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 08-005

Social Topic Models for Community Extraction

Nishith Pathak, Colin Delong, Kendrick Erickson, and Arindam Banerjee

February 11, 2008

Social Topic Models for Community Extraction

Nishith Pathak, Colin Delong, Kendrick Erickson, Arindam Banerjee
Dept. of Computer Science, University of Minnesota, Twin Cities
{npathak,delong,banerjee}@cs.umn.edu, eric0909@umn.edu

Abstract. With social interaction playing an increasingly important role in the online world, the capability to extract latent communities based on such interactions is becoming vital for a wide variety of applications. However, existing literature on community extraction has largely focused on methods based on the link structure of a given social network. Such link-based methods ignore the content of social interactions, which may be crucial for accurate and meaningful community extraction. In this paper, we present a Bayesian generative model for community extraction which naturally incorporates both the link and content information present in the social network. The model assumes that actors in a community communicate on topics of mutual interest, and the topics of communication, in turn, determine the communities. Further, the model naturally allows actors to belong to multiple communities. The model is instantiated in the context of an email network, and a Gibbs sampling algorithm is presented to do inference. Through extensive experiments and visualization on the Enron email corpus, we demonstrate that the model is able to extract well-connected and topically meaningful communities. Additionally, the model extracts relevant topics that can be mapped back to corresponding real-life events involving Enron.

1. Introduction

The last few years have witnessed a dramatic increase in both the prevalence and importance of online social networks, further underscored by the emergence of companies whose business model is centered around social networking portals, such as MySpace, Facebook, and YouTube. Through such collaborative application frameworks, millions of individuals have established online identities defined not only by the content available on their profiles, but also through their social connections with other individuals. These connections, facilitated by myriad person-to-person networking mechanisms (such as blogging, tagging, and uploading videos), and the increasing availability of data defining them (such as the Enron email data set), have paved the way for computer scientists, sociologists, and others to bring data mining techniques to bear on the analysis of social networks. Inherent to social networks are *communities*, which are groups of individuals connected to each other in some way (see Figure 1). Communities play a vital role in understanding the creation, representation, and transfer of knowledge among people, and are an essential building block of all social networks. However, the relationship of one

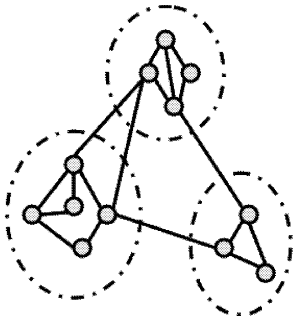

FIGURE 1. Communities in a social network

individual in a community to one another is not easily formalized, or necessarily consistent, and thus the question put to researchers is how, exactly, one extracts communities from a social network. This is one of the more interesting and active research areas of social network analysis.

Traditional methods of community extraction have been primarily link-based. Link-based methods produce communities formed from the explicit links between individuals expressed via some form of measurable interpersonal communication, such as an email or instant message. Actors and their communications are then represented as a graph which is partitioned into different communities ([**10**] and [**14**]). The essential assumption is that intra-community communication is far more dense than inter-community communication. However, community extraction based only on communication links can result in communities which are topically dissimilar, and overly sensitive to individuals who have widely varying philosophies about the frequency of communication and/or the scope of their audience. Thus, it is possible to have two or more latent communities discussing disparate topics merged into a single community since topical information is not utilized. Further, the assumption that every individual belongs to one and only one community does not necessarily hold true in typical social settings. There can also be individuals who are socially inactive and do not belong to any community.

Topic-based methods, on the other hand, can generate communities which are topically similar. In a purely topic-based method, groups of individuals who communicate about the same (or similar) topics become communities in such a framework. A drawback to this approach is that while the communities are topically similar, the individuals contained therein may not share any explicit communication and, as such, may not actually reflect a "community" in the traditional sense. Additionally, issues of synonymy can plague topic-based methods because localized vernacular is not taken into account during extraction, and so while communities are formed which share the same words, the context those words exist in is neglected. This problem is further compounded in social networks utilizing a homogenous language among individuals, such as a company or academic department.

In this paper, we present a probabilistic model for community extraction which allows actors to participate in multiple communities by leveraging both topic and link information from the social network. In particular, we propose a Bayesian model that follows an intuitive generative scheme for modeling email communication. We model the phenomenon of users, belonging to the same community,

exchanging emails among themselves and conversing about topics that are relevant to themselves as well as the community. It can be seen as an extension of the Author-Recipient-Topic Model [**8**] in which users interact with each other based on topics relevant to themselves. We add the *community* element in the ART model, and call it the CART (Community-Author-Recipient-Topic) model. We present a Gibbs sampling based inference scheme for the CART model, and demonstrate its performance on the Enron email corpus.

The rest of the paper is organized as follows: Section 2 presents the CART Model for community extraction as well as the Gibbs sampling updates for the same; Section 3 discusses detailed experimental results on the Enron email corpus; Section 4 discusses issues in evaluating goodness of communities; Section 5 presents related work, followed by the conclusions and future research directions in Section 6.

2. Social Topic Models

In this section we present the CART (Community-Author-Recipient-Topic) model for community extraction. CART is a Bayesian generative model which extends the popular ART (Author-Recipient-Topic) model to discover latent community structure based on authors and recipients. In particular, the observed authors and recipients of an email are assumed to be generated from a latent community. Figures 2(a) and (b) illustrate the ART and CART models respectively.

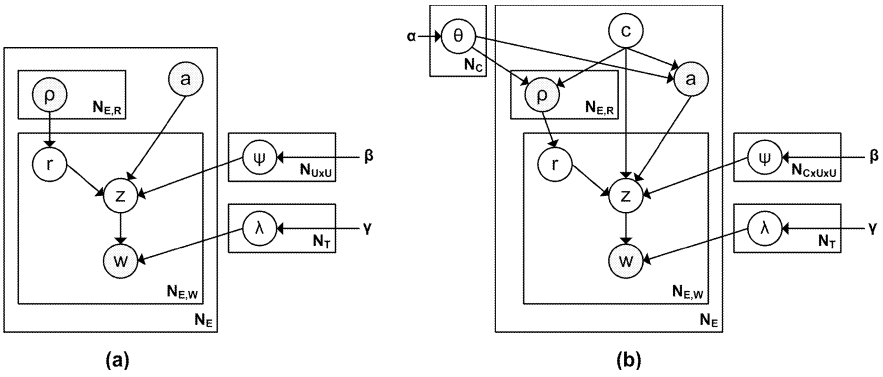

FIGURE 2. Generative models for email networks: (a) The ART model (b) The CART model.

