Graphical models based hierarchical probabilistic community discovery in large-scale social networks

Haizheng Zhang

Microsoft Corporation,
1 Microsoft Way, Redmond, WA 98052, USA
E-mail: haizhengzhang@gmail.com

Ke Ke*

College of Business,
Central Washington University,
2400 S. 240th St., Des Moines, WA 98198, USA
Fax: 206-878-0527 E-mail: keg@cwu.edu
*Corresponding author

Wei Li

Contextual and Display Advertising Sciences Department,
Yahoo! Labs,
4401 Great America Parkway, Santa Clara, CA 95054, USA
E-mail: weili@cs.umass.edu

Xuerui Wang

Yahoo! Labs,
701 First Avenue, Sunnyvale, CA 94089, USA
E-mail: xuerui@cs.umass.edu

Abstract: Real-world social networks, while disparate in nature, often comprise of a set of loose clusters (a.k.a. communities), in which members are better connected to each other than to the rest of the network. In addition, such communities are often hierarchical, reflecting the fact that some communities are composed of a few smaller, sub-communities. Discovering the complicated hierarchical community structure can gain us deeper understanding about the networks and the pertaining communities. This paper describes a hierarchical Bayesian model based scheme namely *hierarchical social network-pachinko allocation model* (*HSN-PAM*), for discovering probabilistic, hierarchical communities in social networks. This scheme is powered by a previously developed hierarchical Bayesian model. In this scheme, communities are classified into two categories: *super-communities* and *regular-communities*. Two different network encoding approaches are explored to evaluate this scheme on research collaborative networks, including *CiteSeer*. The experimental results demonstrate that *HSN-PAM* is effective for discovering hierarchical community structures in large-scale social networks.

Keywords: community discovery; hierarchical; social networks; graphical models; data mining.

Reference to this paper should be made as follows: Zhang, H., Ke, K., Li, W. and Wang, X. (2010) 'Graphical models based hierarchical probabilistic community discovery in large-scale social networks', *Int. J. Data Mining, Modelling and Management*, Vol. 2, No. 2, pp.95–116.

Biographical notes: Haizheng Zhang received his PhD in Computer Science from the University of Massachusetts, Amherst. He currently works at Microsoft as a Research Software Developing Engineer. He has broad interests in many research areas such as machine learning and social networks, information retrieval, multi-agent systems and e-commerce systems. Most recently, he has developed interests in auction market mechanism design and sponsored search research.

Ke Ke received her PhD in Management Science from the University of Massachusetts, Amherst. She is currently an Assistant Professor at the Central Washington University. Her research interests include financial networks with intermediation and electronic transactions and supply chain networks. Her other areas of interest include ARCH modelling, risk analysis, corporate finance, econometrics and asset pricing.

Wei Li received her PhD in Computer Science from the University of Massacusetts, Amherst. She currently works at Yahoo!. Her research focuses on machine learning, information retrieval and large scale data mining.

Xuerui Wang received his PhD in Computer Science from the University of Massacusetts, Amherst. He currently works at Yahoo!, working on machine learning for online advertising.

1 Introduction

Social networks have been studied for decades. In recent years, this line of research has gained even more momentum with the prevalence of online social networking systems, such as *MySpace*, *LiveJournal*, *Friendster* and instant messaging systems. These social networking systems are being used by millions and have gained increasing popularity among very diverse user groups. Despite the vast number of nodes, the heterogeneity of the user bases and the variety of interactions among the members, most of these networks exhibit some common properties, such as the small-world property and power-law degree distribution. In addition, some members in the networks form loose clusters, making them better connected to each other than to the rest of the network. An important task in these emerging networks is community discovery, which is to identify subsets of networks such that connections within each subset are dense and connections among different subsets are relatively sparse. Discovering such inherent community structures can lead to deeper understanding about the networks. Since large-scale complex networks based applications exist in many disciplines, community discovery is appealing to researchers from a variety of areas such as computer science, biology, social science and so on.

While the concept of 'community' is self-explanatory, there is no quantitative, rigorous definition that is commonly accepted. This is partly due to the fact that members in social networks can potentially belong to more than one community and the boundaries between communities are often blurry and difficult to draw. Moreover, the

community structures are seldom flat. Analogous to the human society, most complex social networks imply hierarchical structures. For instance, in computer science research collaboration networks, a researcher may belong to artificial intelligence (AI) community in general. But his specific research interests could be focused on a sub-area in AI with his collaborators working on similar topics. The sub-community this researcher and his collaborators belong to, together with the general AI community, constitute a simple two-level community structure. In order to discover these hierarchical communities from large-scale social networks, we develop a *hierarchical social network-pachinko allocation model* (*HSN-PAM*) scheme by applying the *pachinko allocation model* (*PAM*) (Li and McCallum, 2006), a direct acyclic graph (DAG) structured mixture models, to identify and discover probabilistic hierarchical communities in complex, large-scale social networks. This technique is aligned with two previously developed graphical model approaches, namely: *simple social network-latent Dirichlet allocation* (*SSN-LDA*) (Zhang et al., 2007b) and *generic weighted network-latent Dirichlet allocation* (*GWN-LDA*) (Zhang et al., 2007a), which discover hidden correlations among social actors using hierarchical Bayesian network models. However, the *HSN-PAM* model is able to discover not only correlations among social actors in networks, but also correlations among hidden groups, thus making it possible to uncover complicated, hierarchical community structures.

In this paper, we first describe probabilistic model and the pertaining network encoding approaches are evaluated on three social networks with the sizes varying from extremely small to very large. The experimental results indicate that this probabilistic approach is promising in recovering latent relations in large scale social networks. Note that while this approach is evaluated in the social network domain with co-authorship networks, it can be easily extended to other complex network-based applications.

In conclusion, the contributions of this paper include:

1 applying a DAG-structured mixture model to discover hierarchical, probabilistic communities in large-scale networks that only requires the topological structure of networks

2 the exploration of the impact of two different network encoding schemes, namely *direct neighbour encoding scheme* (*DNES*) and *indirect neighbour encoding scheme* (*IDNES*) on hierarchical community discovery.

The rest of this paper is organised as follows: Section 2 introduces related studies; Section 3 introduces related terminology and notations for social networks and discusses using graphical models to detect single-level group structures in social networks; Section 4 describes the network encoding schemes; Section 5 presents the *HSN-PAM* model and its corresponding learning procedures; Section 6 describes three social networks and corresponding experimental results; Section 7 discusses some possible directions for future work including model extension. Finally, Section 8 concludes the paper.

2 Related work

Community structures exist in different types of networks and have been studied in the context of different applications such as: web communities (Flake et al., 2000, 2004; Jing

et al., 2009; Zhang et al., 2007c), social networks (Clauset et al., 2004; Girvan and Newman, 2004; Hopcroft et al., 2004; Newman, 2004b; Palla et al., 2005; Scott, 2000), co-authorship networks (Börner et al., 2004; Krichel and Bakkalbasi, 2006; Newman, 2004a) and biological networks (Girvan and Newman, 2002; Palla et al., 2005; Wilkinson and Huberman, 2004).

This section introduces the background of this study and describes a series of related work, ranging from graph partition, community discovery, clustering algorithms, to several variants of latent Dirichlet allocation (LDA) models.

