Probabilistic Models for Discovering E-Communities

Ding Zhou[1], Eren Manavoglu[1], Jia Li[2,1], C. Lee Giles[3,1], Hongyuan Zha[1,2]

Department of Computer Science and Engineering[1] The Pennsylvania State University, IST Building University Park, PA 16802

Department of Statistics[2] The Pennsylvania State University, Thomas Building University Park, PA 16802

School of Information Sciences and Technology[3] The Pennsylvania State University, IST Building University Park, PA 16802

ABSTRACT

The increasing amount of communication between individuals in e-formats (e.g. email, Instant messaging and the Web) has motivated computational research in social network analysis (SNA). Previous work in SNA has emphasized the social network (SN) topology measured by communication frequencies while ignoring the semantic information in SNs. In this paper, we propose two generative Bayesian models for semantic community discovery in SNs, combining probabilistic modeling with community detection in SNs. To simulate the generative models, an EnF-Gibbs sampling algorithm is proposed to address the efficiency and performance problems of traditional methods. Experimental studies on Enron email corpus show that our approach successfully detects the communities of individuals and in addition provides semantic topic descriptions of these communities.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; G.3 [**Probability and Statistics**]: Models; J.4 [**Social and Behavioral Sciences**]

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Social Network, Data Mining, Clustering, Email, Gibbs sampling, Statistical Modeling

1. INTRODUCTION

Social network analysis is an established field in sociology [23]. The increasing availability of social network data has led to more computational research in social network analysis (SNA), e.g., discovering interpersonal relationships based on various modes of information carriers, such as emails [22, 27], message boards [10] and the Web [3]. Analysis of social networks (SNs) can be applied to many domains including viral marketing [4, 17] and the evaluation of importance of social actors [24].

WWW, May 23-26, 2006, Edinburgh, Scotland
ACM 1-59593-323-9/06/0005.

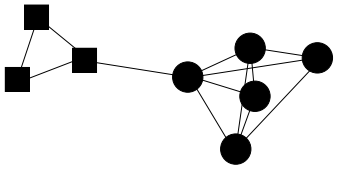


Figure 1: A social network with two communities.

An important characteristic of all SNs is the community graph structure: how social actors gather into groups such that they are intra-group close and inter-group loose [13]. An illustration of a simple two-community SN is sketched in Fig. 1. Here each node represents a social actor in the SN and different node shapes represent different communities. Two nodes share an edge if and only if a relationship exists between them according to social definitions such as their role or participation in the social network. Connections in this case are binary.

Discovering community structures from general networks is of obvious interest. Early methods include graph partitioning [8] and hierarchical clustering [19, 25]. Recent algorithms [6, 14, 2] addressed several problems related to prior knowledge of community size, the precise definition of inter-vertices similarity measure and improved computational efficiency [13]. They have been applied successfully to various areas such as email networks [22] and the Web [5]. Semantic similarity between Web pages can be measured using human-generated topical directories [9]. In general, semantic similarity in SNs is the meaning or reason behind the network connections.

For the extraction of community structures from email corpora [22, 3], the social network is usually constructed measuring the intensity of contacts between email users. In this setting, every email user is a social actor, modeled as a node in the SN. An edge between two nodes indicates that the existing email communication between them is higher than certain frequency threshold.

However, discovering a community simply based purely on communication intensity becomes problematic in some scenarios. (1) Consider a spammer in an email system who sends out a large number of messages. There will be edges between every user and the spammer, in theory presenting

a problem to all community discovery methods which are topology based. (2) Aside from the possible bias in network topology due to unwanted communication, existing methods also suffer from the lack of semantic interpretation. Given a group of email users discovered as a community, a natural question is why these users form a community? Pure graphical methods based on network topology, without the consideration of semantics, fall short in answering to such questions.

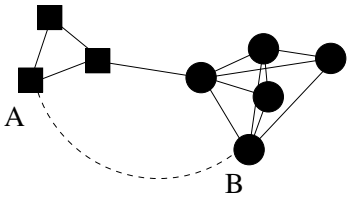


Figure 2: Semantic relationships and hidden communities.

Consider other ways a community can be established, e.g. Fig. 2. From the preset communication intensity, person A and person B belong to two different communities, denoted by squares and circles, based on a simple graph partitioning. However, ignoring the document semantics in their communications, their common interests (denoted by the dashed line) are not considered in traditional community discovery.

In this paper, we examine the inner community property within SNs by analyzing the semantically rich information, such as emails or documents. We approach the problem of community detection using a generative Bayesian network that models the generation of communication in an SN. As suggested in established social science theory [23], we consider the formation of communities as resulting from the similarity among social actors. The generative models we propose introduce such similarity as a hidden layer in the probabilistic model.

As a parallel study in social network with the sociological approaches, our method advances existing algorithms by not exclusively relying the intensity of contacts. Our approach provides topic tags for every community and corresponding users, giving a semantic description to each community. We test our method on the newly disclosed email corpora benchmark – the Enron email dataset and compare with an existing method.

The outline of this paper is as follows: in Section 2, we introduce the previous work which our models are built on. The community detection problem following the line of probabilistic modeling is explained. Section 3 describes our community-user-topic (CUT) models. In Section 4 we introduce the algorithms of Gibbs sampling and EnF-Gibbs sampling (Gibbs sampling with Entropy Filtering). Experimental results are presented in Section 5. We conclude and discuss future work in Section 6.

2. RELATED WORK AND CONTRIBUTIONS

In the first part of this section, we introduce the related work on document modeling. Three related generative models based on which our models are built are described: Topic-Word model, Author-Word model and Author-Topic model. In the second part of this section, we model the generation of SN communications, enabling us to propose a solution to the problem of semantic community identification.

2.1 Related work

Related work on document content characterization [1, 7, 11, 21] introduces a set of probabilistic models to simulate the generation of a document. Several factors in producing a document, either observable (e.g. author [11]) or latent (e.g. topic [7, 1]), are modeled as variables in the generative Bayesian network and have been shown to work well for document content characterization.

