there is an actor in the network who has a large number of positive ratings but only a few negative ratings then the overall rating of this actor would be less than the rating of another who has only a couple of positive ratings and no negative ratings. SocialTrust is a reputation-based trust aggregation framework proposed by Caverlee et al. (2009) which supports tamper-resilient trust establishment. This algorithm is dynamic in nature i.e., it can dynamically revise trust via user feedback mechanism as the network evolves. An additional distinguishing feature of this algorithm is that it differentiates relationship quality and trust between people. Dynamic revision in SocialTrust depends upon the following three factors: the current rating, the history, and the adaptation to change. SocialTrust has been tested on the MySpace dataset.

Haifeng, Liu et al (2006) posted the problem of predicting trust as a machine learning problem. They identified a set of features including network features that can be used to predict trust. Lesani and Bagheri (2006) consider the problem of inferring trust in a large social network where one is likely to encounter contradictory information. They develop a fuzzy framework to address this problem. Kuter and Golbeck (2007) developed a trust inference algorithm that employs a probabilistic sampling technique to determine trust as well as confidence with respect to trust. Unlike TidalTrust which uses the shortest path between two nodes to determine trust between two actors it uses only the nodes with high trust values in a trust path to compute trust. Google's PageRank algorithm (Brin and Page 1998) has also been adopted and modified in EigenTrust (Kamwar 2003) to compute a global measure of trust in a social network.

Mitra and Maheswaran (2007) proposed a protocol for information dissemination which uses trust. It disseminates information from trusted users while at the same time preventing flow of spam messages in the network. Information in this model can be spread by three different methods: Receiver Initiated, Sender Initiated, or using both as Hybrid. They use a Bayesian trust estimation scheme to infer trust in this model. Taherian et al (2008) proposed a trust propagation scheme by mapping the problem of trust inference to electrical resistance where trust is represented as the reverse of the resistance. In this model voltage source connects nodes u to v and electrical current flows between u and v. The current is interpreted as trust relation from u to v. Because resistors allows more current to flow between the nodes if there is a trust relationship between the nodes along a certain path then more trust is said to flow between the nodes. The advantage of this model is that it is quite simple and scalable.

Most of the models described previously assume that explicit trust information is available in the social network which can be used to make more inferences about users for whom trust information is not available. However this is not always the case, to address this problem Kim et al (Kim et al 2008) proposed a method where degree of