2.1 Community discovery algorithms

A closely-related problem is graph partitioning problem whose goal is to find a set of optimal graph partitions, so that the edge weight between the partitions is minimised while maintaining partitions of a minimal size. The NP-complete complexity nature of this problem (Garey and Johnson, 1979) requires approximate solutions. Flake et al. (2000, 2004) developed approximate algorithms to partition the network by solving *s-t* maximum flow techniques. The main idea behind maximum flow is to create clusters that have small inter-cluster cuts and relatively large intra-cluster cuts. This idea was first used to explore the web structure in order to provide guidance for crawlers to identify the authoritative nodes (sinks) and hubs, etc. (Flake et al., 2000). While the graph partitioning problem appears to be similar as the community discovery problem, there exists distinct difference between the two problems. Community discovery usually has no requirements on the size of communities and does not attempt to minimise the number of inter-community edges. Most of the existing community discovery methods can be classified as clustering algorithms and fall into two following categories:

1 *Agglomerative approaches:* In an agglomerative approach, similarity (or distance) measures are calculated and edges are added to an initially empty network starting with the vertex pairs with highest similarity. This process stops on a set of pre-defined criteria and the resulting subgraphs are considered as the discovered communities. However, such approaches tend to find only the cores of communities and leave out the periphery.

2 *Divisive approaches:* A divisive method starts with the original network and removes edges based on similarity/distance measures. In practice, the *centrality indices* or *betweenness* metric has been used. The *betweenness* concept was introduced by Freeman (1977) as a centrality measure. It is defined on a vertex v_i as the number of shortest paths between pairs of other vertices that contain vertex v_i. This measure has been used in many previous studies on co-authorship network (Girvan and Newman, 2002; Wilkinson and Huberman, 2004; Krichel and Bakkalbasi, 2006). Girvan and Newman (2002) extended this measure to edges and designed a clustering algorithm which gradually removed the edges with highest betweenness value. A similar approach was taken to find communities in gene networks by Wilkinson and Huberman (2004), where gene networks were created by collecting gene cooccurrence information from the literature and partitioning them into communities of related genes. However, a major problem with this approach is that the complexity of this approach is $O(m^2n)$, where m is the number of edges in the graph and n is the number of vertices in the network.

While the distance measures employed in these approaches are usually easy to understand, such strategies have difficulty in capturing the overlap among communities, the multiple membership phenomenon and inherent hierarchical communities. Conducting hierarchical clustering algorithms based on these distance measures is one potential way to discover hierarchical structures. However, it still does not provide probabilistic results and the stop criteria are usually hard to determine.

2.2 Probabilistic models

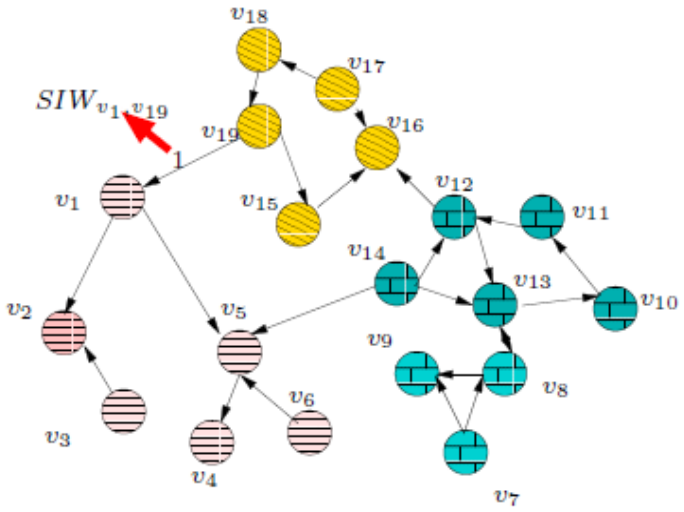
LDA model was first introduced by Blei et al. (2003) for modelling the generative process of a document corpus. Its ability of modelling topics using latent variables has attracted significant interests and it has been applied to many domains such as document modelling (Blei et al., 2003), text classification (Blei et al., 2003), collaborative filtering (Blei et al., 2003), topic models detection (Wang and McCallum, 2006) and social networks (Zhang et al., 2007a, 2007b). Based on LDA model, the PAM model was proposed to capture the correlations among topics by introducing DAG-structured mixture models (Li and McCallum, 2006).

Among the above LDA-based applications, the two approaches proposed in Rosen-Zvi et al. (2004) are both concerned about discovering contextual author groups based on the semantic similarity of their publications. In such models, the evidence is the terms occurring in the documents. In contrast, the two topological community discovery approaches, *SSN-LDA* and *GWN-LDA*, attempt to discover flat communities from social networks (Zhang et al., 2007a, 2007b) by utilising only topological information in social networks. These two models encode the structural information of networks into profiles and discover community structures purely from these social connections among the nodes. With the only input information being the topological structure of a social network, these models can be easily extended to complex networks where no semantic information is available (Bar-Yossef et al., 2006).

3 Single-level community structure discovery

A typical social network G, as shown in Figure 1, is composed of a pair of sets, including the social actor set $V = \{v_1, v_2,..., v_M\}$ and social interaction set $E(e_1, e_2,..., e_N)$, together with a *social interaction weight* function: *SIW:* $(V \times V) \rightarrow I$, where I represents the integer set. The elements of social actor set V are the vertices of G and the elements of social interaction set E are the edges of G, representing the occurrence of social interactions between the corresponding social actors. The number of the social actors in the network is denoted as M. Each social interaction e_i in set E is considered as a binary relation between two social actors, i.e., $e_i(v_{i1}, v_{i2})$ and *SIW* function describes the strength of such interaction. Note that social interaction weight is specified as integer in order to be processed by the *HSN-PAM* model. Throughout this paper, terms *node*, *vertex* and *social actor* are used interchangeably, and so are *edge* and *social interaction*.

Figure 1 A typical social network (see online version for colours)



Notes: Each node represents a social actor and each edge between two nodes represents their social interactions. The weight of an edge, i.e., social interaction weight information, implies the frequency and importance of social interactions between the pertaining social actors.

A node v_i's neighbouring agents are encoded by vector $\overrightarrow{\omega_l}$ and each element $\omega_{ij} \in V$ in the vector represents node v_i's *j*th neighbour. The connectivity of v_i in the network is characterised by its *social interaction profile* (*SIP*), which is defined as a sequence of all v_i's neighbours (ω_{ij}). In this sequence, the frequency of a neighbour ω_{ij} is set as the corresponding social interaction weight information $(SIW(v_i, \omega_{ij}))$. Formally, v_i's SIP is:

$$\overrightarrow{s_l} = (\omega_{i1}, \cdots \omega_{i1}, \omega_{i2}, \cdots \omega_{i2}, \omega_{iN_{v_i}}, \cdots \omega_{iN_{v_i}})$$

where N_i is the number of v_i's neighbouring nodes and the count of a particular neighbouring node $\overrightarrow{s_l}$ in $\rightarrow$ is $SIW(v_i, \omega_{ij})$. Throughout this paper, the variables in sequence $\overrightarrow{s_l}$ is specified as s_{ij}, where $s_{ij} \in \overrightarrow{\omega_l}$. Note that we assume the social interaction elements in this profile are exchangeable and therefore their order will not be concerned. This exchangeability allows these graphical models be used in this application domain (Blei et al., 2003).