Given a set of documents D, each consisting of a sequence of words $\mathbf{w_d}$ of size N_d, the generation of each word $w_{di} \in \mathbf{w_d}$ for a specific document d can be modeled from the perspective of either author or topic, or the combination of both. Fig. 3 illustrates the three possibilities using plate notations. Let ω denote a specific word observed in document d; T and A represent the number of topics and authors; $\mathbf{a_d}$ is the observed set of authors for d. Note that the latent variables are light-colored while the observed ones are shadowed. Fig. 3(a) models documents as generated by a mixture of topics [1]. The prior distributions of topics and words follow Dirichlets parameterized respectively by α and β. Each topic is a probabilistic multinomial distribution over words. Let ϕ denote the topic's distributions over words while θ the document's distribution over topics[1].

In the Topic-Word model, a document is considered as a mixture of topics. Each topic corresponds to a multinomial distribution over the vocabulary. The existence of observed word ω in document d is considered to be drawn from the word distribution ϕ_z, which is specific to topic z. Similarly the topic z was drawn from the document-specific topic distribution θ_d, usually a row in the matrix θ[2].

Similar to the Topic-Word model, an Author-Word model prioritizes the author interest as the origin of a word [11]. In Fig. 3(b), $\mathbf{a_d}$ is the author set that composes document d. Each word in this d is chosen from the author-specific distribution over words. Note that in this Author-Word model, the author responsible for a certain word is chosen at random from $\mathbf{a_d}$.

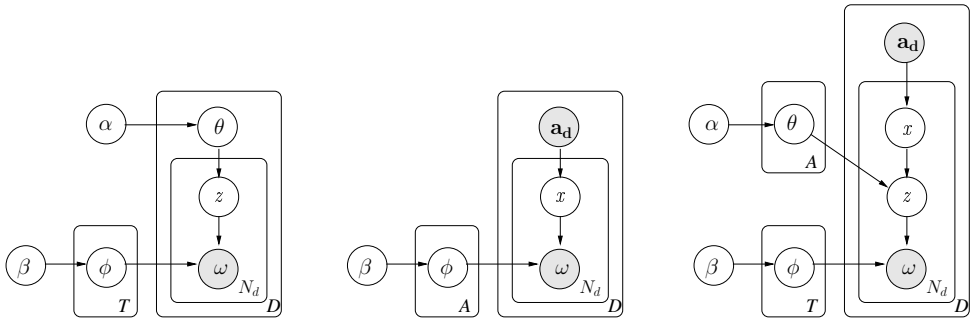
An influential work following this model [21] introduces the Author-Topic model combined with the Topic-Word and Author-Word models and regards the generation of a document as affected by both factors in a hierarchical manner. Fig. 3(c) presents the hierarchical Bayesian structure.

According to the Author-Topic model in Fig. 3(c), for each observed word ω in document d, an author x is drawn uniformly from the corresponding author group $\mathbf{a_d}$. Then with the probability distribution of topics conditioned on x, $\theta_{x\cdot}$, a topic z is generated. Finally the z produces ω as observed in document d.

The Author-Topic model has been shown to perform well for document content characterization because it involves two essential factors in producing a general document: the

[1]Usually the ϕ is represented using $T \times V$ matrix, where T and V are the number of topics and size of vocabulary. Similarly is θ modeled as $D \times T$ matrix.

[2]The Topic-Word model was firstly introduced with name Latent Dirichlet Allocation (LDA). In consistency with the line of research, we use the alternative name.



(a) Topic-Word (LDA) (b) Author-Word (c) Author-Topic

Figure 3: Three Bayesian network models for document content generation

author and the topic. Modeling both factors as variables in the Bayesian network provides the model with capacity to group the words used in a document corpus into semantic topics. Based on the posterior probability obtained after the network is set up, a document can be denoted as a mixture of topic distributions. In addition, each author's preference of using words and involvement in topics can be discovered.

The estimation of the Bayesian network in the aforementioned models typically reply on the observed pairs of author and words in documents. Each word is treated as an instance generated following the probabilistic hierarchy in the models. Some layers in the Bayesian hierarchy are observed, such as authors and words. Other layers are hidden that is to be estimated, such as topics.

2.2 Contributions

We address the problem of identifying social actors based on the semantics of the communications, in particular the semantics as they are related to their communication documents.

Much communication in SNs usually occurs by exchanging documents, such as emails, instant messages or posts on message boards [15]. Such content rich documents naturally serve as an indicator of the innate semantics in the communication among an SN. Consider an information scenario where all communications rely on email. Such email documents usually reflect nearly every aspect of and reasons for this communication. For example, the recipient list records the social actors that are associated with this email and the message body stores the topics they are interested in.

We define such a document carrier of communication as a *communication document*. Our main contribution is resolving the SN communication modeling problem into the modeling of generation of the *communication documents*, based on whose features the social actors associate with each other.

Modeling communication based on *communication document* takes into consideration the semantic information of the document as well as the interactions among social actors. Many features of the SN can be revealed from the parameterized models such as the leader-follower relation [12]. Using such models, we can avoid the effect of meaningless *communication documents*, such as those generated by a network spammer, in producing communities.

Our models accentuate the impact of community on the SN communications by introducing community as a latent variable in the generative models for *communication documents*. One direct application of the models is semantic community detection from SNs. Rather than studying network topology, we address the problem of community exploration and generation in SNs following the line of aforementioned research in probabilistic modeling.

3. COMMUNITY-USER-TOPIC MODELS

Our definition for a *semantic community* in a social network is:

DEFINITION 1. *A semantic community in a social network includes users with similar communication interests and topics that are associated with their communications.*

We study the community structure of an SN by modeling the *communication documents* among its social actors and the format of *communication documents* we model is email because emails embody valuable information regarding shared knowledge and the SN infrastructure [22].

Our Community-User-Topic (CUT) model[3] builds on the Author-Topic model. However, the modeling of a *communication document* includes more factors than the combination of authors and topics.