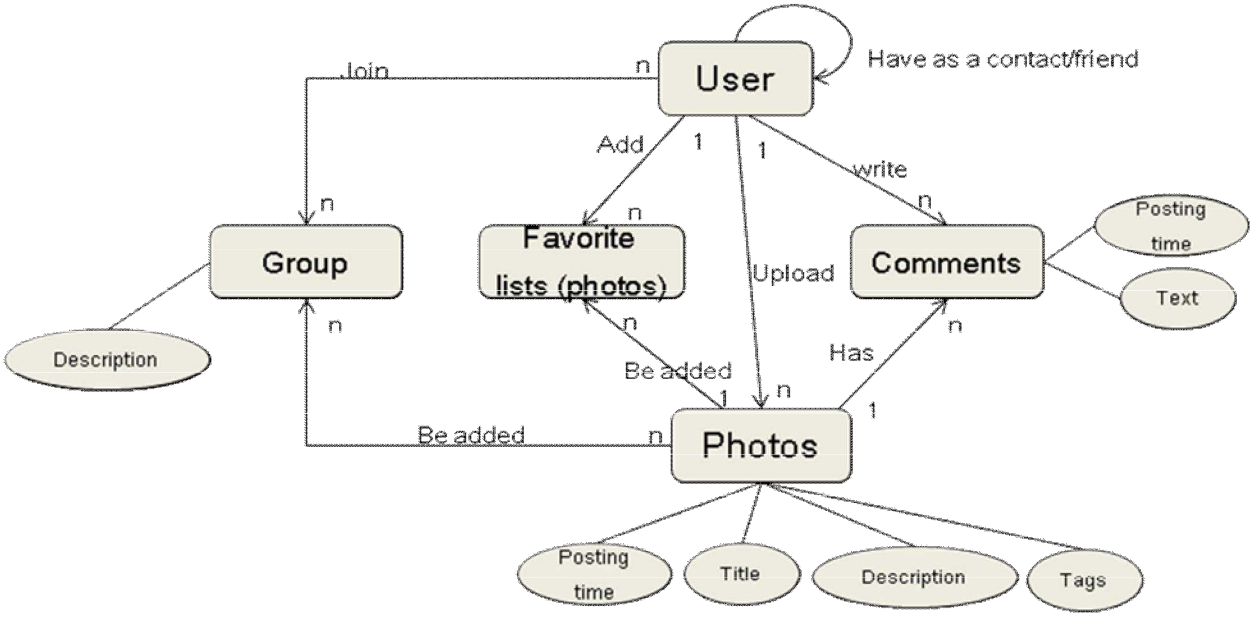
trust between users can be predicted based on users' expertise and users' affinity for certain contexts (topics). The main idea is that the social network formed by interaction of people is much denser than the social network formed by trust and distrust information. By using data from the ratings website epinions they were able to show that interaction information can be used as a proxy to predict trust. Another assumption which is made in almost all the models previously described is that the trust level between two actors does not change over time. This is not always true in the real world. To address this problem Golbeck and Kutter (Golbeck and Kutter 2008) evaluated how different algorithms for propagating trust behave if there were changes in the trust network. Another facet of trust which has been explored is less detail is distrust. We already mentioned that transitivity assumptions in case of trust in social networks have to be modified but in case of distrust in network it is even less clear how distrust show be handled in a network setting. Kunegis et al (Kunegis et al 2009) extended many of the concepts from social network analysis to cases where negative edges are present in the network. However the problem of propagating distrust is still an open problem.

Appendix 2. Additional social networking sites for data collection

A case study of Flickr

Flickr is a photo-sharing social network community which allows users to upload and share photos and videos with others. As shown in the below Figure, a user who creates a Flickr account can upload their own photos with title, tags and short description of the photo and write comments to photos shared by other users. Tags added to a photo allow searchers to find images related to particular topics or keywords. The more positive comments a photo receives, the higher chance a photo is visible in a main page because the number of positive comments means popularity.

When a user likes the shared photo, that user can add the photo his/her favorite list and moreover add the photo sharing user as a contact list in order to follow up his/her photos. A list of contacts for each user is a kind of trust list in terms of sharing high quality photos. Then, if a user A adds a user B in his/her contact list and write many comments on a user B's photos, we can predict 'a user A trusts a user B'. Flickr provides a function of groups that is a pool for photos with a similar topics or subjects. The groups can be created by any Flickr user including individual users and companies in public, public (invite only) or completely private. In a group with specific objective, subjective or theme, users share related photos with other group members who have similar topic interests. Then, we can predict the interesting area that a user is interested in by analyzing the groups a user joins and actively participates in by sharing photos or writing comments.



.

Figure A1. Entity Relationship Diagram showing the major salient features of Flickr

Flickr may not be useful and appropriate to analyze adolescent health behavior through photo sharing based social interactions because it's hard to identify a group of adolescent and the shared photos don't explain enough information of health behavior. Nevertheless, it gives an opportunity to connect Flickr users with Twitter users and to supplement activities, interest areas, social interaction and affiliations in health related communities (or users) in Twitter. For example, when two users join a group of 'workout and exercise' together in Flickr and follow each other reciprocally in Twitter, we can predict they are positively influenced together.

A case study of social game

Mafia Wars and Farmville are Facebook verified online social games provided by Zynga Game Network Inc. The popular social games allow users to invite their friends in Facebook and play the games together. Therefore, it is possible to discover deeper social connections with friends in Facebook and the relationship of game-playing behavior and health behavior. Mafia Wars can be connected to MySpace, Yahoo, iPhone, Tagged as well as Facebook in order to provide a platform for users to form active social relationship with their friends by playing the game.