The CART model has the following generation scheme:

(1) To generate email e_d, a community c_d is chosen uniformly at random.
(2) Based the community c_d, the author a_d and the set of recipients ρ_d are chosen.
(3) To generate every word $w_{(d,i)}$ in that email, a recipient $r_{(d,i)}$ is chosen uniformly at random from the set of recipients ρ_d.
(4) Based on the community c_d, author a_d, and recipient $r_{(d,i)}$, a topic $z_{(d,i)}$ is chosen.
(5) The word $w_{(d,i)}$ itself is chosen based on the topic $z_{(d,i)}$.

Other than the uniform distributions for sampling communities c_d, and recipients r_d from ρ_d, all other discrete distributions used in the generative model have Dirichlet

priors as shown in Figure 2(b). The author a_d, set of recipients ρ_d, and sequence of words $\mathbf{w}_d$ used in every email e_d are observable from the email log data, and all other variables are latent. The total number of words (W) and users (U) can be determined from the email log data and the number of communities (C) and topics (T) are provided as inputs. From the model, we can see that every email is constrained to belong to one community. This constrains all users involved and the topics of conversation to belong to the same community in the context of that particular email. A subset of the same users and topics may get assigned to a different community in the context of a different email. The basic intuition behind such a model is that users within a community communicate with each other on topics relevant to themselves as well as the community. Thus, we incorporate link as well as content based information in our community extraction model. The joint probability distribution for the various entities (i.e., communities, authors, recipients, topics and words) for a given email e_d is given as

$$\begin{aligned} &p(c_d, a_d, \rho_d, \mathbf{r}_d, \mathbf{z}_d, \mathbf{w}_d) \\ &= p(c_d)p(a_d|c_d) \prod_{r\in\rho_d} p(r|c_d) \prod_{i=1}^{N_d} p(w_{(d,i)}|z_{(d,i)})p(z_{(d,i)}|c_d, a_d, r_{(d,i)}) \ , \end{aligned} \tag{2.1}$$

where $\mathbf{r}_d$ is the sequence of latent recipients (selected from ρ_d), $\mathbf{z}_d$ is the sequence of latent topic corresponding to word sequence $\mathbf{w}_d$ in the email, $r_{(d,i)}$ is the latent recipient and $z_{(d,i)}$ is the latent topic corresponding to the i^{th} word $w_{(d,i)}$, and N_d is the total number of words in the email.

Given an email corpus over a network of users, the CART model enables the discovery of latent communities in the network, as well as the latent social topics of discussion in the corpus. From a Bayesian network perspective, given the set of observable nodes $(\mathbf{a}, \rho, \mathbf{w})$, such latent structure discovery can be carried out by doing inference over the latent nodes $(\mathbf{c}, \mathbf{r}, \mathbf{z})$. Motivated by recent work on sampling based inference for hierarchical Bayesian models [7], inference in the CART model is carried out using Gibbs Sampling. For CART, the Gibbs sampling updates alternate between updating latent communities c_d conditioned on other variables, and updating recipient-topic tuples $(r_{(d,i)}, z_{(d,i)})$ for each word conditioned on other variables. In particular, the conditional distribution of the community assignment of an email e_d is given by

$$\begin{aligned} &p(c_d = c|\mathbf{c}_{-d}, \boldsymbol{\rho}, \mathbf{r}, \mathbf{a}, \mathbf{z}, \mathbf{w}) \\ &\propto \frac{\prod_{u_i\in\{\rho_d, a_d\}}(n^{CU}_{-d,cu_i} + \alpha)}{\prod_{i=0}^{|\rho_d|}\sum_{u=1}^{U}(n^{CU}_{-d,cu} + U\alpha + i)} \\ &\times \prod_{r\in\rho_d}\left(\frac{\prod_{z=1}^{T}\Gamma(e_{d,rz} + n^{(CUU)T}_{-d,(c_d a_d r)z} + \beta)}{\Gamma\left(\sum_{z=1}^{T}(e_{d,rz} + n^{(CUU)T}_{-d,(c_d a_d r)z})\right) + T\beta}\right) \ , \end{aligned} \tag{2.2}$$

where n^{CU}_{-d,cu_i} is the number of times user u_i was generated from community c other than email d, $e_{d,rz}$ is the number of times topic z was generated from recipient r in email d, and $n^{(CUU)T}_{-d,(c_d a_d, r)z}$ is the number of times topic z was generated from community, author, recipient (c_d, a_d, r) other than email d. Further, the conditional

distribution of the recipient-topic tuple assignment for a word $w_{(d,i)}$ is given by

$$p(r_{(d,i)} = r, z_{(d,i)} = z | \mathbf{c}_{-d}, \mathbf{a}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-(d,i)}, \mathbf{z}_{-(d,i)}, \mathbf{w}_{-(d,i)}, c_d, a_d, \rho_d, w_{(d,i)} = w)$$

$$\propto \frac{n^{TW}_{-(d,i),zw} + \gamma}{\sum_{v=1}^{W} n^{TW}_{-(d,i),zv} + W\gamma} \times \frac{n^{(CUU)T}_{-(d,i),(c_d a_d r)z} + \beta}{\sum_{h=1}^{T} n^{(CUU)T}_{-(d,i),(c_d a_d r)h} + T\beta} \tag{2.3}$$

where, $n^{XY}_{-(d,i),xy}$ is the number of times $y \in Y$ was generated by $x \in X$ excluding the i^{th} instance in email d. Detailed derivations of the conditional distribution based updates are in Appendix A. It is important to note that the presence of a latent node (community c) higher up in the Bayesian network makes the CART model markedly different from much of the recent literature on non-parametric hierarchical Bayesian models. In particular, the Gibbs sampling update for the node c is not as straightforward as latent nodes (such as (r, z)) corresponding to the lower nodes in the Bayesian network.

Using the above updates, a Gibbs sampling simulation is carried till convergence, and the latent node assignments for every email are determined. For a given assignment of latent node values, the communities can be determined as:

$$p(u|c) = \frac{n^{CU}_{cu} + \alpha}{\sum_i n^{CU}_{cu_i} + U\alpha} \tag{2.4}$$

where, n^{CU}_{cu} is the number of times user u was generated from community c. The above equation associates a degree of membership for every user belonging to a community. Note that the model allows for mixed membership, i.e., a user is allowed to participate in more than one community. By counting how many times a user is assigned to a particular community, we can determine the top users for every community. Similarly we can also determine the topmost words for every topic. These topmost words and users can be used to analyze (or to put it more correctly *visualize*) the different topics and communities respectively.

3. Experiments

We demonstrate the performance of our model on the Enron email corpus. The Enron email corpus[1] is a set of emails belonging to 151 users, mostly senior management of Enron, exchanged between mid-1998 and mid-2002 (approximately 4 years), which includes the Enron crisis that broke out in October 2001. In the current experimental setup, a cleaned version is chosen, in which duplicate, erroneous and junk emails have been removed [**13**]. The dataset consists of 252,759 email messages. For experimental analysis only those emails (approximately 20,311) which are exchanged between these 151 users were selected. Results were compiled for 8 communities and 25 topics. All the model hyperparameters were initialized with a value of 1. The model was run for a total of 500 iterations and after stabilizing the Markov chain (around 20 iterations), samples were drawn after every 5 iterations.