Serving as an information carrier for communication, a *communication document* is usually generated to share some information within a group of individuals. But unlike publication documents such as technical reports, journal papers, etc., the *communication documents* are inaccessible for people who are not in the recipient list. The issue of a *communication document* indicates the activities of and is also conditioned on the community structure within an SN. Therefore we consider the community as an extra latent variable in the Bayesian network in addition to the author and topic variables. By doing so, we guarantee

[3]In order to fit our model literally to the social network built on email communication, we change the name "Author" to "User". An alternative name of our model is *Community-Author-Topic* Model: *CAT*.

that the issue of a *communication document* is purposeful in terms of the existing communities. As a result, the communities in an SN can be revealed and also semantically explanable.

We will use generative Bayesian networks to simulate the generation of emails in SNs. Differing in weighting the impact of a community on users and topics, two versions of CUT are proposed.

3.1 CUT_1: Modeling community with users

Given the impact of community in the generation of communication, the first step is to determine the interrelationships among this latent variable, the email users and the topics, i.e. the structure of the Bayesian network.

We first consider an SN community as no more than a group of users. This is a notion similar to that assumed in a topology-based method. For a specific topology-based graph partitioning algorithm such as *Modularity* [13], the connection between two users can be simply weighted by the frequency of their communications. In our first model CUT_1, we treat each community as a multinomial distribution over users. Each user u is associated with a conditional probability $P(u|c)$ which measures the degree that u belongs to community c. The goal is therefore to find out the conditional probability of a user given each community. Then users can be tagged with a set of topics, each of which is a distribution over words. A community discovered by CUT_1 is typically in the structure as shown in Fig. 8.

Fig. 4 presents the hierarchy of the Bayesian network for CUT_1. Let us use the same notations in Author-Topic model: α and β parameterizing the prior Dirichlets for topics and words. Let ψ denote the multinomial distribution over users for each community c, each marginal of which is a Dirichlet parameterized by γ. Let the prior probabilities for c be uniform. Let C, U, T denote the number of community, users and topics.

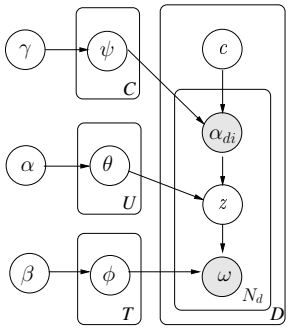


Figure 4: Modeling community with users

Typically, an email message d is generated by four steps: (1) there is a need for a community c to issue an act of communication by sending an email d; (2) a user u is chosen from c as observed in the recipient list in d; (3) u presents to read d since a topic z is concerned, which is drawn from the conditional probability on u over topics; (4) given topic z, a word ω is created in d. By iterating the same procedure, an email message d is composed word by word.

Note that the u is not necessarily the composer of the message in our models. This differs from existing literatures which assume α as the author of document. The assumption is that a user is concerned with any word in a *communication document* as long as the user is on the recipient list.

To compute $P(c, u, z|\omega)$, the posterior probability of assigning each word ω to a certain community c, user u and topic z, consider the joint distribution of all variables in the model:

$$\begin{aligned} P(c, u, z, \omega) =& P(\omega|z)P(c, u, z) \\ =& P(\omega|z)P(z|u)P(c, u) \\ =& P(\omega|z)P(z|u)P(u|c)P(c) \end{aligned} \quad (1)$$

Theoretically, the conditional probability $P(c, u, z|\omega)$ can be computed using the joint distribution $P(c, u, z, \omega)$.

A possible side-effect of CUT^1, which considers a community c solely as a multinomial distribution over users, is it relaxes the community's impact on the generated topics. Intrinsically, a community forms because its users communicate frequently and in addition they share common topics in discussions as well. In CUT_1 where community only generates users and the topics are generated conditioned on users, the relaxation is propagated, leading to a loose connection between community and topic. We will see in the experiments that the communities discovered by CUT_1 is similar to the topology-based algorithm Modularity proposed in [13].

3.2 CUT_2: Modeling community with topics

In contrast to CUT_1, our second model introduces the notion that an SN community consists of a set of topics, which are of concern to respective user groups.

As illustrated in Fig. 5, each word ω observed in email d is finally chosen from the multinomial distribution of a user α_{di}, which is from the recipient list of d. Before that, α_{di} is sampled from another multinomial of topic z and z is drawn from community c's distribution over topics.

Analogously, the products of CUT_2 are a set of conditional probability $P(z|c)$ that determines which of the topics are most likely to be discussed in community c. Given a topic group that c associates for each topic z, the users who refer to z can be discovered by measuring $P(u|z)$.

CUT_2 differs from CUT_1 in strengthing the relation between community and topic. In CUT_2, semantics play a more important role in the discovery of communities. Similar to CUT_1, the side-effect of advancing topic z in the generative process might lead to loose ties between community and users. An obvious phenomena of using CUT_2 is that some users are grouped to the same community when they share common topics even if they correspond rarely, leading to the different scenarios for which the CUT models are most appropriate. For CUT_1, users often tend to be grouped to the same communities while CUT_2 accentuates the topic similarities between users even if their communication seem less frequent.

Derived from Fig. 5, define in CUT^2 the joint distribution of community c, user u, topic t and word ω:

$$P(c, u, z, \omega) = P(\omega|u)P(u|z)P(z|c)P(c) \quad (2)$$

Let us see how these models can be used to discover the communities that consist of users and topics. Con-

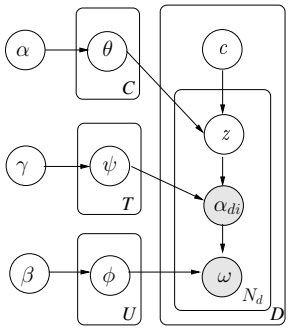


Figure 5: Modeling community with topics

sider the conditional probability $P(c, u, z|\omega)$, a word ω associates three variables: community, user and topic. Our interpretation of the semantic meaning of $P(c, u, z|\omega)$ is the probability that word ω is generated by user u under topic z, in community c.