Currently Mafia Wars has 25 million monthly active users and Farmville has 62 million monthly active users. When a user signs up on Facebook and allows the game access, it will pull user's profile information, photos, friends' information; a user is then able to start playing games. It automatically shows a list of user's friends who already play the game and also provides a function to easily invite user's friend to play together.

In Mafia wars, players have limited energy, health and stamina meters for dong jobs and fighting other players in order to eventually establish and advanced their crime empire. Then, players recruit other mafias from within their friends' network and by using other mafia members, they can fight against bigger mafias and earn experience points. Through such collaboration with other mafias, users can level up and expand their criminal empire. The social interaction in Mafia Wars like other RPG (Role Playing

Game) game will be stronger than that of Facebook because it is constructed for achieving a specific objective such as winning against enemies or expanding a criminal empire. Given the rich interactions that are made in virtual social game, we can develop a trust model based on them and analyze how much a trust network built in social game environment overlap with the other network (i.e., healthy behavior (sports/hobby) network, unhealthy behavior (alcohol, drugs, smoking) network).

Farmville is a real-time farm simulation game where users plant, grow and harvest virtual crops, tress and livestock. In this game, users can invite their friend to be neighbors and by visiting and helping neighbors' farms with neighbors' allowance, users can earn money and experience to expand and own a larger farm. They also can send gifts such as trees and animals to their neighbors. Such interactions with neighbors (i.e., the number of helping a neighbor, the number of sending gifts, the amount of gifts) can be used to measure a level of trust to a neighbor.

References

Adafre, S. F. , Rijke, M., Discovering missing links in Wikipedia, LinkKDD, 2005

Albanesius, C., More Americans go to Facebook than MySpace, PC Magazine, http://www.pcmag.com/article2/0,2817,2348822,00.asp, 2009

Amazon's Mechanical Turk, https://www.mturk.com/mturk/welcome

Baeza-Yates, R.A., Ribeiro-Neto, B., Modern information retrieval, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1999

Bausch, S., Han, L., Social networking sites grow 47%, year over year, reaching 45% of Web users, according to Nielsen//NetRatings. Nielsen Ratings Web site. http://www.nielsen-netratings.com/pr/pr_060511.pdf. Accessed October 27, 2009

Berkowitz, A.D., An overview of the social norms approach. In: Cresskill editor(s), Changing the Culture of College Drinking: A Socially Situated Health Communication Campaign. NJ: Hampton Press, 2005

Blei, D. M., Ng, A.Y., Jordan, M.I., Latent Dirichlet allocation, Journal of Machine Learning Research, 3, 993–1022, 2003

Boyd, D., Viewing American class divisions through Facebook and MySpace, Apophenia Blog Essay. June 24, 2007

Brin, S., Page, L., The anatomy of a large-scale hypertextual Web search engine, Proceedings of the seventh international conference on World Wide Web, Brisbane, Australia. (Section 2.1.1 Description of PageRank Calculation), 107-117, 1998

Bronfenbrenner, U., The ecology of human development: Experiments by nature and design, Cambridge, MA: Harvard University Press, 1979

Caverlee, J., Liu, L., Webb, S., Towards robust trust establishment in web-based social networks with socialtrust, Proceeding of the 17th international conference on World Wide Web. Beijing, China, 2008

Centers for Disease Control and Prevention. Audience insights: Communicating to teens (Aged 12–17), U.S. department of health and human services, National Center for Health Marketing, 2009

Centers for Disease Control and Prevention, Youth risk behavior surveillance — United States, 2007, Surveillance Summaries, MMWR, 57(No. SS-4), June 6, 2008

Chau, M., Xu, J., Mining communities and their relationships in blogs: A study of online hate groups, International Journal of Human-Computer Studies, 65, 57-70, 2005

Christakis, N.A., Fowler, J.H., The collective dynamics of smoking in a large social network, N Engl J Med, 358, 2249-58, 2008

Christakis, N.A., Fowler, J.H., The spread of obesity in a large social network over 32

years, N Engl J Med, 357, 370-9, 2007

Christophe, V., Joshi, Y.V., New product diffusion with influentials and imitators, Marketing Science, 2007