Of the eight communities extracted, communities 1-4 were more likely to be observed than communities 5-8 (the exact probabilities for communities 1-8 are 0.14, 0.14, 0.16, 0.24, 0.06, 0.09, 0.07 and 0.1 respectively). We observed that for

[1] http://www.cs.cmu.edu/∼enron/

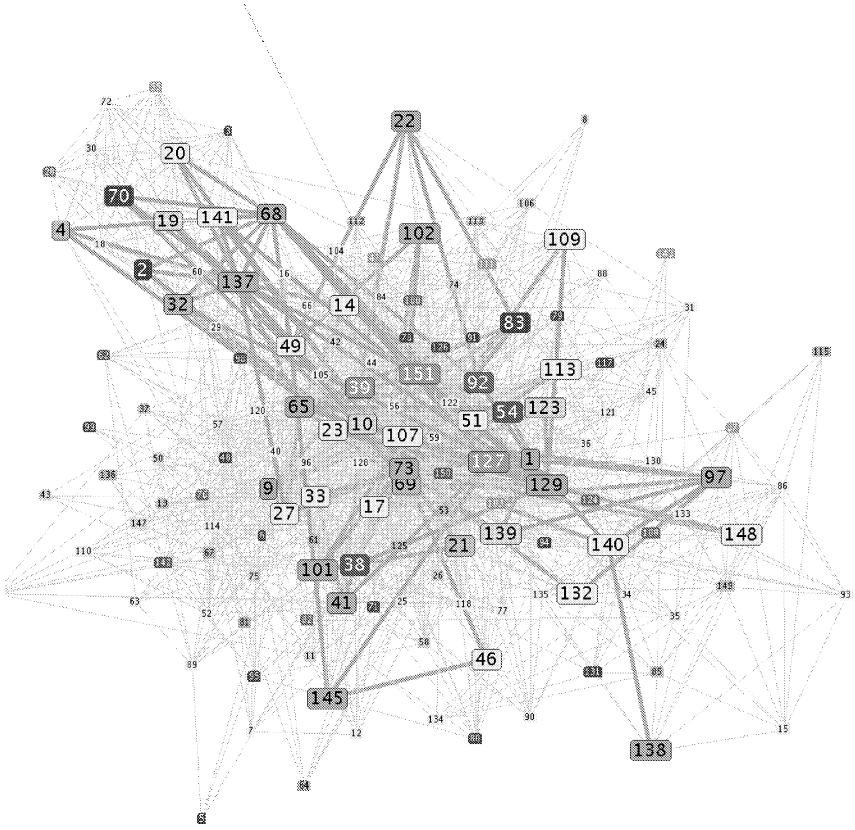

FIGURE 3. Visualization of the red community extracted by the CART model based on Top 15 users.

each community, certain *central actors* (central actors are prominent and communicatively active [**16**]) connect almost all the other actors in that community. This is due to their active communication habits. In our visualizations (see Figure 3), these actors tend to be situated in the central region of the graph.

3.1. Community Visualization. Figures 3 and 4 provide visualizations for communities 2 (red) and 4 (green) respectively. The visualization is based on a spring tension model that uses edge weights (based on the number of emails sent between actors). For each community we consider the top 15 users, each of which are assigned colors indicating community membership (green, red, pink, blue, etc.). Any user among the top 15 for more than one community is colored black and any user not among the top 15 for any community is colored white. In Figure 3 and 4, for both the communities we highlight the subgraph within edge-distance 2 of a chosen central actor. For example, for the green community we highlight the

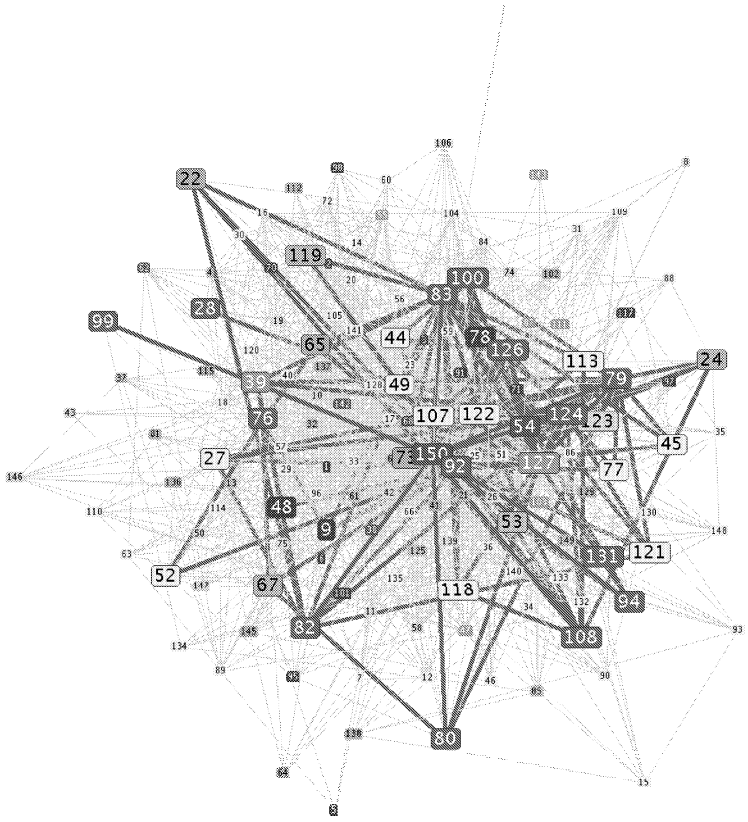

FIGURE 4. Visualization of the green community extracted by the CART model based on Top 15 users.

subgraph within edge-distance 2 of node 92,[2] whereas for the red community we do the same using node 73.

Community Structure: From the figures, we can see that when such a central node is picked, most of the top 15 nodes in the community are highlighted and thus can be reached within a distance of 2 (For both communities, 14 out of the top 15 users are reachable). It is important to observe that all the nodes belonging to the same community that are located in the central portion of the graph are always highlighted. Due to the spring tension model of visualization, any node away from the center has little communication and is relatively (when compared to the central ones) not an active member of the community. Such nodes are the ones missed out when we highlight a subgraph of distance 2 from a central node. This is expected as they are not well connected in general, and need to be reached through a more circuitous path.