Unfortunately, this conditional probability cannot be computed directly. To get $P(c, u, z|\omega)$,we have:

$$P(c, u, z|\omega) = \frac{P(c, u, z, \omega)}{\Sigma_{c,u,z} P(c, u, z, \omega)} \quad (3)$$

Consider the denominator in Eq. 3, summing over all c, u and z makes the computation impractical in terms of efficiency. In addition, as shown in [7], the summing doesn't factorize, which makes the manipulation of denominator difficult. In the following section, we will show how an approximate approach of Gibbs sampling will provide solutions to such problems. A faster algorithm EnF-Gibbs sampling will also be introduced.

4. SEMANTIC COMMUNITY DISCOVERY: THE ALGORITHMS

In this section, we first introduce the Gibbs sampling algorithm. Then we address the problem of semantic community discovery by adapting Gibbs sampling framework to our models. Finally, we combine two powerful ideas: Gibbs sampling and entropy filtering to improve efficiency and performance, yielding a new algorithm: EnF-Gibbs sampling.

4.1 Gibbs sampling

Gibbs sampling is an algorithm to approximate the joint distribution of multiple variables by drawing a sequence of samples. As a special case of the Metropolis-Hastings algorithm [18], Gibbs sampling is a Markov chain Monte Carlo algorithm and usually applies when the conditional probability distribution of each variable can be evaluated. Rather than explicitly parameterizing the distributions for variables, Gibbs sampling integrates out the parameters and estimates the corresponding posterior probability.

Gibbs sampling was first introduced to estimate the Topic-Word model in [7]. In Gibbs sampling, a Markov chain is formed, the transition between successive states of which is simulated by repeatedly drawing a topic for each observed word from its conditional probability on all other variables. In the Author-Topic model, the algorithm goes over all documents word by word. For each word ω_i, the topic z_i and the author x_i responsible for this word are assigned based on the posterior probability conditioned on all other variables: $P(z_i, x_i|\omega_i, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}, \mathbf{a_d})$. z_i and x_i denote the topic and author assigned to ω_i, while $\mathbf{z_{-i}}$ and $\mathbf{x_{-i}}$ are all other assignments of topic and author excluding current instance. $\mathbf{w_{-i}}$ represents other observed words in the document set and $\mathbf{a_d}$ is the observed author set for this document.

A key issue in using Gibbs sampling for distribution approximation is the evaluation of conditional posterior probability. In Author-Topic model, given T topics and V words, $P(z_i, x_i|\omega_i, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}, \mathbf{a_d})$ is estimated by:

$$P(z_i = j, x_i = k|\omega_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}, \mathbf{a_d}) \propto \quad (4)$$

$$P(\omega_i = m|x_i = k) P(x_i = k|z_i = j) \propto \quad (5)$$

$$\frac{C^{WT}_{mj} + \beta}{\Sigma_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{AT}_{kj} + \alpha}{\Sigma_{j'} C^{AT}_{kj'} + T\alpha} \quad (6)$$

where $m' \neq m$ and $j' \neq j$, α and β are prior parameters for word and topic Dirichlets, C^{WT}_{mj} represents the number of times that word $\omega_i = m$ is assigned to topic $z_i = j$, C^{AT}_{kj} represents the number of times that author $x_i = k$ is assigned to topic j.

The transformation from Eq. 4 to Eq. 5 drops the variables, $\mathbf{z_{-i}}$, $\mathbf{x_{-i}}$, $\mathbf{w_{-i}}$, $\mathbf{a_d}$, because each instance of ω_i is assumed independent of the other words in a message.

4.2 Semantic community discovery

By applying the Gibbs sampling, we can discover the semantic communities by using the CUT models. Consider the conditional probability $P(c, u, z|\omega)$, where three variables in the model, community, user[4] and topic, are associated by a word ω. The semantic meaning of $P(c, u, z|\omega)$ is the probability that ω belongs to user u under topic z, in community c. By estimation of $P(c, u, z|\omega)$, we can label a community with semantic tags (topics) in addition to the affiliated users. The problem of semantic community discovery is thus reduced to the estimation of $P(c, u, z|\omega)$.

```
(1)  /* Initialization */
(2)  for each email d
(3)      for each word ω_i in d
(4)          assign ω_i to random community, topic and user;
(5)          /* user in the list observed from d */
(6)  /* Markov chain convergence */
(7)  i ← 0;
(8)  I ← desired number of iterations;
(9)  while i < I
(10)     for each email d
(11)         for each ω_i ∈ d
(12)             estimate P(c_i, u_i, z_i|ω_i), u ∈ α_d;
(13)             (p, q, r) ← argmax(P(c_p, u_q, z_r|ω_i));
(14)             /*assign community p,user q, topic r to ω_i*/
(15)             record assignment τ(c_p, u_q, z_r, ω_i);
(16)     i++;
```

Figure 6: Gibbs sampling for CUT models

[4]Note we denote user with u in our models instead of x as in previous work.

The framework of Gibbs sampling is illustrated in Fig. 6. Given the set of users U, set of email documents D, the number of desired topic $|T|$, number of desired community $|C|$ are input, the algorithm starts with randomly assigning words to a community, user and topic. A Markov chain is constructed to converge to the target distribution. In each trial of this Monte Carlo simulation, a block of $(community, user, topic)$ is assigned to the observed word ω_i. After a number of states in the chain, the joint distribution $P(c, u, z|\omega)$ approximates the targeted distribution.

To adapt Gibbs sampling for CUT models, the key step is estimation of $P(c_i, u_i, z_i|w_i)$. For the two CUT models, we describe the estimation methods respectively.

Let $P(c_i = p, u_i = q, z_i = r|\omega_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}})$ be the probability that ω_i is generated by community p, user q on topic r, which is conditioned on all the assignments of words excluding the current observation of ω_i. $\mathbf{z_{-i}}$, $\mathbf{x_{-i}}$ and $\mathbf{w_{-i}}$ represent all the assignments of topic, user and word not including current assignment of word ω_i.