Clark, A.E., Loheac, Y., "It wasn't me, it was them!" Social influence in risky behavior by adolescents, Journal of Health Economics, 26, 763–784, 2007

Cohen-Cole, E., Fletcher, J.M., Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic, Journal of Health Economics, 27, 1382–1387, 2008

Coleman, J. S., Katz, E., Menzel, H., Medical innovation: A diffusion study, Indianapolis: Bobbs-Merrill, 1966

Cotterell, J., Social networks in youth and adolescence (2nd edition) Taylor and Francis, Inc. 2006.

Klovdahl, A.S., Potterat, J.J., Woodhouse, D.E., Muth, J.B., Buth, S.Q., Darrow, W.W., Social networks and infectious disease – the Colorado Springs study, Social Science and Medicine, 38(1), 79-88, 1994

Ebreo, A., Feist-Price, S., Siewe, Y., Zimmerman, R.S., Effects of peer education on the peer educators in a school-based HIV prevention program: where should peer education research go from here?, Health Education & Behavior.,29(4), 411-23, 2002 Aug

Facebook, www.facebook.com

Facebook Statistics, http://www.facebook.com/press/info.php?statistics

Fowler, J.H., Christakis, N.A., Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study, BMJ, 337: 2338, 2008

Fowler, J.H., Christakis, N.A., Estimating peer effects on health in social networks: A response to Cohen-Cole and Fletcher; and Trogdon, Nonnemaker, and Pais, Journal of Health Economics 27, 1400–1405, 2008

Garey, M. R., Johnson, D. S., Computers and intractability: A guide to the theory of NP-completeness freeman, San Francisco, 1979

George, K., Kumar, V., A fast and high quality multilevel scheme for partitioning irregular graphs, SIAM Journal on Scientific Computing, 20-1, 359 - 392, 1999

Goetz, Thomas Practicing Patients New York Times 03/23/2009

Golbeck, J., Computing and applying trust in web-based social networks PhD thesis, University of Maryland, College Park, 2005

Golbeck, J., Hendler, J., Inferring binary trust relationships in web-based social networks, ACM Transactions on Internet Technology, 6(4), New York, NY, USA. November 2006

Golbeck, J., Kuter, U. The ripple effect: Change in trust and its impact over a social network. In Computing with Social Trust. J. Golbeck (ed.), Springer, 2008

Green, H., Google: harnessing the power of cliques, Business Week, October 6, 50, 2008

Hammond, A., Sloboda, Z., Tonkin, P., Stephens, R., Teasdale, B., Grey, S.F., Williams, J., Do adolescents perceive police officers as credible instructors of substance abuse prevention programs?, Health Education Research, 23(4), 682-96, 2008

Harden, A., Oakley, A., Oliver, S., Peer-delivered health promotion for young people: A systematic review of difference study designs, Health Education Journal, 60, 339-353, 2001

Hasan, M. A., Chaoji, V., Salem, S., Zaki, M., Link prediction using supervised learning, Workshop on Link Analysis, Counter-terrorism and Security (at SIAM Data Mining Conference), Bethesda, MD, 2006

Havelin, K., Peer pressure: How can I say no?, Capstone Press, 2000

http://www.blog.compete.com/2007/11/12/connecting-the-social-graph-member-overlap-at-opensocial-and-facebook/

http://www.danah.org/papers/essays/ClassDivisions.html

http://www.epic.org/privacy/socialnet/

Huang, Z., Link prediction based on graph topology : The predictive value of the generalized clustering coefficient, In Proceedings of LinkKDD'06, Philadelphia, Pennsylvania, 2006

Jain, A.K., Fundamentals of digital image processing, Prentice-Hall, 1989

Jantsch, J., Using Twitter for Business, Duct Tape Marketing, 2009

Kamvar, S. D. , Schlosser, M. T., Garcia-Molina,H., The EigenTrust algorithm for reputation ,management in P2P Networks, In Proceedings of the Twelfth International World Wide Web Conference, 2003