[2]We chose not to pick a node too much towards the center as such nodes are highly connected and end up highlighting a large portion of the graph along with the community itself, making it more difficult to visualize the communities

Bridging Nodes: There are a few instances where non-community nodes (particularly white colored nodes) act as bridging points between two nodes of the same community (since the highlighted portion includes all nodes within a distance of 2, the number of bridging nodes is actually quite less than the number of non-community nodes highlighted). The presence of these bridging nodes can be explained by the following: (a) Some of these nodes are articulating points, i.e., important hubs and so are responsible for maintaining connectivity. These hubs either do not participate in any community and simply facilitate communication (e.g., node 122 who is a chief operating officer [**5**]) or are important hubs but still participate in certain communities (e.g., node 150 is the assistant of Enron president Greg Whalley); (b) They are not in the top 25 but if a larger range was considered they would be included.

The choice of the highlighting node does not affect the results as long as they are close to the center. Communicatively inactive actors, when picked, would highlight the community nodes close to them as well as some of the central nodes, but they often miss nodes which are located further away. These results suggest the proposed model does manage to extract communities such that the communicatively active nodes in them are generally well-connected with each other and act as hubs for connecting the inactive or non-central members of the community.

TABLE 1. Topics and their probabilities extracted from the Enron email corpus. Each topic is identified by the top 7 words and their probabilities given the corresponding topic.

Topic 5	0.155	**Topic 9**	0.039	**Topic 10**	0.022	**Topic 11**	0.021
enron	0.021	company	0.006	Sonat	0.003	taliban	0.0008
message	0.011	3d	0.005	dominion	0.002	html	0.0007
original	0.011	germany	0.005	germany	0.002	afghanistan	0.0006
gas	0.008	trading	0.004	mcmichael	0.001	mughniyeh	0.0005
pmto	0.007	nymex	0.003	boyt	0.001	htm	0.0004
fw	0.005	stock	0.002	dth	0.001	terrorist	0.0004
amto	0.005	exchange	0.002	petition	0.001	http	0.0003
Topic 15	0.038	**Topic 16**	0.11	**Topic 17**	0.185	**Topic 21**	0.024
louise	0.007	enron	0.022	enron	0.02	ces	0.005
kitchen	0.005	agreement	0.009	mail	0.01	germany	0.004
john	0.004	sara	0.009	energy	0.008	columbia gas	0.003
mike	0.004	ect	0.008	california	0.007	chris	0.003
meeting	0.003	subject	0.007	power	0.007	cng	0.002
ubs	0.003	corp	0.007	jeff	0.006	transco	0.002
lavorato	0.003	master	0.006	ees	0.006	columbia energy group	0.002

3.2. Social Topics. Table 1 shows the top 7 words for some of the interesting topics discovered by CART in the Enron email dataset. The table also shows the probabilities of occurrence of each topic as well as the probabilities corresponding

to the top 7 words given the topic. The dominant topics in the corpus are *Topic 5*, *Topic 16*, and *Topic 17*.
Topic 5 typically consists of common junk words that are encountered in email communication. Many emails contain the terms 'fw' (forwards) and 'original message'. 'Enron' is expected to be quite common and so is also placed in this topic. In many of the emails, the time field is immediately followed by the To field, and as a result the 'am' or 'pm' suffix of the time value is concatenated with 'to' (hence the 'amto' and 'pmto' terms). Note that detection of such a topic shows that, among other things, CART may be helpful in data cleaning.
Topic 16 is more interesting and is about the master agreement for Enron following its filing for bankruptcy. Sara is one of the employees who is actively involved in communications involving the master agreement and so her name shows up as well. 'ect' is short for Enron Capital Resources, one of Enron's subsidiaries. Although a strong topic, it is not as dominating as topics 5 and 17.
Topic 17 is yet another interesting topic which represents the California power crisis. 'Jeff' is the first name of Jeff Dasovich, Enron's Governmental Affairs Executive and 'ees' stands for Enron Energy Services, which played a major role in the California power crisis. The presence of other terms such as 'California', 'power' and 'Enron' is self-explanatory.

Since *Topic 5* consists of junk terms commonly occurring in emails, it is ubiquitous in the social network. *Topic 17* and, to a certain extent, *Topic 16* are related to the Enron crisis, and hence dominate a large percentage of the Enron email corpus.

Certain other less prominent topics were also extracted from the Enron corpus. *Topic 9* is about Enron's participation in the NYMEX (New York Mercantile Exchange). *Topic 10* is regarding Enron's dealings with Sonet (Southern Natural Gas) and Dominion. 'germany', 'mcmichael' and 'boyt' are last names of employees involved in these dealings and 'dth' (decatherm) is a unit of measure for energy widely used by the energy industry. *Topic 11* was about the war in Afghanistan. Communication regarding this topic consisted of html sources of web documents and so certain terms such as 'http' and 'htm' were also picked up by this topic. *Topic 15* is about UBS's (Union Bank of Switzerland) takeover of Enron Online Services. Louis Kitchen was the president and creator of Enron Online Services. 'Lavorato' is the then Enron CEO's last name and 'Mike' as well as 'John' possibly also represent other people involved. *Topic 16* is similar to *Topic 10* and is in regards to Enron's dealings with Columbia Energy Services ('ces') and the energy transportation firm Transco. Once again the employee Chris Germany emerges.

The dominant topic of communication in the Enron email corpus is the Enron crisis, and this is supported by our results as *Topics 16* and *17* are directly related to the same. The model also picks up on several other less important topics and it is likely that in a large organization like Enron many such smaller topics will exist. Overall, the community and author-recipient based topics extracted by the proposed model are meaningful and can be mapped back to their corresponding real-life events involving Enron.

3.3. Community Profiles. We also present the profile of topics across communities. Figure 5 presents plots for topic probabilities for communities 1,2,3 and 6 (profiles for communities 4,5,7 and 8 are similar to the ones for 2,3 and 6). From the plots we can see that *Topic 17* is very prominent in *Community 1* as opposed to

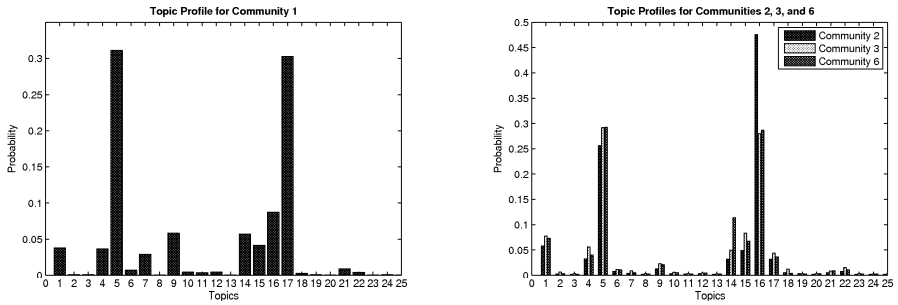

FIGURE 5. Topic profiles $p(z|c)$ for community 1 (on the left) and communities 2,3 and 6 (on the right). Community 1 focuses on *Topic 17*, whereas most other communities focus on *Topic 16*. *Topic 5* is present across all communities.