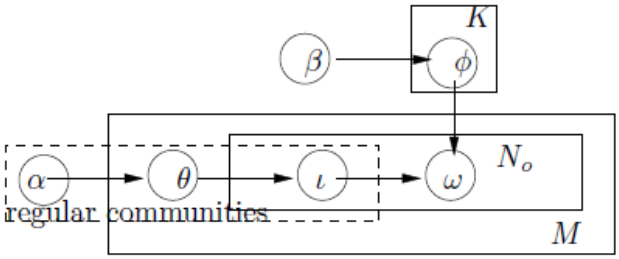
Probabilistic graphical models such as Bayesian networks have been widely used as an important machine learning technique to represent dependency relations between visible and hidden random variables. Among others, document clustering is a well-known application of graphical models where words are modelled as visible variables and clusters are modelled as hidden variables. This paper applies graphical models to community discovery in complex networks. In our model, nodes connectivity information is modelled as visible variables while communities are modelled as hidden variables.

The connections between nodes in social networks are seldom random or casual as there usually exist very manifest group structures in such networks. The essence of community detection using graphical models is to learn and discover relations among

hidden groups from social networks based on visible social interaction information. In these models, these groups are often modelled as latent variables and the dependency relations among groups and social interactions are captured by introducing a variety of graphical structures among these variables. Subsequently, the social network in question can be generated from a generative process based on such graphical structures.

In particular, the recently proposed *SSN-LDA* model (Zhang et al., 2007b), as depicted in Figure 2, applies a hierarchical Bayesian network model, LDA model, to discover communities from large scale social networks. This model includes a hidden community layer which contains a set of community variables $\iota(\iota_1, \iota_\exists, \iota_k)$, as well as a social actor layer, $\vec{s_t}$, which represents the occurrence of social actors in SIPs. Each social actor contributes a part, big or small to every community in the society. The community proportion variable θ is regulated by a Dirichlet distribution with a known parameter α. Meanwhile, each social actor belongs to every community with different probabilities and therefore its SIPs can be represented as random mixtures over latent communities' variables.

Figure 2 Graphical model for *SSN-LDA*



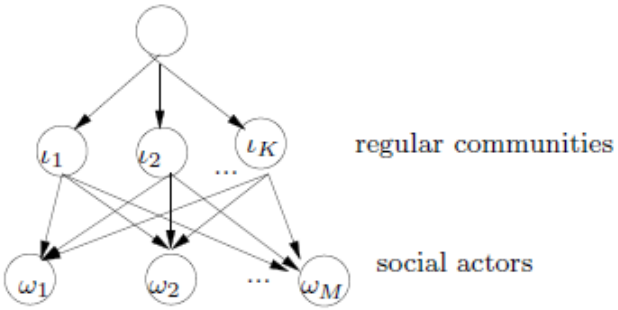
The notation for all the variables in Figure 2 is listed in Table 1. The distribution of communities in SIPs and the social actors over communities are two multinomial distributions with two Dirichlet priors, whose hyperparameters are $\vec{\alpha}$ and $\vec{\beta}$ respectively. The dimensionality K of the Dirichlet distribution, which is also the number of community component distributions, is assumed to be known and fixed. M is the number of social actors (SIPs) in the social network; and N_i is the number of social interactions in a SIP $\vec{s_t}$. $\vec{\alpha}$ is the hyperparameter(known) of the Dirichlet prior distribution of the mixing proportion; $\vec{\beta}$ is the Dirichlet prior hyperparameter (known) on the mixture component distributions for *SSN-LDA*.

Table 1 Notation for symbols in *SSN-LDA*

ι	Hidden community variables
ω_{ij}	Social interaction variables in $\vec{s_t}$
ι_{ij}	Community assignment of s_{ij}
$\vec{\theta}$	$p(\iota \vert \vec{s_j})$, community mixture proportion for $\vec{s_t}$
$\vec{\varnothing}_k$	$p(s_{ki} \mid \iota_k)$, the mixture component of community k

Figure 3 illustrates *SSN-LDA* approach in a more intuitive way with a tree-like structure, which consists of one root node at the top, a set of regular communities in the middle and a set of social actors at the bottom.

Figure 3 Tree structure of *SSN-LDA*, including *K* communities and *M* social actors



However, while the communities discovered by *SSN-LDA* capture co-occurrences among social actors, they do not explicitly model correlations among communities. This limitation arises because the community proportions for each SIP are regulated by a single Dirichlet distribution. Correspondingly, *SSN-LDA* has difficulties in modelling data in which some communities are closer to other communities. This paper presents an alternative graphical model, namely *HSN-PAM*, to represent and learn nested community correlations and identify hierarchical communities in large-scale social networks based on a DAG-based graphical model.

4 Network encoding schemes

The set of SIPs collectively determines the topological structure of a social network. The *HSN-PAM* model depends on the profile information to learn the graphical model and identify hidden communities in the pertaining social networks. In this paper, we explore two different encoding schemes, namely *DNES* and *IDNES*, to generate SIPs.

In the *DNES* scheme, a social actor v_i's SIP contains all directly connected neighbours and the count of each neighbour in the profile is one. Hence, the SIPs of all the social actors constitute the adjacent matrix of the social network. Many previous studies on social networks use this simple representation (Freeman, 1977; Wilkinson and Huberman, 2004). More formally, the SIW function is defined as:

$$SIW_D(v_{i1}, v_{i2}) = \begin{cases} 1 & e(v_{i1}, v_{i2}) \in E; \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

However, one of the disadvantages of the *DNES* scheme is that the SIPs give no consideration to those social actors that are close, but not directly connected to the node in question. The *IDNES* scheme addresses this problem by taking node's indirect neighbours into account. This way, the SIPs reflect the proximity of the nodes in the network more accurately. Furthermore, the final matrix defined by the SIPs is less sparse which can improve the performance of the graphical models (Si and Jin, 2005). While theoretically, such encoding scheme can factor into any indirect neighbours that are

arbitrary hops away; in this paper, the *IDNES* approach only gives weights to a node's direct neighbours and neighbours' neighbours. The SIW function for *IDNES* is defined as follows:

$$SIW_D(v_{i1}, v_{i2}) = \begin{cases} 1 & \text{if } e(v_{i1}, v_{in}) \in E \\ & \text{And } e(v_{in}, v_{i2}) \in E \\ & \text{And } e(v_{i1}, v_{i2}) \notin E; \\ 2 & \text{if } e(v_{i1}, v_{i2}) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

5 HSN-PAM model

In the hierarchical community structure that will be described in this section, namely *HSN-PAM*, the concept of communities is extended to include two different types of communities, namely *regular communities* and *super communities*. The two types of communities are denoted as ι^s (*super communities*) and ι^r (*regular communities*). A *regular community* is defined as a distribution on the social actor space while a *super community* is considered as a distribution on the *regular communities* or *super communities*. There can be arbitrary number of *super community* levels in *HSN-PAM*. In this section, we first introduce the generic *HSN-PAM* model in Section 5.1 and describe a simplified *HSN-PAM* model, namely *TLC-HSN-PAM*, with a two-level community structure. Finally, the Gibbs sampler for solving *TLC-HSN-PAM* model is presented in Section 5.3.