Karcher, M.J., Brown, B.B., Elliot, D.W., Enlisting peers in developmental interventions: Principles and practices, In The Youth Development Handbook: Coming of Age in American Communities, Hamilton SF and Hamilton MA, Editors, Sage Publications, Inc.: Thousand Oaks, CA, 2004

Katz, E., Lazarsfeld, P. F., Personal influence; the part played by people in the flow of mass communications, Glencoe, IL: Free Press, 1955

Kim, Y.A., Le, M.T., Lauw, H.W., Lim, E.P., Liu, H., Srivastava, J., Building a web of trust without explicit trust ratings, Workshop on Data Engineering for Blogs, Social Media, and Web 2.0 at ICDE'08, Apr 2008

Kimura, M., Saito, K., Nakano, R., Motoda, H., Finding influential nodes in a social network from information diffusion data, Social Computing and Behavioral Modeling , 1-8, 2009

Kleinberg, J., Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5): 604–632, 1999

Kobayashi, M., Takeda, K., Information retrieval on the web, ACM Computing Surveys (ACM Press) 32 (2), 144–173, 2000

Komro, K.A., Perry, C., Peer-planned social activities for preventing alcohol use among young adolescents, Journal of School Health, 66, 328-334, 1996

Kunegis, J., Lommatzsch, A., Bauckhage, C., The slashdot zoo: Mining a social network with negative edges, WWW, 2009

Kuter, U., Golbeck, J., SUNNY: A new algorithm for trust inference in social networks, using probabilistic confidence models, Proceedings of the Twenty-Second National Conference on Artificial Intelligence (AAAI-07). Vancouver, British Columbia, July, 2007

Lazarsfeld, P. F., Berelson, B., Gaudet, H., The people's choice: How the voter makes up his mind in a presidential campaign, New York: Columbia University Press, 1968

Lenhart, A. Madden, M., Social networking websites and teens: An overview, Pew Internet and American Life Project, http:// www.pewinternet.org/PPF/r/198/report_display.asp, Jan. 2007

Lenhart, A., Madden, M., Hitlin, P., Teens and technology, Pew Internet and American Life Project, http://www.pewinternet.org/Reports/2005/Teens-and-Technology.aspx, July, 2005

Lenhart, S., Madden, M., Macgill, A.R., Smith, A., Teens and social media, Pew Internet and American Life Project, Washington, DC. December 19, 2007

Leicht, E. A., Newman, M. E. J., Community structure in directed networks, Phys. Rev. Lett. 100, 118703, 2008

Lesani, M., Bagheri, S., Fuzzy trust inference in trust graphs and its application in semantic web social networks, World Automation Congress, WAC '06, Sharif University of Technology, Iran, 2006

Levien, R. L, Attack resistant trust metrics, PhD thesis, Department of Computer Science, University of California, Berkeley, 2004

Liben-Nowell, D., Kleinberg, J., The link prediction problem for social networks. In Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM), 2003

Licciardone, J.C., Perceptions of drinking and related findings from the Nationwide Campuses Study, Journal of American College Health, 51(6), 238-45, 2003

Liu, H., Lim, E.P., Lauw, H.W., Le, M.T., Sun, A., Srivastava, J., Kim, Y.A., Predicting trusts among users of online communities: an epinions case study, ACM Conference on Electronic Commerce 2008: 310-319, 2008

Mashable, MySpace Number One, http://mashable.com/2006/07/11/myspace-americas-number-one/

Matsuo, Y., Mori, J., Hamasaki, M., POLYPHONET: An advanced social network extraction system from the web, The International World Wide Web Conference (WWW 2006), (Edinburgh, Scotland, May 23-26, 2006), 397-406, 2006

McCallum, A. , Andres, C.-E., Wang, X., Topic and role discovery in social networks. In IJCAI, 2005

McLeroy, K.R., Bibeau, D., Steckler, A., Glanz, K., An ecological perspective on health promotion programs, Health Education Quarterly,15(4), 351-377, 1998