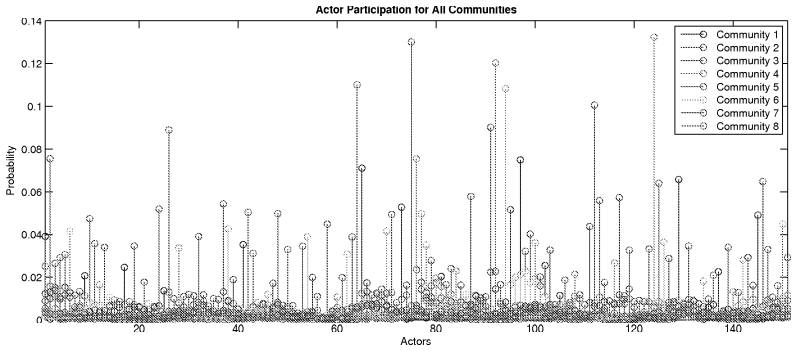

FIGURE 6. Actor profiles $p(u|c)$ for each of the 8 communities. From the above plot we can see that the actor memberships across the community are diverse with different actors having peak memberships in different communities.

Topic 16 which is far more prominent in all other communities. Note that despite being dominant in a single community, *Topic 17* still dominates *Topic 16* in the entire corpus. This is because each word in an email is associated with a topic and so the length of emails will play an important part in deciding the dominance of a topic. It is quite likely that the number of words assigned to *Topic 17* in emails in Community 1 are more than the number of words assigned to *Topic 16* in emails in the other communities. Apart from Community 1, all the other communities have similar topic profiles. This is to be expected due to the heavy dominance of *Topics 16* and *17* in the entire corpus, and even though there are some differences in the prominence of lower strength topics, their effects are mitigated. This is expected as *topic 16* and *topic 17* are related to the Enron crisis, which is the most dominant and widely discussed topic in the corpus.

Figure 6 presents plots of actor probabilities given a community, for all communities. It is readily apparent that communities have different profiles for actor participation. The diversity of profiles implies that several actors, though members

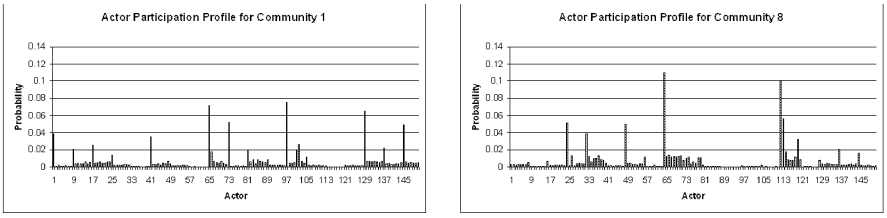

FIGURE 7. Actor profiles $p(u|c)$ for community 1 (left) and community 8 (right).

of different communities, are talking about similar topics within a particular community assignment. Further, the differences in actor participation profiles across communities 2 to 8 can be accounted for by the social links between actors, which also play an important part in determining community structures. In figure 7 we present specific actor profiles for community 1 and 8. In community 1 the topmost actors consisting of numbers 97, 65, 129 and 73. Actor number 97 is one of the traders involved in the infamous Enron tapes that became famous during the California power crisis. Actors 65, 129 and 73 have designations of Managing Director Legal Dpartment, Chief Operating Officer and Government Relation Executive respectively. The topmost topic for community 1 i.e. *topic 17*, the California energy crisis explains the presence of these users at the top of the community. In community 8, the top users tend to be grouped by sets of related broadcast emails pertaining to trading or meetings. While not centered around any one topic, the emails shared by these top users appear to occur most frequently in the case of notification-style emails regarding, for instance, notes from a previous meeting and upcoming meeting times and locations. For instance, emails shared by the top 2 users, actors 112 and 64, were all broadcast emails, with topics ranging from class locations, exotic options information, and a large number of emails recounting conference call content. For actors 64 and 24 (the third most probable user in the community) the emails were again of the broadcast type, this time pertaining to seat assignments for the Enron Center South move. The connectivity factor was important in grouping these users into a community.

From the results presented, we can observe that the proposed CART model is capable of extracting well-linked and topically meaningful communities using both the social link and communication content information. Moreover, the probabilistic nature of our model also allows actors to participate in multiple communities, a more realistic assumption compared to limiting each actor to a single community.

4. Discussion

In this section we briefly discuss the main issues with the evaluation of community extraction methods. Techniques developed in the physics and social science domains (such as [**6**], [**14**] [**10**] and [**4**]) demonstrate their methods on computer-generated random graphs and real world graphs whose community structures are already known. Moreover, these techniques are based on connectivity only and compute hard partitions of the social network graph in order to extract community structures. In such a scenario it is easier to evaluate community extraction methodologies since methods extracting communities which are most closely aligned with

the known community structure are preferred over those that do not. In the absence of communities known a priori, methods extracting communities having high intra-community connectivity and low inter-community connectivity are generally considered superior. In our case, however, community extraction considers both links and topics, essentially imbuing the CART model with a more generalized definition of connectivity. Further, each user can belong to more than one community (i.e. it is not a hard partitioning of the social network graph). As such, traditional graph-based measures are inappropriate for evaluating the quality of such a community extraction method since it attempts to find the best trade-off between high connectivity and topic similarity between users in a community. For example, suppose there is a book reading club consisting of a community in which 20 people follow author YYZ's work. Even if these 20 people do not interact much with each other (i.e. low connectivity) they are still extracted as a community due to their high topic similarity with respect to the work of author YYZ. In such a case purely graph-based measures might not be indicative of good community structure and thus novel methods taking into account links as well as topics for evaluating goodness of community structure are desirable. The Community User Topic (CUT) model [**17**], which extracts communities based on users and topics, also uses an ad-hoc evaluation strategy in which semantic networks, which illustrate users, communities, and topics, are provided and the most probable words generated from a topic are given. The authors also provide a measure of how similar their results are to Newman's modularity based method [**6**]. However, this measure does not indicate whether or not any of the methods are performing better than the other. As such, constructing a measure or a technique for evaluating similar community extraction methods is a non-trivial issue and requires further research effort.

5. Related Work

The existing literature on community extraction from social networks is primarily based on the link information of the network. [**6**] and [**14**] follow an approach based on iteratively removing highest *betweenness* edges from the social network graph, where betweenness is the number of shortest paths traversing through an edge. The graph is broken into connected components, and each component is checked to see if it is a meaningful community. A second approach, discussed in [**10**] and [**4**], is an agglomerative hierarchical algorithm where each node starts out as an individual community and at each step two communities whose amalgamation produces the largest change in *modularity* are merged. Modularity for a given division of nodes into communities C_1 to C_k is defined as $Q = \sum_{i=1}^{k} (e_{ii} - a_i^2)$, where e_{ii} is the fraction of edges that join a vertex in C_i to another vertex in C_i, and a_i is the fraction of total edges that are attached to a vertex in C_i. Recently, [**12**] has presented an extension of this modularity based approach. Other existing approaches are typically based on such graph partition schemes, and do not take communication content information into account. Another limitation of such approaches is that each actor's participation is limited to just one community.