5.1 Generic PAM model description

The *HSN-PAM* model uses a direct acrylic graph (DAG) structure to represent and learn arbitrary-arity, nested and possibly sparse correlations among communities in social networks in contrast to the single-level hierarchy structure in *SSN-LDA*. In Figure 5, each community ι_i is associated with a distribution g_{ι_i} over its children. In general, g_{ι_i} could be any distribution over discrete variables such as logistic normal. In this paper, we assume the distribution associated with communities is Dirichlet component multinomials (DCM) Dir_t (Minka and Lafferty, 2002). A DCM distribution is defined as a distribution hierarchy, including a multinomial distribution and a Dirichlet prior. Dirichlet is often used as the prior distributions for multinomial distributions in Bayesian statistics in order to obtain close-form solutions. In the context of *HSN-PAM*, this means that the SIP is generated by a multinomial distribution whose parameters are generated by its Dirichlet prior distribution.

5.2 Two-level community HSN-PAM model

This paper focuses on a simplified, two-level community *HSN-PAM* structure, which is shown in Figure 5. The two level community structure consists of two types of communities: super community $\vec{\iota^s} = \{\iota_1^s, \iota_2^s, \ldots, \iota_{k_1}^s\}$ and regular community

$\vec{\iota^r} = \{\iota_1^r, \iota_2^r, \ldots, \iota_{k_1}^r\}$. Figure 5 demonstrates that the root community γ is connected to all super communities and all super communities are fully connected to regular communities. Finally, regular communities are fully connected to all the social actors in the social network.

Table 2 Notation for quantities in *HSN-PAM*

ι	Hidden community variable
ι_{Sc}	Super community variable
ι_c	Regular community variable
$\vec{\theta}$	$p(\iota \vert \vec{s_j})$, community mixture proportion for $\vec{s_j}$
$\vec{\varnothing}_k$	$p(s_{ki} \vert \iota_k)$, the mixture component of community k

Figure 4 Graphical model for *TLC-HSN-PAM*

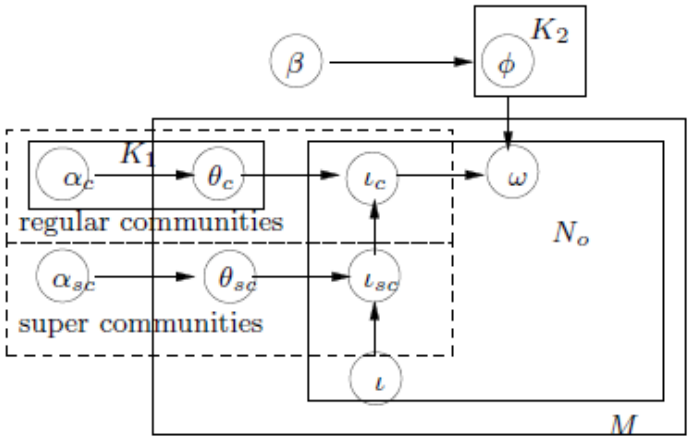
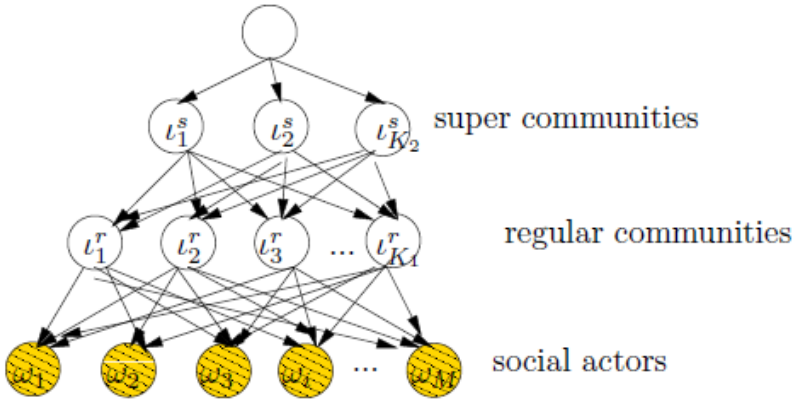


Figure 5 Tree structure of a two-level community structure *HSN-PAM* model, including K_2 super communities, K_1 regular communities and M social actors (see online version for colours)



Two different types of distributions are used in this two-level community structure. We specify that the distributions of root and super communities are still DCM distribution while the distributions of regular communities are modelled with fixed multinomial distributions $\varnothing_{\iota_i^r}$, sampled once for the whole social network from a single Dirichlet distribution $Dir(\beta)$. The corresponding graphical model is shown in Figure 4 and the notations are listed in Table 2.

The multinomials for the root and super communities are still sampled individually for each SIP. Each community ι_i is associated with a Dirichlet distribution.

Based on the graphical model in Figure 4, the generative process for a social actor's SIP $\overrightarrow{s_j}$ is a two-step process:

1 Sample $\overrightarrow{\theta_r^j}$ from the root $Dir_t(\alpha_t)$, where $\overrightarrow{\theta_t^j}$ is a multinomial distribution over super-communities.

2 For each super-community ι_i^s, sample $\overrightarrow{\theta_{\iota_i^s}}$ from $Dir_l(\alpha_i)$, where $\overrightarrow{\theta_{\iota_i^s}}$ is a multinomial distribution over regular communities.

3 For each social actor in the SIP,

   1 sample a super-community ι_w^s from $\overrightarrow{\theta_r}$

   2 sample a regular community ι_j^r from $\overrightarrow{\theta_{\iota_w}}$

   3 sample a social actor ω from $\overrightarrow{\varnothing_{\iota_i^r}}$.

The model structure and the generative process for this special setting are similar to *SSN-LDA* approach. The major difference is that it has one additional layer of super-topics modelled with Dirichlet multinomials, which are the key component capturing correlations among communities here. Another way to interpret this is that given the regular communities, each super-community is essentially an individual *SSN-LDA* structure. Therefore, this can be viewed as a mixture over a set of *SSN-LDA* models.

Following this process, the joint probability of generating a SIP, the community assignment i and the multinomial distribution $\vec{\theta}$ is:

$$P(\overrightarrow{s_t},\vec{\iota},\vec{\theta}|\alpha,\phi) = P(\overrightarrow{\theta_t}|\alpha_t)\Pi_{t=1}^{s}P(\vec{\theta}_{\iota_i}|\alpha_i)\times\Pi_{\omega}(P(\iota_\omega|\overrightarrow{\theta_r})P(\iota_\omega|\overrightarrow{\theta_r})P(\iota_j^r|\vec{\theta}_{\iota_w}P(\omega|\varnothing_{\iota_j^r}))) \quad (3)$$

Integrating out and summing over, we calculate the marginal probability of a SIP as:

$$P(\overrightarrow{s_t},|\alpha,\Phi) = \int P(\overrightarrow{\theta_t}|\alpha_t)\Pi_{t=1}^{s}P(\vec{\theta}_{\iota_i}|\alpha_i)\times\Pi_{\omega}\Sigma_{\iota_\omega}(P(\iota_\omega|\overrightarrow{\theta_r})P(\iota_j^r|\overrightarrow{\theta_{l_w}})P(\omega|\Phi_{\iota_j^r}))\,d\vec{\theta} \quad (4)$$

The probability of generating the entire social network $\vec{S}$ is the product of the probability for every SIP $\vec{S}_t$, integrating out the multinomial distributions for regular communities Φ:

$$P(\vec{S}|\alpha,\beta) = \int \Pi_j\, P(\varnothing_{\iota_j^r}|\beta)\,\Pi_{\overrightarrow{s_t}}P(\overrightarrow{s_t}|\alpha,\Phi)d\varnothing$$

5.3 Gibbs samplers for HSN-PAM

Exact inference is generally intractable for even the two-level community *HSN-PAM* model. There have been three major approximate approaches for solving this type of hierarchical Bayesian network models, including variational expectation maximisation (Blei et al., 2003), expectation propagation (Minka and Lafferty, 2002) and Gibbs sampling (Andrieu et al., 2003; Griffiths and Steyvers, 2004; Heinrich, 2004). Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) simulation (MacKay, 2002) where the dimension K of the distribution are sampled alternately one at a time, conditioned on the values of all other dimensions (Heinrich, 2004). We employ Gibbs sampling to learn *HSN-PAM* models because it often yields relatively simple algorithms for approximate inference in high-dimensional models.