Mitra, A., Maheswaran, M., Trusted gossip: A rumor resistant dissemination mechanism for peer-to-peer information sharing, Advanced Information Networking and Applications, AINA '07. 21st International Conference on. Dept. of Comput. Sci., Manitoba Univ., Winnipeg, MB, 2007

Moreira, M.T., Smith, L.A., Foxcroft, D., Social norms interventions to reduce alcohol misuse in University or College students, Cochrane Database of Systematic Reviews, Issue 3, 2009

Moreno, M.A., Parks, M., Richardson, L.P., What are adolescents showing the world about their health risk behaviors on MySpace?, Medscape General Medicine, 9(4), 9, 2007

Moreno, M.A., Parks, M.R., Zimmerman, F.J., Brito, T.E., Christakis, D.A., Display of health risk behaviors on MySpace by adolescents: Prevalence and associations, Arch Pediatr Adolesc Med, 163(1), 27-34, 2009

Moturu, S. T., Liu, H., Johnson, W. G., Trust evaluation in health information on the World Wide Web, 30th Annual International IEEE EMBS Conference, Vancouver, British Columbia, Canada, August 20-24, 2008

Nagoshi, C. T., Wood, M. D., Cote, C. C., Abbit, S. M., College drinking game participation within the context of other predictors of other alcohol use and problems, Psychology of Addictive Behaviors 8:203–13, 1994

Newman, M. E. J., Girvan, M., Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004

Newman, M. E. J., Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103: 8577–8582, 2006

Nielsen, Global faces and networked places, A Nielsen Report on Social Networkings: New Global Footprint Nielson Online. March, 2009.

Nishith P., Delong, C., Banerjee, A., Erickson, K., Social topic models for community extraction export, In The 2nd SNA-KDD Workshop '08 (SNA-KDD'08), August 2008

Ochieng, B.M., Health promotion strategy for adolescents' sexual behavior, Journal of Child Health Care, 5(2):77-81, 2001

Palmer, J., Emergency 2.0 is coming to a website near you, New Scientist. http://www.newscientist.com/article/mg19826545.900-emergency-20-is-coming-to-a-website-near-you.html., 2009

Patriquin, A., Connecting the social graph: Member overlap at opensocial and Facebook, Compete.com blog

Paxton, S.J., Schutz, H.K.,Wertheim, E.H., Muir, S.L., Friendship clique and peer influences on body image concerns, dietary restraint, extreme weight-loss behaviors, and binge eating in adolescent girls, Journal of Abnormal Psychology, 108 (2), 255–266, 1999

Perkins, H.W., Craig, D.W., A successful social norms campaign to reduce alcohol misuse among college student-athletes, Journal of Studies on Alcohol, 67(6), 880-9, 2006

Perkins, H.W., Wechsler, H., Variation in perceived college drinking norms and its impact on alcohol abuse: A nationwide study, Journal of Drug Issues, 26(4), 961–974, 1996

Popescul, A., Lyle H. U., Statistical relational learning for link prediction, IJCAI-2003

Prentice, D. A., Dale T. M., Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm, Journal of Personality and Social Psychology 64 (2): 243–56, 1993

Rew, L. Adolescent health: A multidisciplinary approach to theory, research and intervention, Thousand Oaks, CA: Sage Publications, 2005

Rogers, E. M., Diffusion of innovations, New York: Free Press,1995

Rose, G., The strategy of preventive medicine, New York, NY: Oxford University Press, 1992

Sallis, J., Owen, N., Ecological models of health behavior, In: K G, BK R, FM L, eds. Health Behavior and Health Education: Theory, Research, and Practice, 3rd ed. San Francisco, CA: Jossey-Bass, 462-484, 2002

Sawyer, R.G., Pinciaro, P., Bedwell, D., How peer education changed peer sexuality educators' self-esteem, personal development, and sexual behavior, Journal of American College Health, 45(5), 211-7, 1997

Scholly, K., Katz, A.R., Gascoigne, J., Holck, P.S., Using social norms theory to explain perceptions and sexual health behaviors of undergraduate college students: An exploratory study, Journal of American College Health, 53, 159–166, 2005