In CUT_1, combining Eq. 1 and Eq. 3, assuming uniform prior probabilities on community c, we can compute $P(c = p, u = q, z = r|\omega_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}})$ for CUT_1 by:

$$P(c_i = p, u_i = q, z_i = r|\omega_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}) \propto$$
$$P(\omega_i = m|z_i = r)P(z_i = r|u_i = q)P(u_i = q|c_i = p) \propto$$
$$\frac{C^{WT}_{mr} + \beta}{\Sigma_{m'} C^{WT}_{m'r} + V\beta} \frac{C^{TU}_{rq} + \alpha}{\Sigma_{r'} C^{TU}_{r'q} + T\alpha} \frac{C^{UC}_{qp} + \gamma}{\Sigma_{q'} C^{UC}_{q'p} + U\gamma} \tag{7}$$

where $P(\omega_i = m|z = r)$, $P(z_i = r|u_i = q)$ and $P(u_i = q|c_i = p)$ are estimated via:

$$P(\omega_i = m|z_i = r) \propto \frac{C^{WT}_{mr} + \beta}{\Sigma_{m'} C^{WT}_{m'r} + V\beta} \tag{8}$$

$$P(z_i = r|u_i = q) \propto \frac{C^{TU}_{rq} + \alpha}{\Sigma_{r'} C^{TU}_{r'q} + T\alpha} \tag{9}$$

$$P(u_i = q|c_i = p) \propto \frac{C^{UC}_{qp} + \gamma}{\Sigma_{q'} C^{UC}_{q'p} + U\gamma}. \tag{10}$$

In the equations above, C^{WT}_{mr} is the number of times that word $\omega_i = m$ is assigned to topic $z_i = r$, not including the current instance. C^{TU}_{rq} is the number of times that topic $z = r$ is associated with user $u = q$ and C^{UC}_{qp} is the number of times that user $u = q$ belongs to community $c = p$, both not including the current instance. C is the number of communities in the social network given as an argument.

The computation for Eq. 8 requires keeping a $W \times T$ matrix C^{WT}, each entry C^{WT}_{ij} of which records the number of times that word i is assigned to topic j. Similarly, a $T \times U$ matrix C^{TU} and a $U \times C$ matrix C^{UC} are needed for computation in Eq. 9 and Eq. 10.

Similarly, $P(c_i = p, u_i = q, z_i = r|\omega_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}})$ is estimated based on the Bayesian structure in CUT_2:

$$P(c = p, u = q, z = r|\omega_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}) \propto$$
$$\frac{C^{WU}_{mq} + \beta}{\Sigma_{m'} C^{WU}_{m'q} + V\beta} \frac{C^{UT}_{qr} + \gamma}{\Sigma_{q'} C^{UT}_{q'r} + U\gamma} \frac{C^{TC}_{rp} + \alpha}{\Sigma_{r'} C^{TC}_{r'p} + T\alpha} \tag{11}$$

Hence the computation of CUT_2 demands the storage of three 2-D matrices: C^{WU}, C^{UT} and C^{TC}.

With the set of matrices obtained after successive states in the Markov chain, the semantic communities can be discovered and tagged with semantic labels. For example, in CUT_1, the users belonging to each community c can be discovered by maximizing $P(u|c)$ in C^{UC}. Then the topics that these users concern are similarly obtained from C^{TU} and explanation for each topic can be retrieved from C^{WT}.

4.3 Gibbs sampling with entropy filtering

In this section, we further develop Gibbs sampling to improve computational efficiency and performance.

Consider two problems with Gibbs sampling illustrated in Fig. 6: (1) efficiency: Gibbs sampling has been known to suffer from high computational complexity. Given a textual corpus with $N = 10^6$ words. Let there be $U = 150$ users, $C = 10$ communities and $T = 20$ topics. An $I = 1000$ iteration Gibbs sampling has the worst time complexity $O(I * N * (U * C * T))$, which in this case is about $3 * 10^{13}$ computations. (2) performance: unless performed explicitly before Gibbs sampling, the algorithm may yield poor performance by including many nondescriptive words. For Gibbs sampling, some common words like 'the', 'you', 'and' must be cleaned before Gibbs sampling. However, the EnF-Gibbs sampling saves such overhead by automatically removing the non-informative words based on entropy measure.

```
(1)  /* Initialization */
(2)  assign each ω_i to random topic, user and community;
(3)
(4)  /* Markov chain convergence */
(5)  i ← 0;   TrashCan ← φ;
(6)  I ← desired number of iterations;
(7)  while i < I
(8)      for each observed ω_i
(9)          if i < A /* in early iterations */
(10)             estimate P(c, u, z|ω_i), u ∈ α_d;
(11)             (p, q, r) ← argmax(P(c_p, u_q, z_r|ω_i));
(12)             record assignment τ(c_p, u_q, z_r, ω_i);
(13)         else /* removing non-informative words */
(14)         if ω_i ∉ TrashCan
(15)             if Entropy(ω_i) ≤ θ
(16)                 (p, q, r) ← argmax(P(c_p, u_q, z_r|ω_i));
(17)                 record assignment τ(c_p, u_q, z_r, ω_i);
(18)             else
(19)                 TrashCan ← TrashCan ∪ {ω_i};
(20)     i++;
```

Figure 7: EnF-Gibbs sampling

Fig. 7 illustrates the EnF-Gibbs sampling algorithm we propose. We incorporate the idea of entropy filtering into Gibbs sampling. During the interactions of EnF-Gibbs sampling, the algorithm keeps in $TrashCan$ an index of words that are not informative. After A times of iterations, we start to ignore the words that are either already in the $TrashCan$ or are non-informative. In Step 15 of Fig. 7, we quantify the informativeness of a word ω_i by the entropy of this word over another variable. For example, in CUT_1

where C^{WT} keeps the numbers of times ω_i is assigned to all topics, we calculate the entropy on the ith row of the matrix.