For an arbitrary DAG, we need to sample a community path for each social actor given other variable assignments enumerating all possible paths and calculating their conditional probabilities. In the two-level community structure *HSN-PAM* model, each path contains the root, a super-community and a regular community. Since the root is fixed, we only need to jointly sample the super-community and regular community assignments for each social actor based on their conditional probability given observations and other assignments, integrating out the multinomial distributions, Θ (thus the time for each sample is in the number of possible paths). The following equation shows the conditional probability given the assignment of other regular and super communities. For social actor ω_j in SIP $\vec{s_t}$, we have:

$$p(\iota_{\omega 2} = k_2, \iota_{\omega 3} = k_3 \,|\, D, \iota_{-\omega}, \alpha, \beta) \propto \frac{n_{1k}^{(d)} + \alpha_{\alpha k}}{n_1^{(d)} + \sum_{k' \alpha_1 k'}} \times \frac{n^{(d)} + \alpha_{kp}}{n_k^{(d)} + \sum_{p'} \alpha_{kp'}} \times \frac{n_{pw} + \beta_w}{n_p + \sum_m \beta_m}.$$

Here, we assume that the root community is k_1, $\iota_{\omega 2}$ and $\iota_{\omega 3}$ correspond to super community and regular community assignments respectively. $\iota_{-\omega}$ is the community assignments or all other social actors. Excluding the current social actor, $n_x^{(d)}$ is the number of occurrences of community k_x in social interaction profile *sip*; $n_{xy}^{(d)}$ is the number of times community k_y is sampled from its parent k_x in SIP; n_x is the number of occurrences of regular-community k_x in the whole network and n_{xw} is the number of occurrences of social actor ω in regular-community k_x. Furthermore, α_{xy} is the yth component in α_x and β_w is the component for social actor ω in β.

Note that in the Gibbs sampling equation, we assume that the Dirichlet parameters are given. While *SSN-LDA* can produce reasonable results with a simple uniform Dirichlet, we have to learn these parameters for the super-communities in *TLD-HSN-PAM* since they capture different correlations among regular-communities. As for the root, we assume a fixed Dirichlet parameter. To learn α, we could use maximum likelihood or maximum a posterior estimation. However, since there are no closed-form solutions for these methods and we wish to avoid iterative methods for the sake of simplicity and speed, we approximate it by moment matching. In each iteration of Gibbs sampling, we update

$$\mu_{xy} = \frac{1}{N}\sum_d \frac{n_{xy}^{(d)}}{n_x^{(d)}};$$

$$\sigma_{xy} = \frac{1}{N}\sum_d \left(\frac{n_{xy}^{(d)}}{n_x^{(d)}} - \mu_{xy}\right)^2;$$

$$m_{xy} = \frac{\mu_{xy}(1-\mu_{xy})}{\sigma_{xy}} - 1;$$

$$\alpha x_y \propto \mu_{xy};$$

$$\sum_y (\alpha_{xy}) = \frac{1}{5}\exp\left(\frac{\sum_y \log(m_{yx})}{s_2 - 1}\right).$$

For each super-community k_x and regular-community k_y, we first calculate the sample mean μ_{xy} and sample covariance σ_{xy}^2. $n_{xy}^{(d)}$ and $n_x^{(d)}$ are the same as defined above. Then we estimate α_{xy}, the yth component in α_x from sample mean and variance. N is the number of social actors and s_2 is the number of regular communities.

Smoothing is important when we estimate the Dirichlet parameters with moment matching. From the equations above, we can see that when one regular community y does not get sampled from super community x in one iteration, α_{xy} will become zero. Furthermore, from the Gibbs sampling equation, we know that this regular community will never have the chance to be sampled again by this super community. We introduce a prior in the calculation of sample means so that μ_{xy} will not be zero even if $n_{xy}^{(d)}$ is zero for every SIP.

6 Experiments and evaluation

We evaluate two-level community structure *HSN-PAM* on three social network data collections. The first network is Zachary club network, a well-studied case in traditional social network analysis and the other two are collaboration networks. The three networks are representative in terms of sizes, which range from extremely small (34 nodes) to very large (398,831 nodes). The evaluation for this model is conducted in both descriptive and quantitative ways. First, we demonstrate the exemplary communities discovered by the algorithms for three social networks and briefly discuss the results. Thereafter, we calculate the likelihood values for a set of community numbers.

Throughout the experiments, we assume a fixed Dirichlet distribution with parameter 0.01 for the root node. We can change this parameter to adjust the variance in the sampled multinomial distributions. We choose a small value so that the variance is high and each SIP contains only a small number of super communities, which tends to make the super communities more interpretable. We treat the regular communities in the same

way as *SSN-LDA* and assume that they are sampled once for the whole corpus from a given Dirichlet with parameter 0.01. So the only parameters we need to learn are the Dirichlet parameters for the super communities and multinomial parameters for the regular communities.

6.1 Datasets

6.1.1 Zachary – a toy data set

The first dataset used in this paper is a small network, namely *Zachary Karate Club Network*, which has been used as a test case in a number of community discovery algorithms. *Zachary's Karate Club Network* was created by Wayne Zachary and had as few as 34 nodes in the network. Over the course of two years in the early 1970s, Wayne Zachary observed social interactions between the members of a karate club at a US university (Zachary, 1977). He constructed networks of ties between members of the club based on their social interactions both within the club and away from it. By chance, a dispute arose during the course of his study between the club's administrator and its principal karate teacher over whether to raise club fees, and as a result, the club eventually was split into two smaller clubs, centred on the administrator and the teacher.

6.1.2 CiteSeer

CiteSeer is a free public resource created by Kurt Bollacker, Lee Giles and Steve Lawrence in 1997–1998 at NEC Research Institute (now NEC Labs), Princeton, NJ. It contains rich information on the citation, co-authorship, semantic information for computer science literature. In this paper, we only consider the co-authorship information which constitutes a large-scale social network regarding academic collaboration with diversities spanning in time, research fields and countries.

Table 3 lists the statistics for *CiteSeer*. *CiteSeer* contains unconnected subnetworks. In particular, *CiteSeer* has 31,998 subgraphs and the size of the largest connected subnetwork of *CiteSeer* is 249,866. In this paper, we are only interested in discovering communities in the two largest subnetworks. Therefore, unless specially specify, we always mean the two subnetworks when referring *CiteSeer*.