Our proposed approach is based on a Bayesian generative model, which have gained significant popularity in recent years [**3, 7, 1, 9**]. Much of the recent work on such Bayesian models have focused on topic models based on textual content [**3, 7, 2**]. A recent approach that works with relations between entities is the group-topic (GT) model proposed by [**15**]. The goal of the GT model is to

cluster entities such that entities within a group exhibit similar interaction patterns with entities in another group. Broadly, the GT-model generalizes stochastic block models widely studied in the social network analysis literature [**11**], by discovering blocks conditioned on topics of relations between entities. If a GT model were to be applied to email communication data, then the result would be a summarization of the underlying social network where for each topic of conversation one would get groups such that actors in one group exhibit similar communication habits with actors in another group. Thus, although the GT model works with related entities with textual attributes on the relations, it attempts to solve a completely different problem and as such is not applicable to community extraction.

The author-recipient-topic (ART) model, recently proposed by [**8**], extracts topics based on communication between people. The ART-model works with relations as observed through content communication, and models topics based on author and sets of recipients. As explained earlier, our model naturally builds on the ART model by assuming that the author and recipients of an email belong to the same community in the context of the topic of the email.

The community-user-topic (CUT) models were recently proposed by [**17**], where a community is modeled as a joint distribution of topic distributions and user distributions. The model uses Gibbs sampling and entropy computation to filter non-informative samples. The main ideas behind the CUT models are very much similar to our proposed idea in that the CUT models attempt to leverage link as well as communication content information in order to extract communities. However, the underlying semantics of the CUT models are such that there is a loose coupling between how topics and links affect community structure. Specifically, of the two proposed models CUT1 and CUT2, the former is biased towards extracting communities from just the link information and the latter is biased towards extracting communities from just the content information. One of the underlying assumptions of the CUT models is that the topic-user pairs associated with an email are generated from the same community. However, the updates provided, for the CUT model, are for a model which associates a different community with every topic-user pair in an email. In the Bayesian plate model, the semantic assumption of the community generating the topic-user pairs for every word would translate into the community node being outside the plate which contains the topic and word nodes (our model also makes the same assumption, see Figure 2). As noted in Section 2, such semantics lead to a non-trivial update for the conditional distribution corresponding to the community node. The community node Gibbs updates provided in the CUT paper [**17**] are actually for a model which would have the community node inside the plate containing the topic and word nodes, and thus inconsistent with the underlying semantic assumption. The semantics corresponding to the update is equivalent to assuming that an instance of communication (i.e., an email) between two users can belong to many different communities. This is an inappropriate because individual instances of communication tend to focus on a theme or a focused and related set of topics which are indicative of the community within which the users are participating. Corrected community node updates for the CUT models can be derived along similar lines by following the Gibbs sampling update derivations for our proposed model, or the update derivations for the Group-Topic model (see Appendix B).

6. Conclusion

In this paper, a Bayesian generative model for community extraction from social networks was presented. Unlike much of existing literature, the CART model extracts communities based on both communication link as well as content information. The underlying assumption behind the model is that actors in a community communicate on topics of mutual interest, and the topics of communication, in turn, determine the communities. The proposed model is non-parametric, and does not involve any parameter learning or thresholds. Further, the model is probabilistic, and allows actors to be a part of multiple communities. Through extensive experiments and visualization on the Enron email corpus, we demonstrate that the model is able to extract well connected and topically meaningful communities. Additionally, the model extracts relevant topics that can be mapped back to corresponding real life events involving Enron.

In addition to links and textual content information, real social networks, such as Myspace, Facebook, and Youtube, often have additional information on both individual actors as well as their communication patterns. For individual actors, most real networks have an actor profile as well as content (blogs, videos) created by the actor. Further, in addition to emails, actors are allowed to exchange scraps, comments, photos, etc., with other actors. An important direction of future work will be to investigate if such heterogenous observable data can be seamlessly integrated by non-parametric Bayesian models resulting in significantly more powerful latent community extraction methodologies. Further, since communities as well as topics of discussion evolve over time, it will be interesting to see if dynamic versions of the CART model can be used to track the evolution of latent communities and corresponding social topics. Other future work will consist of investigating methods for evaluating the goodness of link and topic based communities.

References

[1] D. Blei, T. Griffiths, M. Jordan, and J. Tanenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2003.

[2] D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 2007. To appear.

[3] David Blei, Andrew Y Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] A Clauset, Mark Newman, and C Moore. Finding community structure in very large networks. *Phys. Rev. E 70*, 0066111, 2004.

[5] J Diesner and K Carley. Exploration of communication networks from the enron email corpus. In *Workshop on Link Analysis Counter-Terrorism and Security, In SIAM International Conference on Data Mining*, 2005.

[6] M Girvan and Mark Newman. Community structure in social and biological networks. In *Proc. National Academy of Science*, pages 7821–7826, 2002.

[7] T Griffiths and M Steyvers. Finding scientific topics. In *Proc. National Academy of Science*, pages 5228–5235, 2004.

[8] Andrew McCallum, Corrada-Emmanuel Andres, and Xuerui Wang. Topic and role discovery in social networks. In *IJCAI*, 2005.

[9] D. Navarro, T. Griffiths, S. Steyvers, and D. Lee. Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, 50:101–122, 2006.

[10] M Newman. Fast algorithms for detecting community structure. *Phys. Rev. E 69*, 066133, 2004.

[11] K. Nowicki and T. Snijders. Estimation and prediction for stochastic block structures. *Journal of American Statistical Association*, 96(455):1077–1087, 2001.

[12] J Reichardt and S Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E 74*, 2006.
[13] J Shetty and J Abidi. Enron email dataset. Technical report, USC Information Sciences Institute, 2004.
[14] J Tyler, D Wilkinson, and B Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *Communities and Technologies*, 2003.
[15] Xueriu Wang, N Mohanty, and Andrew McCallum. Group and topic discovery from relations and their attributes. In *NIPS*, 2006.
[16] S Wasserman and F Katherine. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
[17] D Zhou, E Manavoglu, J Li, and L Giles. Probabilistic models for discovering e-communities. In *WWW*, 2006.