5. EXPERIMENTS

We present experimental results of our models with the Enron email corpus. Enron email dataset was made public by the Federal Energy Regulatory Commission during its investigations and subsequently made available [20].

In this section, we present examples of discovered semantic communities. Then we compare our communities with those discovered by the topology-based algorithm Modularity [2] by comparing groupings of users. Finally we evaluate the computational complexity of Gibbs sampling and EnF-Gibbs sampling for our models.

We implemented all algorithms in JAVA and all experiments have been executed on Pentium IV 2.6GHz machines with 1024MB DDR of main memory and Linux as operating system.

5.1 Semantic community representation

We preprocessed the Enron email dataset by removing the common stop words. Each employee in Enron is identified by an email address. For brevity, we use only the email ids without organization suffixes hereafter.

In all of our simulations, we fixed the number of communities C at 6 and the number of topics T at 20. The smoothing hyper-parameters α, β and γ were set at $5/T$, 0.01 and 0.1 respectively. We ran 1000 iterations for both our Gibbs sampling and EnF-Gibbs sampling with the MySQL database support. Because the quality of results produced by Gibbs sampling and our EnF-Gibbs sampling are very close, we simply present the results of EnF-Gibbs sampling hereafter.

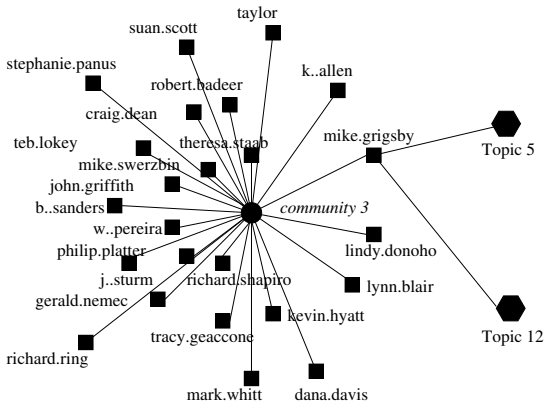


Figure 8: A Community Discovered by CUT_1

The ontologies for both models are illustrated in Fig. 8 and Fig. 11. In both figures, we denote user, topic and community by square, hexagon and dot respectively. In CUT_1 results, a community connects a group of users and each user is associated with a set of topics. A probability threshold can be set to tune the number of users and topics desired for description of a community. In Fig. 8, we present all the users and two topics of one user for a discovered community. By merging all the topics for the desired users of a community, we can tag a community with topic labels.

Topic 3	Topic 5	Topic 12	Topic 14
rate	dynegy	budget	contract
cash	gas	plan	monitor
balance	transmission	chart	litigation
number	energy	deal	agreement
price	transco	project	trade
analysis	calpx	report	cpuc
database	power	group	pressure
deals	california	meeting	utility
letter	reliant	draft	materials
fax	electric	discussion	citizen

Table 1: Topics Discovered by CUT_1

Fig. 8 shows that user mike.grigsby is one of the users in community 3. Two of the topics that is mostly concerned with mike.grigsby are topic 5 and topic 12. Table 1 shows explanations for some of the topics discovered for this community. We obtain the word description for a topic by choosing 10 from the top 20 words that maximize $P(w|z)$. We only choose 10 words out of 20 because there exist some names with large conditional probability on a topic that for privacy concern we do not want to disclose.

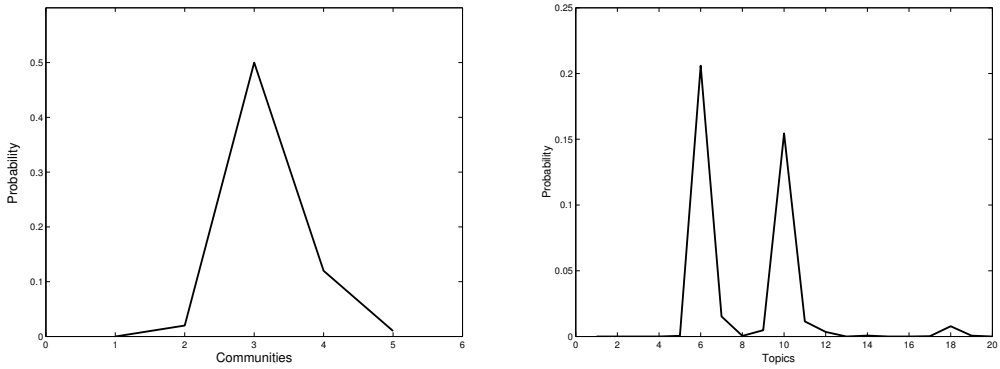
abbreviations	organizations
dynegy	An electricity, natural gas provider
transco	A gas transportation company
calpx	California Power Exchange Corp.
cpuc	California Public Utilities Commission
ferc	Federal Energy Regulatory Commission
epsa	Electric Power Supply Association
naruc	National Association of Regulatory Utility Commissioners

Table 2: Abbreviations

We can see from Table 1 that words with similar semantics are nicely grouped to the same topics. For better understanding of some abbreviate names popularly used in Enron emails, we list the abbreviations with corresponding complete names in Table 2.

For a single user, Fig. 9 illustrates its probability distribution over communities and topics as learned from the CUT_1 model. We can see the multinomial distribution we assumed was nicely discovered in both figures. The distribution over topics for all users are presented in Fig. 10. From Fig. 10, we can see some Enron employees are highly active to be involved in certain topics while some are relatively inactive, varying in heights of peaks over users.