Table 3 Statistics for datasets *CiteSeer* and *NanoSCI*

Dataset	*Size*	*PN*	*EN*	*AAP*	*SLC*
CiteSeer	398,831	716,793	1,181,133	1.65	249,866
NetSci	1,589	N/A	2741	N/A	1,589

Notes: *PN* denotes the number of papers; *EN* denotes the number of edges; *AAP* denotes the average author number per paper and *SLC* denotes the size of the largest component.

6.2 Empirical results

10% of the original datasets is held out as test set and we run the Gibbs sampling process on the training set for i iterations. In particular, in generating the exemplary communities, we set the number of the communities as 50 and the iteration numbers i as 1,000. In

perplexity computation, i is set as 300 in order to shorten the computation time. In both cases, α is set as $\frac{1}{k}$ and β is set as 0.01, where K is the number of the communities.

Figure 6 shows a consensus network structure extracted from Zachary's observation before the split. Encoding this network with both *DNES* and *IDNES* schemes, the results are shown in Table 4. The administrator is represented by node 1 and the instructor is represented by node 33.

Figure 6 The social interaction between the members of *Zachary's Karate Club*

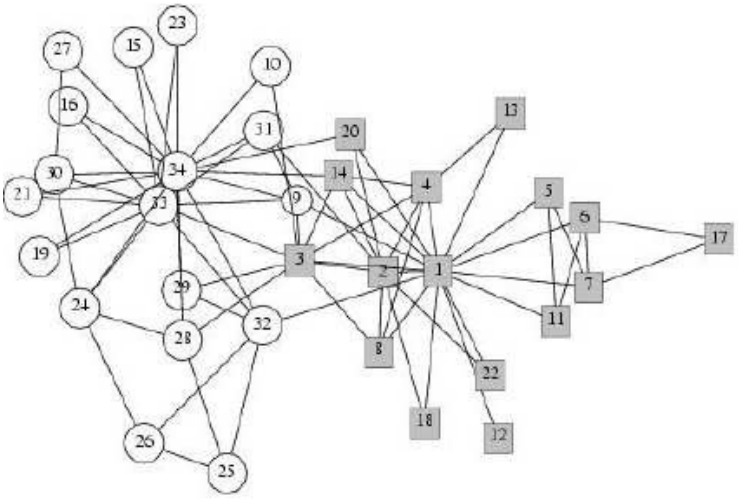


Table 4 demonstrates the final four regular communities discovered from the Zachary club. In particular, regular communities ι_1^r and ι_4^r belong to the same super community ι_1^s while regular communities ι_2^r and ι_3^r belong to the other super community ι_2^s. From Figure 6, super community ι_1^s corresponds to the cluster led by the administrator (node 1) while super community ι_2^s corresponds to the cluster led by the instructor (node 33). Note that there is only one node (node 10) that is misclassified by the *TLC-HSN-PAM* algorithm and node 9 is identified as a member for both super communities by the algorithm.

Table 4 Four regular communities discovered in Zachary Club

ι_1^r	1, 2, 3, 4, 5, 6, 10, 12
ι_2^r	9, 15, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34
ι_3^r	16, 19, 21, 23, 31
ι_4^r	7, 8, 9, 11, 13, 14, 17, 20, 22

Notes: The two super communities, ι_1^s is mainly composed of community ι_1^r and ι_4^s and ι_2^s is mainly composed of community ι_2^r and ι_3^r.

6.2.1 CiteSeer results

Tables 5, 6 and 7 demonstrate some exemplary communities that are discovered by *TLC-HSN-PAM* algorithm for the *CiteSeer* dataset with SIPs being created using *DNES* encoding scheme. Each community is shown with the top five researchers that have the highest probability conditioned on the community. Note that *CiteSeer* dataset was crawled from web and some authors were not recovered correctly, we keep the results in an 'as is' fashion. In this dataset, the number of super communities is set as 50 while the number of regular communities is set as 200.

These results illustrate that researchers from the regular communities that belong to the same super community are often interested in related subjects. For instance, the four top regular communities in ι_{48}^{s}, as shown in Table 5, include researchers that are working on 'signal processing' (ι_{63}^{s}), 'robot and learning' (ι_{19}^{r}), 'medical and image processing' (ι_{40}^{r}) and 'multimedia and learning' (ι_{185}^{r}) topics. Similarly, Table 6 lists four regular communities that belong to super community ι_{36}^{s}, including four relevant areas such as 'gent and AI' (ι_{179}^{r}), 'algorithm theory' (ι_{33}^{r}), 'multi-agent and distributed systems' (ι_{165}^{r}) and 'multimedia and learning' (ι_{185}^{r}). Table 7 demonstrates the four regular communities inside super community ι_{46}^{s}, including four relevant areas such as 'theory and distributed systems' (ι_{71}^{r}), 'distributed systems and cryptography theory' (ι_{192}^{r}), 'architecture and networks' (ι_{149}^{r}) and 'spatial databases' (ι_{136}^{r}). Note that a regular community can belong to many related super communities. For instance, regular community ι_{185}^{r} belongs to both super communities ι_{48}^{s} and ι_{36}^{s}.

Table 5 An illustration of four regular communities that belong to the 48th super community (ι_{48}^{s}) for the *CiteSeer* dataset after 1,000 iterations

Community 63	*Community 19*
Signal processing	*Learning, robot*
Marc Moonen	Manuela Veloso
Robert W. Dutton	Peter Stone
Brian L. Evans	Anthony Skjellum
Thomas H. Lee	Boi Faltings
Jung suk Goo	Edmund Burke
Community 140	*Community 185*
Medical, image	*Multimedia, learning*
Ron Kikinis	Thomas S. Huang
Ferenc A. Jolesz	Shih fu Chang
Simon K. Warfield	Anoop Gupta
Mark A. Musen	Gonzalo Navarro
Martha Shenton	Kathleen R. Mckeown

Notes: Each community is shown with the five researchers. The regular communities in ι_{48}^{s} are largely on learning and signal processing.

Table 6 An illustration of four regular communities that belong to the 36th super community for the *CiteSeer* dataset after 1,000 iterations

Community 179	*Community 33*
Agent AI	*Algorithm theory*
Nicholas R. Jennings	Micha Sharir
Simon Parsons	Pankaj K. Agarwal
Michael Wooldridge	John H. Reif
Peter Mcburney	Boris Aronov
Timothy J. Norman	Leonidas J. Guibas
Community 165	*Community 185*
Multi-agent, distributed	*Multimedia, learning*
Victor Lesser	Thomas S. Huang
Thomas Wagner	Shih fu Chang
David Kotz	Anoop Gupta
Michael Gerndt	Gonzalo Navarro
Heinz Stockinger	Kathleen R. Mckeown

Notes: Each community is shown with the five researchers.