Appendix A. Gibbs Sampling for CART

The joint probability distribution for the various entities (i.e. communities, authors, recipients, topics and words) for a given email e_d is given as

$$\begin{aligned} &p(c_d, a_d, \rho_d, \mathbf{r}_d, \mathbf{z}_d, \mathbf{w}_d) \\ &= p(c_d)p(a_d|c_d) \prod_{r\in\rho_d} p(r|c_d) \prod_{i=1}^{N_d} p(w_{d,i}|z_{d,i})p(z_{d,i}|c_d, a_d, r_{d,i}) \ , \end{aligned} \tag{A.1}$$

where ρ_d is the set of observed unique recipients in the email, $\mathbf{r}_d$ is the sequence of latent recipients (selected from ρ_d) and $\mathbf{z}_d$ is the sequence of latent topic corresponding to each word in the email, and N_d is the number of words in the email.

LEMMA 1. *For a given email e_d,*

$$\begin{aligned} &p(c_d = c|\mathbf{c}_{-d}, \boldsymbol{\rho}, \mathbf{r}, \mathbf{a}, \mathbf{z}, \mathbf{w}) \\ &\propto \frac{\prod_{u_i\in\{\rho_d,a_d\}}(n^{CU}_{-d,cu_i} + \alpha)}{\prod_{i=0}^{|\rho_d|}\sum_{u=1}^{U}(n^{CU}_{-d,cu} + U\alpha + i)} \\ &\times \prod_{r\in\rho_d}\left(\frac{\prod_{z=1}^{T}\Gamma(e_{d,rz} + n^{(CUU)T}_{-d,(c_d a_d r)z} + \beta)}{\Gamma\left(\sum_{z=1}^{T}(e_{d,rz} + n^{(CUU)T}_{-d,(c_d a_d r)z})\right) + T\beta}\right) \ , \end{aligned} \tag{A.2}$$

where n^{CU}_{-d,cu_i} is the number of times user u_i was generated from community c other than email d, $e_{d,rz}$ is the number of times topic z was generated from recipient r in email d, and $n^{(CUU)T}_{-d,(c_d a_d,r)z}$ is the number of times topic z was generated from community, author, recipient (c_d, a_d, r) other than email d.

PROOF. Using Bayes rule,

$$\begin{aligned} &p(c_d = c|\mathbf{c}_{-d}, \boldsymbol{\rho}, \mathbf{r}, \mathbf{a}, \mathbf{z}, \mathbf{w}) = p(c_d = c|\mathbf{c}_{-d}, \boldsymbol{\rho}, \mathbf{r}, \mathbf{a}, \mathbf{z}) \\ &\propto p(a_d, \rho_d, \mathbf{r}_d, \mathbf{z}_d|c_d = c, \mathbf{c}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-d}, \mathbf{a}_{-d}, \mathbf{z}_{-d}) \\ &= p(a_d, \rho_d|c_d = c, \mathbf{c}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-d}, \mathbf{a}_{-d}, \mathbf{z}_{-d}) \times p(\mathbf{r}_d|\rho_d) \\ &\quad \times p(\mathbf{z}_d|c_d = c, \mathbf{c}_{-d}, \rho_d, \boldsymbol{\rho}_{-d}, a_d, \mathbf{a}_{-d}, \mathbf{r}_d, \mathbf{r}_{-d}, \mathbf{z}_{-d}) \\ &\propto p(a_d, \rho_d|c_d = c, \mathbf{c}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-d}, \mathbf{a}_{-d}) \\ &\quad \times p(\mathbf{z}_d|c_d = c, \mathbf{c}_{-d}, a_d, \mathbf{a}_{-d}, \mathbf{r}_d, \mathbf{r}_{-d}, \mathbf{z}_{-d}) \\ &= \mathcal{I}_1 \times \mathcal{I}_2 \ . \end{aligned}$$

Now,

$$
\begin{aligned}
T_1 &= p(a_d, \rho_d | c_d = c, \mathbf{c}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-d}, \mathbf{a}_{-d}) \\
&= \prod_{\substack{i=0 \\ u_i \in \{a_d, \rho_d\}}}^{|\rho_d|} p(u_i | c_d = c, \mathbf{c}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-d}, \mathbf{a}_{-d}, u_0, \ldots, u_{i-1}) \\
&= \prod_{i=0}^{|\rho_d|} \int_{\phi_c} \Big(p(u_i | c_d = c, \phi_c) \\
&\qquad \times p(\phi_c | c_d = c, \mathbf{c}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-d}, \mathbf{a}_{-d}, u_0, \ldots, u_{i-1}) \Big) \, d\phi_c \\
&= \frac{n^{CU}_{-d,cu_0} + \alpha}{\sum_u n^{CU}_{-d,cu} + U\alpha} \times \frac{n^{CU}_{-d,cu_1} + \alpha}{\sum_u n^{CU}_{-d,cu} + U\alpha + 1} \times \cdots \\
&\qquad \times \frac{n^{CU}_{-d,cu_{|\rho_d|}} + \alpha}{\sum_u n^{CU}_{-d,cu} + U\alpha + |\rho_d|} \\
&= \frac{\prod_{u_i \in \{\rho_d, a_d\}} (n^{CU}_{-d,cu_i} + \alpha)}{\prod_{i=0}^{|\rho_d|} \sum_{u=1}^{U} (n^{CU}_{-d,cu} + U\alpha + i)} \, .
\end{aligned}
$$

Further,

$$
\begin{aligned}
T2 &= p(\mathbf{z}_d | c_d = c, \mathbf{c}_{-d}, a_d, \mathbf{a}_{-d}, \mathbf{r}_d, \mathbf{r}_{-d}, \mathbf{z}_{-d}) \\
&= \prod_{r \in r_d} \int_{\psi_{ca_dr}} \Big(p(\mathbf{z}_{d,r} | c_d = c, a_d, r, \psi_{ca_dr}) \\
&\qquad \times p(\psi_{ca_dr} | c_d = c, \mathbf{c}_{-d}, a_d, \mathbf{a}_{-d}, \mathbf{r}_d, \mathbf{r}_{-d}, \mathbf{z}_{-d}) \Big) \, d\psi_{ca_dr} \\
&= \prod_{r \in r_d} \int_{\psi_{ca_dr}} \left(\prod_{z=1}^{T} \psi_{(ca_dr)z}^{e_{d,zr} + n^{(CUU)T}_{-d,z(c_da_dr)} + \beta} \right) \, d\psi_{ca_dr} \\
&= \prod_{r \in r_d} \left(\frac{\prod_{z=1}^{T} \Gamma(e_{d,zr} + n^{(CUU)T}_{-d,(c_da_dr)z} + \beta)}{\Gamma\left(\sum_{z=1}^{T} (e_{d,zr} + n^{(CUU)T}_{-d,(c_da_dr)z}) \right) + T\beta} \right) \, .
\end{aligned}
$$

That completes the proof. □

LEMMA 2. *For a given email e_d,*

$$
\begin{aligned}
p(r_{(d,i)} = r, z_{(d,i)} = z | &\mathbf{c}_{-d}, \mathbf{a}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-(d,i)}, \mathbf{z}_{-(d,i)}, \\
&\mathbf{w}_{-(d,i)}, c_d, a_d, \rho_d, w_{(d,i)} = w)
\end{aligned}
$$

(A.3)

$$
\propto \frac{n^{TW}_{-(d,i),zw} + \gamma}{\sum_{v=1}^{W} n^{TW}_{-(d,i),zv} + W\gamma} \times \frac{n^{(CUU)T}_{-(d,i),(c_da_dr)z} + \beta}{\sum_{h=1}^{T} n^{(CUU)T}_{-(d,i),(c_da_dr)h} + T\beta}
$$

where, $n^{XY}_{-(d,i),xy}$ is the number of times $y \in Y$ was generated by $y \in Y$ excluding the i^{th} instance in email d.