Fig. 11 illustrates a community discovered by CUT_2. According to the figure, Topic 8 belongs to the semantic community and this topic concerns a set of users, which includes rick.buy whose frequently used words are more or less related to business and risk. Surprisingly enough, we found the words our CUT_2 learned to describe such users were very appropriate after we checked the original positions of these employees in Enron. For the four users presented in Table 3, d..steffes was the vice president of Enron in charge of government affairs; cara.semperger was



(a) Over communities (b) Over topics

Figure 9: Communities/topics of an employee

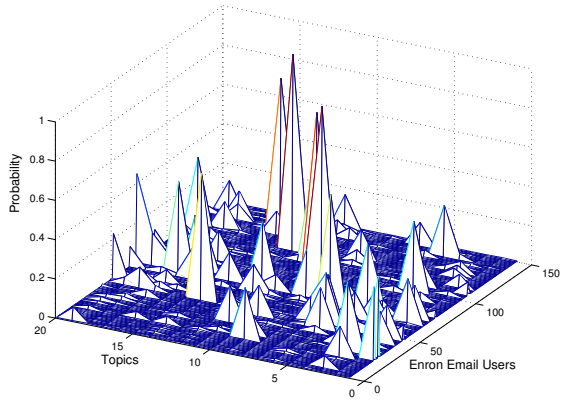


Figure 10: Distribution over topics for all users

a senior analyst; mike.grigsby was a marketing manager and rick.buy was the chief risk management officer.

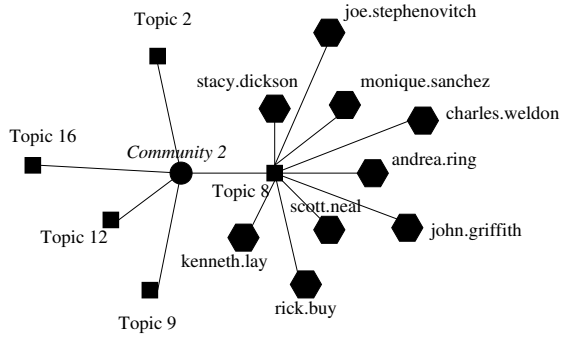


Figure 11: A Community Discovered by CUT_2

d..steffes	cara.s	mike.grigsby	rick.buy
power	number	file	corp
transmission	cash	trader	loss
epsa	ferc	report	risk
ferc	database	price	activity
generator	peak	customer	validation
government	deal	meeting	off
california	bilat	market	business
cpuc	caps	sources	possible
electric	points	position	increase
naruc	analysis	project	natural

Table 3: Distribution over words of some users

5.2 Semantic community discovery quality

We evaluate the quality of discovered communities against the topology-based algorithm in [2], a hierarchical agglomeration algorithm for community structure detection. The algorithm is based on Modularity, which is a measurement of whether a division of a network is a good one, in the sense that there are many edges within communities and only a few between them. We employ the clustering comparison method in [16] to measure the similarity between our communities and the clusters of users produced by [2].

Given N data objects, the similarity between two clustering results λ is defined[5]:

$$\lambda = \frac{N_{00} + N_{11}}{N(N-1)/2}$$

where N_{00} denotes the count of object pairs that are in different clusters for both clustering and N_{11} is the count of pair that are in the same cluster.

The similarities between three CUT models and Modularity are illustrated in Fig. 12. We can see that as we expected the similarity between CUT_1 and $Modularity$ is large while that between CUT_2 and $Modularity$ is small. This is because the CUT_1 is more similar to $Modularity$

[5]Another recent work on comparing clusterings is defined introduced in [26]. But for our problem where cluster labels are categorical, both clustering comparison perform similarly as suggested.

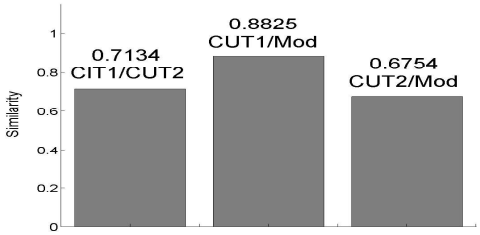


Figure 12: Community similarity comparisons

than CUT_2 by defining a community as no more than a group of users.

We also test the similarity among topics(users) for the users(topics) which are discovered as a community by CUT_1 (CUT_2). Typically the topics(users) associated with the users(topics) in a community represent high similarities. For example, in Fig. 8, Topic 5 and Topic 12 that concern mike.grigsby are both contained in the topic set of lindy.donoho, who is the community companion of Mike Grigsby.

5.3 Computational complexity and EnF-Gibbs sampling

We evaluate the computational complexity of Gibbs sampling and EnF-Gibbs sampling for our models. For the two metrics we measure the computational complexity based on are total running time and iteration-wise running time. For overall running time we sampled different scales of subsets of messages from Enron email corpus. For the iteration-wise evaluation, we ran both Gibbs sampling and EnF-Gibbs sampling on complete dataset.

In Fig. 13(a), the running time of both sampling algorithms on two models are illustrated. We can see that generally learning CUT_2 is more efficient than CUT_1. It is a reasonable result considering the matrices for CUT_1 are larger in scales than CUT_2. Also entropy filtering in Gibbs sampling leads to 4 to 5 times speedup overall.

The step-wise running time comparison between Gibbs sampling and EnF-Gibbs sampling is shown in Fig. 13(b). We perform the entropy filtering removal after 8 iterations in the Markov chain. We can see the EnF-Gibbs sampling well outperforms Gibbs sampling in efficiency. Our experimental results also show that the quality of EnF-Gibbs sampling and Gibbs sampling are almost the same.

6. CONCLUSIONS AND FUTURE WORK

We present two versions of Community-User-Topic models for semantic community discovery in social networks. Our models combine the generative probabilistic modeling with community detection. To simulate the generative models, we introduce EnF-Gibbs sampling which extends Gibbs sampling based on entropy filtering. Experiments have shown that our method effectively tags communities with topic semantics with better efficiency than Gibbs sampling.

Future work would consider the possible expansion of our CUT models as illustrated in Fig. 14. The two CUT models we proposed either emphasize the relation between community and users or between community and topics. It would be interesting to see how the community structure

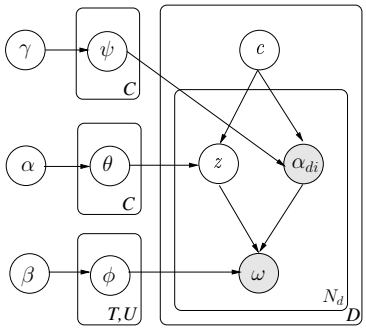


Figure 14: Modeling community with topics and users

changes when both factors are simultaneously considered. One would expect new communities to emerge. The model in Fig. 14 constrains the community as a joint distribution over topic and users. However, such nonlinear generative models require larger computational resources, asking for more efficient yet approximate solutions. It would also be interesting to explore the predictive performance of these models on new communications between strange social actors in SNs.