Table 7 An illustration of 4 regular communities that belong to the 46th super community for the *CiteSeer* dataset after 1000 iterations

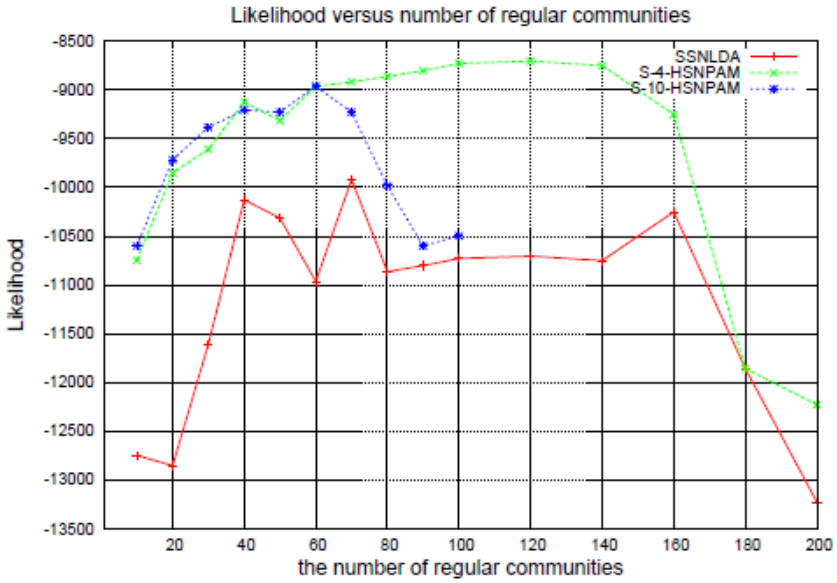
Community 71	*Community 192*
Theory distributed	*Distributed cryptography*
Nancy Lynch	Oded Goldreich
Danny Dolev	Moti Yung
John W. Lockwood	Manuel Hermenegildo
Jason Cong	Mihir Bellare
Riccardo Poli	Ran Canetti
Community 149	*Community 136*
Architecture network	*Spatial database*
William J. Dally	Michael H. Bhlen
Bernhard Steffen	Kristian Torp
Tiziana Margaria	Christian S. Jensen
Alon Y. Halevy	Heidi Gregersen
Daniel S. Weld	Daniel Thalmann

Notes: Each community is shown with the five researchers.

6.3 Likelihood analysis

In addition to empirical analysis on discovered communities, we also provide quantitative measurements to compare *HSN-PAM* with *SSN-LDA* approach. In this experiment, we use the same *CiteSeer* dataset and split it into two subsets with 90% and 10% of the data respectively. Then we learn the models from the larger data set and calculate likelihood for the smaller set. This is a common criterion for measuring the performance of statistical models in information theory. It indicates the uncertainty in predicting the occurrence of a particular social interaction given the parameter settings, and hence it reflects the ability of a model to generalise unseen data. In Figure 7, *SSN-LDA*, *S-4-HSNPAM* and *S-10-HSNPAM* illustrate the likelihood for *SSN-LDA* and *HSN-PAM* models when the number of super communities is set as four and ten, respectively. The *x*-axis represents the number of regular communities. This figure demonstrates that in general *HSN-PAM* is able to produce higher likelihood value. These curves can be used to detect the approximate optimal regular communities given the number of super communities.

Figure 7 Likelihood versus the number of communities (see online version for colours)

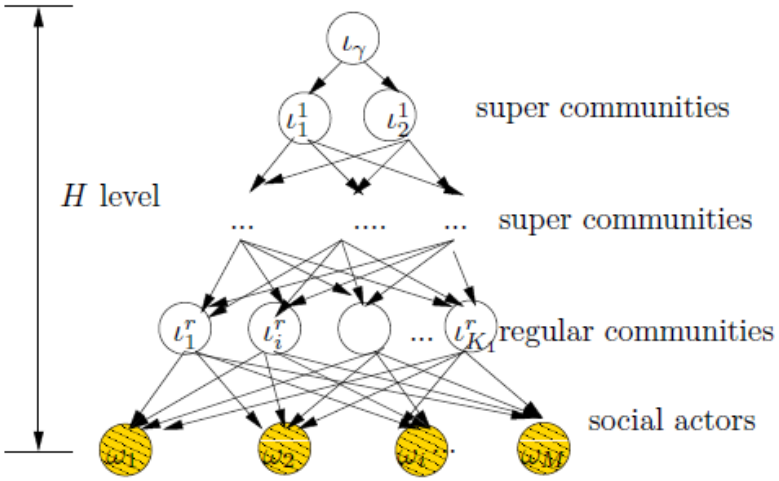


7 Discussions and future work

This paper focuses on two-level HSM-PAM model which indeed could be extended to arbitrary-level community scenario. Figure 8 demonstrates the DAG structure of an arbitrary level *HSN-PAM* model where each non-leaf interior node represents a community and a leaf node represents a social actor in social networks. The *HSN-PAM* model consists of two types of communities: regular communities and super communities. An interior node whose children are all leaves corresponds to a regular

community as in the *SSN-LDA* model. In addition to regular communities, the *HSN-PAM* model also includes super communities whose children contain interior nodes, or communities, thus representing a mixture over communities. With introduction of 'super communities', the *HSN-PAM* model is able to capture the correlation among social actors as well as correlations among communities. *Community path* is defined as a sequence of community ι_s of length H from the root to any regular community following the tree structure specified in Figure 8, $L_\omega : \left\langle \iota_{\omega_1}, \iota_{\omega_2}, \ldots, \iota_{\omega_{L_\omega}} \right\rangle$. ι_{ω_1} is always the root and ι_{ω_2} through $\iota_{\omega_{L_\omega}}$ are community nodes. ι_{ω_i} is a child of $\iota_{\omega_{(i-1)}}$.

Figure 8 Tree structure of arbitrary DAG-based *HSN-PAM* model, including arbitrary levels of super communities, one level of regular communities, and *M* social actors (see online version for colours)



To generate a SIP using DCM model, a sample is first drawn from the Dirichlet to get a multinomial distribution, then social actors in the SIPs are iteratively drawn based on the multinomial distribution. Each Dirichlet prior distribution Dir_{ι_i} in the DCM hierarchy is parameterised with a vector $\alpha^p_{\iota_i}$, which has the same dimension as the number of children in ι_i. Based on the graphical model in Figure 8, the generative process for a social actor's SIP $\overrightarrow{s_j}$ is a two-step process. The first step of the process is to sample the multinomial distribution for $\overrightarrow{s_j}$ based on the community variables' Dirichlet prior distributions. Subsequently, for each social interaction variable s_{jk} in $\overrightarrow{s_j}$, sample a potential community path leading from root node to the leaf node and then sample the social actor from the leaf node. Specifically, the process is described as follows:

1 *Sampling multinomial distribution:* Sample $\overrightarrow{\theta_{t1}}, \overrightarrow{\theta_{t2}}, \ldots, \overrightarrow{\theta_{tN}}$ from $Dir_1(\alpha_1)$, $Dir_2(\alpha_2)$,..., $Dir_N(\alpha_N)$, where $\overrightarrow{\theta_{tt}}$ is a multinomial distribution of community ι_i over its children.

2 *For each social actor s_k in the community:*

1 Sample a community path ι_ω of length $L_\omega : \left\langle \iota_{\omega_1}, \iota_{\omega_2}, \ldots, \iota_{\omega_{L_\omega}} \right\rangle$. ι_{ω_1} is always the root and ι_{ω_2} through $\iota_{\omega_{L_\omega}}$ are community nodes. ι_{ω_i} is a child of $\iota_{\omega_{(i-1)}}$ and it is sampled according to the multinomial distribution $\theta^{(j)}_{\iota_{\omega(i-1)}}$.