PROOF. Using Bayes rule,

$$\begin{aligned}
&p(r_{(d,i)} = r, z_{(d,i)} = z|\mathbf{c}_{-d}, \mathbf{a}_{-d}, \boldsymbol{\rho}_{-d}, \mathbf{r}_{-(d,i)}, \mathbf{z}_{-(d,i)}, \\
&\qquad\qquad \mathbf{w}_{-(d,i)}, c_d, a_d, \rho_d, w_{(d,i)} = w) \\
&= p(r_{(d,i)} = r|\rho_d) \times p(z_{(d,i)} = z|\mathbf{c}_{-d}, \mathbf{a}_{-d}, \mathbf{r}_{-(d,i)}, \mathbf{z}_{-(d,i)}, \\
&\qquad\qquad \mathbf{w}_{-(d,i)}, c_d, a_d, \rho_d, r_{(d,i)} = r, w_{(d,i)} = w) \\
&\propto p(z_{(d,i)} = z|\mathbf{c}_{-d}, \mathbf{a}_{-d}, \mathbf{r}_{-(d,i)}, \mathbf{z}_{-(d,i)}, c_d, a_d, \rho_d, r_{(d,i)} = r) \\
&\times p(w_{(d,i)} = w|\mathbf{z}_{-(d,i)}, \mathbf{w}_{-(d,i)}, z_{(d,i)} = z) \\
&= T_1 \times T_2 \ .
\end{aligned}$$

Now,

$$\begin{aligned}
T_1 &= p(z_{(d,i)} = z|\mathbf{c}_{-d}, \mathbf{a}_{-d}, \mathbf{r}_{-(d,i)}, \mathbf{z}_{-(d,i)}, c_d, a_d, \rho_d, r_{(d,i)} = r) \\
&= \int_{\psi_{c_d a_d r}} \Big(p(z_{(d,i)} = z|c_d, a_d, r, \psi_{c_d a_d r}) \\
&\qquad \times p(\psi_{c_d a_d r}|\mathbf{c}_{-d}, \mathbf{a}_{-d}, \mathbf{r}_{-(d,i)}, \mathbf{z}_{-(d,i)}, c_d, a_d, r) \ d\psi_{c_d a_d r} \Big) \\
&= \frac{n^{(CUU)T}_{-(d,i),(c_d a_d r)z} + \beta}{\sum_{h=1}^{T} n^{(CUU)T}_{-(d,i),(c_d a_d r)h} + T\beta}
\end{aligned}$$

Further,

$$\begin{aligned}
T2 &= p(w_{(d,i)} = w|\mathbf{z}_{-(d,i)}, \mathbf{w}_{-(d,i)}, z_{(d,i)} = z) \\
&= \int_{\phi_z} p(w_{d,i} = w|\phi_z) p(\phi_z|w_{(d,i)}, z_{-(d,i)}) \ d\phi_z \\
&= \frac{n^{TW}_{-(d,i)zw} + \gamma}{\sum_{v=1}^{W} n_{-(d,i),zv} + W\gamma} \ .
\end{aligned}$$

That completes the proof. □

Appendix B. Corrected Updates for CUT

In the CUT1 (Community-User-Topic) Model the joint probability distribution for the entities communities, users, topics and words for a given email e_d is given as

$$\begin{aligned}
&p(c_d, \mathbf{u}_d, \mathbf{z}_d, \mathbf{w}_d) \\
&= p(c_d) \prod_{i=1}^{N_d} p(w_{d,i}|z_{d,i}) p(z_{d,i}|u_{d,i}) p(u_{d,i}|c_d) \ ,
\end{aligned} \tag{B.1}$$

where μ_d is the set of observed users in the email, $\mathbf{u}_d$ is the sequence of latent recipients (selected from μ_d) and $\mathbf{z}_d$ is the sequence of latent topic corresponding to each word in the email, and N_d is the number of words in the email.

LEMMA 3. *For a given email e_d,*

$$\begin{aligned}
&p(c_d = c|\mathbf{c}_{-d}, \mathbf{u}, \mathbf{z}, \mathbf{w}) \\
&\propto \left(\frac{\prod_{u=1}^{U} \Gamma(e_{d,u} + n^{CU}_{-d,cu} + \alpha}{\Gamma\left(\sum_{u=1}^{U} (e_{d,u} + n^{CU}_{-d,cu})\right) + U\alpha} \right) \ ,
\end{aligned} \tag{B.2}$$

where $n^{CU}_{-d,cu}$ is the number of times user u was generated from community c other than email d and $e_{d,u}$ is the number of times user u was generated in email d.

PROOF. Using Bayes rule,

$$\begin{aligned} p(c_d = c|\mathbf{c}_{-d}, \mathbf{u}, \mathbf{z}, \mathbf{w}) &= p(c_d = c|\mathbf{c}_{-d}, \mathbf{u}) \\ &\propto p(\mathbf{u}_d|c_d = c, \mathbf{c}_{-d}, \mathbf{u}_{-d}) \end{aligned}$$

Now,

$$\begin{aligned} & p(\mathbf{u}_d|c_d = c, \mathbf{c}_{-d}, \mathbf{u}_{-d}) \\ &= \int_{\phi_c} p(\mathbf{u}_d|c_d = c, \phi_c) p(\phi_c|c_d = c, \mathbf{c}_{-d}, \mathbf{u}_{-d}) \; d\phi_c \\ &= \int_{\phi_c} \prod_{u=1}^{U} \phi_{cu}^{e_{d,u}+n^{CU}_{-d,cu}+\alpha} \; d\phi_c \\ &= \left(\frac{\prod_{u=1}^{U} \Gamma(e_{d,u} + n^{CU}_{-d,cu} + \alpha)}{\Gamma\left(\sum_{u=1}^{U}(e_{d,u} + n^{CU}_{-d,cu})\right) + U\alpha} \right) \; . \end{aligned}$$

That completes the proof. □

The community Gibbs updates for the CUT2 model can be obtained in a similar manner where the users and topics are switched in the above. The Gibbs updates for topics z_d and users u_d are the same as discussed in the CUT paper [**17**].