7. REFERENCES

[1] D. Blei, et al., Latent Dirichlet allocation, *Journal of Machine Learning Research*, 3, 993-1022, 2003.

[2] Aaron Clauset, et al., Finding community structure in very large networks, *Phys. Rev.* E 70, 066111, 2004.

[3] Aron Culotta, et al., Extracting social networks and contact information from email and the Web, In *First Conference on Email and Anti-Spam*, Mountain View, CA, USA. July 2005.

[4] P. Domingos, et al., Mining the network value of customers, In *7th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 57–66, ACM Press, 2001.

[5] G. Flake, S. Lawrence and Lee Giles, Efficient Identification of Web Communities, In *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , 150-160, 2000.

[6] M. Girvan and M. Newman, Community structure in social and biological networks. In Proceedings of *National Academic Science*, USA 99, 7821–7826, 2002.

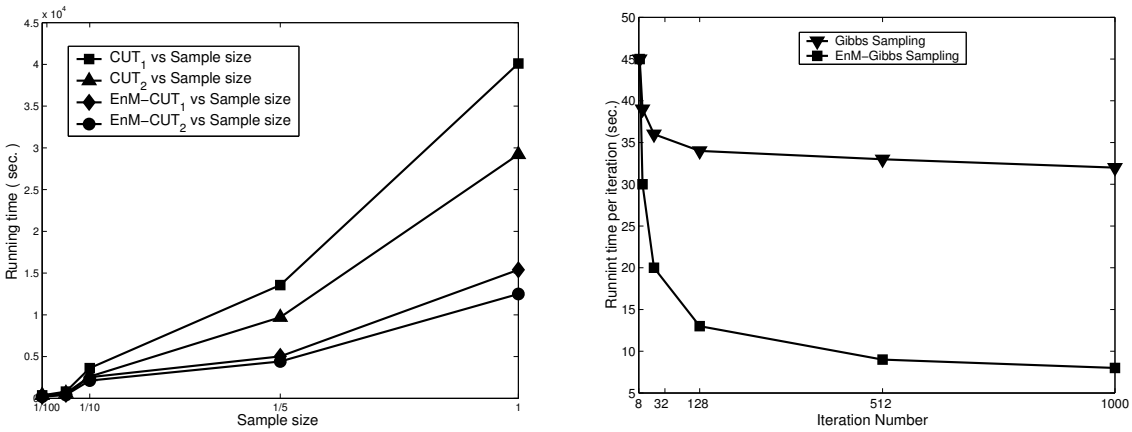
[7] T. Griffiths, Finding scientific topics, In *National Academy of Sciences*, 5228-5235, 2004.

[8] B. W. Kernighan, An efficient heuristic procedure for partitioning graphs, *Bell System Technical Journal*, 49, 291–307, 1970.

[9] Ana Maguitman, et al., Algorithmic detection of semantic similarity, In *Proceedings of the 14th International Conference on World Wide Web (WWW 2005))*. Chiba, Japan, May. 2005.

[10] Naohiro Matsumura, et al., Mining Social Networks in Message Boards, In *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*, Chiba, Japan, May. 2005.

[11] A. McCallum, Multi-label text classification with a mixture model trained by EM, In *AAAI Workshop on*



(a) Time vs. sample size

(b) Time vs. iterations

Figure 13: Computational complexity

Text Learning, 1999.

[12] A. McCallum, et al., The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email, Technical Report, Computer Science, University of Massachusetts Amherst, 2004.

[13] Mark Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev.*, E, 2004.

[14] Mark Newman, Detecting community structure in networks, *Eur. Phys.* 38, 321-330, 2004.

[15] Mike Perkowitz, et al., Mining models of human activities from the Web, In *Proceedings of the 13th International Conference on World Wide Web (WWW 2004)*, New York, NY, USA, 2004.

[16] W. M. Rand, Objective criteria for the evaluation of clustering methods, *Journal of American Statistical Association*, 66:846-850, 1971.

[17] M. Richardson, et al., Mining knowledge-sharing sites for viral marketing, In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 61-70, ACM Press, 2002.

[18] Christian P. Robert and George Casella, Monte Carlo Statistical Methods, Springer Publisher.

[19] J. Scott, Social Network Analysis: A Handbook, Sage, London, 2nd edition, 2000.

[20] Jitesh Shetty, et al., The Enron Email Dataset Database Schema and Brief Statistical Report, Information Sciences Institute, 2004.

[21] Mark Steyvers, et al., Probabilistic author-topic models for information discovery, In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 306-315, Seattle,WA, 2004.

[22] Joshua R. Tyler, et al., Email as Spectroscopy: Automated Discovery of Community Structure within Organizations, *Communities and Technologies*, 81-96, 2003.

[23] Stanley Wasserman and Katherine Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.

[24] S. White, et al., Algorithms for estimating relative importance in networks, In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 266-275, 2003.

[25] Andrew Y. Wu, et al., Mining scale-free networks using geodesic clustering, In *Proceedings of the 10th SIGKDD International Conference on Knowledge Discovery and Data Mining*, 719–724, Seattle, Washington, USA, 2004.

[26] Ding Zhou, et al., A New Mallows distance based Metric for Comparing Clusterings, In *Proceedings of the 22rd International Conference on Machine Learning (ICML 2005)*, 8pp, Bonn, Germany, 2005.

[27] Ding Zhou, et al., Towards Discovering Organizational Structure from Email Corpus, In *Proceedings of the 4th IEEE International Conference on Machine Learning and Applications*, 8 pp, Los Angeles, CA, USA, 2005.