2 Sample social actor s_k from $\theta^{(j)}_{\iota_{s_{L_s}}}$.

Following this generative process, the joint probability of generating a SIP $\overrightarrow{s_j}$, the community assignments ι^j and the multinomial distributions $\theta^{(j)}$ is

$$P(\overrightarrow{s_j}, \overrightarrow{\iota_j}, \overrightarrow{\theta_j}, |\alpha) = \Pi_{i=1}^{K} P(\vec{\theta}_{\iota_i} | \alpha_i) \times \Pi_{s_j k \in \overrightarrow{s_j}} (\Pi_{i=2}^{H} P(\iota_{si} | \vec{\theta}_{\iota_s(i-1)}) (P(S_{jk} | \overrightarrow{\theta_{\iota_{\omega H}}}))$$

Integrating out $\theta^{(j)}$ and summing over ι^j, we calculate the marginal probability of a SIP $\overrightarrow{s_j}$ as:

$$P(\overrightarrow{s_j} | \vec{\alpha}) = \int \Pi_{i=1}^{K} P(\vec{\theta}_{\iota_i} | \alpha_i) \times \Pi_s \Sigma_{\iota_\omega} (\Pi_{i=2}^{L_\omega} P(\iota_{\omega i} | \vec{\theta}_{\iota_\omega(i-1)}) (P(\omega | \vec{\theta}_{L_\omega})) d\vec{\theta} \quad (5)$$

Finally, the likelihood of generating the complete network $\vec{S} = \{\overrightarrow{\omega_m}\}_{m-1}^{M}$ is determined by the product of the likelihoods of the independent nodes:

$$P(\vec{S} | \alpha) = \prod_{\overrightarrow{s_j}} (\overrightarrow{s_j} | \vec{\alpha})$$

8 Conclusions

Real-world social networks, while disparate in nature, often comprise of a set of loose clusters (a.k.a. communities), in which members are better connected to each other than to the rest of the network. In addition, such community structures are often hierarchical, reflecting the fact that some communities are composed of a few smaller sub-communities. Discovering the complicated hierarchical community structure can gain us deeper understanding about the networks and the pertaining community structures. This paper describes a hierarchical Bayesian model based scheme, namely *HSN-PAM*, for discovering probabilistic, hierarchical communities in social networks. This scheme is powered by a previously developed hierarchical Bayesian model. In this scheme, communities are classified into two categories: *super communities* and *regular communities*. Two different network encoding approaches are explored to evaluate this scheme on research collaborative networks, including *CiteSeer*. The experimental results demonstrate that *HSN-PAM* is effective for discovering hierarchical community structures in large-scale social networks.

References

Andrieu, C., Freitas, N., de Doucet, A. and Jordan, M.I. (2003) 'An introduction to MCMC for machine learning', *Machine Learning*, Vol. 50, Nos. 1–2, pp.5–43.

Bar-Yossef, Z., Guy, I., Lempel, R., Maarek, Y.S. and Soroka, V. (2006) 'Cluster ranking with an application to mining mailbox networks', in *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pp.63–74.

Blei, D.M., Ng, A.Y. and Jordan. M.I. (2003) 'Latent dirichlet allocation', *Journal of Machine Learning Research*, Vol. 3, pp.993–1022.

Börner, K., Maru, J.T. and Goldstone, R.L. (2004) *The Simultaneous Evolution of Author and Paper Networks*.

Clauset, A., Newman, M.E.J. and Moore, C. (2004) 'Finding community structure in very large networks', *Physical Review E*, Vol. 70, p.066111.

Flake, G.W., Lawrence, S. and Giles, C.L. (2000) 'Efficient identification of web communities', in *KDD'00: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.150–160.

Flake, G.W., Tarjan, R.E. and Tsioutsiouliklis, K. (2004) 'Graph clustering and minimum cut trees', *Internet Mathematics*, Vol. 1, No. 4, pp.385–408.

Freeman, L. (1977) 'A set of measures of centrality based upon betweenness', *Sociometry*, pp.35–41.

Garey, M.R. and Johnson, D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman & Co., New York, NY, USA.

Girvan, M. and Newman, M.E. (2002) 'Community structure in social and biological networks', *Proc Natl Acad Sci USA*, Vol. 99, No. 12, pp.7821–7826.

Griffiths, T. and Steyvers, M. (2004) 'Finding scientific topics', *Proceedings of the National Academy of Sciences*.

Heinrich, G. (2004) 'Parameter estimation for text analysis', Technical report.

Hopcroft, J., Khan, O., Kulis, B. and Selman, B. (2004) 'Tracking evolving communities in large linked networks', *Proc Natl Acad Sci USA*, Vol. 101, No. 1, pp.5249–5253.

Jing, L., Li, J., Ng, M.K., Cheung, Y.M. and Huang, J. (2009) 'Smart: a subspace clustering algorithm that automatically identifies the appropriate number of clusters', *International Journal of Data Mining, Modelling and Management*, Vol. 1, pp.149–177.

Krichel, T. and Bakkalbasi, N. (2006) 'A social network analysis of research collaboration in the economics community', in the *International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting*, 10–12 May 2006, Nancy, France.

Li, W. and McCallum, A. (2006) 'Pachinko allocation: DAG-structured mixture models of topic correlations', in *ICML*, pp.577–584.

MacKay, D.J.C. (2002) *Information Theory, Inference & Learning Algorithms*, Cambridge University Press, New York, NY, USA.

Minka, T. and Lafferty, J. (2002) 'Expectation-propagation for the generative aspect model'.

Newman, M.E. (2004a) 'Coauthorship networks and patterns of scientific collaboration', *Proc Natl Acad Sci USA*, Vol. 101, No. 1, pp.5200–5205.

Newman, M.E. (2004b) 'Fast algorithm for detecting community structure in networks', *Physical Review E*, Vol. 69, p.066133.

Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. (2005) 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature*, pp.435–814.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004) 'The author-topic model for authors and documents', in *AUAI'04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp.487–494, Arlington, Virginia, United States.

Scott, J.P. (2000) *Social Network Analysis: A Handbook*, SAGE Publications.

Si, L. and Jin, R. (2005) 'Adjusting mixture weights of Gaussian mixture model via regularized probabilistic latent semantic analysis', in *PAKDD*, pp.622–631.

Wang, X. and McCallum, A. (2006) 'Topics over time: a non-Markov continuous-time model of topical trends', in *KDD*, pp.424–433.

Wilkinson, D.M. and Huberman, B.A. (2004) 'A method for finding communities of related genes', *Proc Natl Acad Sci USA*, Vol. 101, No. 1, pp.5241–5248.

Zachary, W. (1977) 'An information flow model for conflict and fission in small groups', *Journal of Anthropological Research*, pp.452–473.

Zhang, H., Giles, C.L., Foley, H.C. and Yen, J. (2007a) 'Probabilistic community discovery using hierarchical latent Gaussian mixture model', in *AAAI*, pp.663–668.

Zhang, H., Qiu, B., Giles, C.L., Foley, H.C. and Yen, J. (2007b) 'An LDA-based community structure discovery approach for large-scale social networks', in *IEEE International Conference on Intelligence and Security Informatics*, pp.200–207.

Zhang, J., Ackerman, M.S. and Adamic, L. (2007c) 'Expertise networks in online communities: structure and algorithms', in *WWW'07: Proceedings of the 16th International Conference on World Wide Web*, pp.221–230, New York, NY, USA.