Seitz, H. H., Seasonal influenza vaccination promotion through interactive media, www.preventinfluenza.org%2FNIVS_2009%2F6%2520-%2520Seitz.pdf, 2009

Scott, J., Social network analysis: A handbook, 2nd ed, Sage Publications, London, 2000

Shiner, M., Defining peer education, Journal of Adolescence, 22, 555-566, 1999

Srivastava, J., Pathak, N., Mane, S., Ahmad, M. A., Data mining for social network analysis, IEEE International Conference on Data Mining (ICDM 2006), Hong Kong, December 18-22, 2006

Stokols, D., Establishing and maintaining healthy environments: Toward a social ecology of health promotion, American Psychologist, 47(1), 6-22, 1992

Sun, P., Unger, J.B., Palmer, P.H., Internet accessibility and usage among urban adolescents in Southern California: implications for web-based health research, Cyberpsychol Behav. 8(5), 441-453, 2005

Taherian, M., Amini, M., Jalili, R., Trust inference in web-based social networks using resistive networks, ICIW, 233-238, 2008

Target Market. Target Market Ads Feature Real Teens Targeting Tobacco Back. Minnesota Department of Health. http://www.health.state.mn.us/news/pressrel/target.htm. Accessed March 2, 2010.

Taskar, B., Abbeel, P. , Wong, M. , Koller, D. Label and link prediction in relational data, In Advances in Neural Information Processing Systems (NIPS) 16, 2003

Tomasz T., Angelova, R., Bedathur, S., Towards time-aware link prediction in evolving social networks, SNA-KDD, 2009

Turney, P., Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), 417-424, 2002

Trogdon, J.G., Nonnemaker, J., Pais, J., Peer effects in adolescent overweight, Journal of Health Economics, 27, 1388–1399, 2008

Twitter, www.twitter.com

Valente, T. W., Network models of the diffusion of innovations, Cresskill, NJ: Hampton, 1995

Vartak, S., A survey on link prediction, http://sourabhvartak.com/pdf/A Survey on Link Prediction.pdf

von Ahn, L., Dabbish, L., Labeling images with a computer game, CHI 2004, ACM Press,319-326, 2004

von Ahn, L., Games with a purpose, IEEE Computer, 6(39), 92–94, 2006

Wang, C., Satuluri, V., Parthasarathy, S., Local probabilistic models for link prediction, IEEE ICDM 2007 (NS, DM), 2007

Wasserman, S., Faust, K., Social network analysis: Methods and applications, Cambridge University Press, 1994

Watts, D.J., Dodds, P.S., Influential networks and public opinion formation, Journal of Consumer Research, 34, 441-458, 2007

Werch, C.E., Pappas, D.M., Carlson, J.M., DiClemente, C.C., Chally, P.S., Sinder, J.A.,

Results of a social norm intervention to prevent binge drinking among first-year residential college students, Journal of American College Health, 49(2), 85-92, 2000 Sep

Wicker, A.W., An introduction to ecological psychology, Pacific Grove, CA. Brooks/Cole, 1979

Williams, A.L., Merten, M.J., A review of online social networking profiles by adolescents: Implications for future research and intervention, Adolescence, 43(170), 2008

Zhang, Q., Yu, T., Irwin, K., A classification scheme for trust functions in reputation-based trust management, in: Proceedings of ISWC Workshop on Trust, Security, and Reputation on the Semantic Web, 2004

Xiang, E.W., Link Prediction Tutorial, http://ihome.ust.hk/~wxiang/Tutorial/LinkPrediction.pdf

Zhou, D., Manavoglu, E., Li, J., Giles, L., Probabilistic models for discovering e-communities, In WWW, 2006

Zhu, X., Goldberg, A. Introduction to semi-supervised learning. Morgan & Claypool Publishers, 2009

Ziegler, C.N., Golbeck, J., Investigating interactions of trust and interest similarity, Decision Support Systems 43(2): 460-475, 2007

Ziegler, C. N., Lausen, G., Spreading activation models for trust propagation, Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04), 2004

Zynga, www.zynga.com