Chapter 1

MULTI-RELATIONAL CHARACTERIZATION OF DYNAMIC SOCIAL NETWORK COMMUNITIES

Yu-Ru Lin, Hari Sundaram and Aisling Kelliher
School of Arts, Media and Engineering
Arizona State University, Tempe AZ, USA

1. Introduction

The emergence of the mediated social web – a distributed network of participants creating rich media content and engaging in interactive conversations through Internet-based communication technologies – has contributed to the evolution of powerful social, economic and cultural change. Online social network sites and blogs, such as Facebook, Twitter, Flickr and LiveJournal, thrive due to their fundamental sense of "community". The growth of online communities offers both opportunities and challenges for researchers and practitioners. Participation in online communities has been observed to influence people's behavior in diverse ways ranging from financial decision-making to political choices, suggesting the rich potential for diverse applications. However, although studies on the social web have been extensive, discovering communities from online social media remains challenging, due to the interdisciplinary nature of this subject. In this article, we present our recent work on characterization of communities in online social media using computational approaches grounded on the observations from social science.

Motivation – human community as meaning-making eco-system. A key idea from situated cognition is that knowledge is fundamentally situated within the activity from which it is developed [6]. Brown, Collins and Duguid [6] offers an analysis of how we make meaning with the lexical and grammatical resources of language – people can interpret indexical expressions (containing such words as *I*, *you* , *here*, *now*, *that*, etc.) only when they can find what the indexed words might refer to. The concept of *indexicality* suggests [6] that "*knowledge, which comes coded by and connected to the activity and environment in which it is developed, is spread across its component parts, some of which are in the mind and some in the world much as the final picture on a jigsaw is spread across its component pieces.*" In other words, knowledge does not solely reside in the mind of an individual, but is distributed and shared among co-participants in authentic situations. The meaning-making eco-social systems, denoted by Lemke [25], shape and create meaning not by individual components (people, media, objects, etc.), but by their co-participation in an activity situation.

Being influenced by this theory, we believe that semantics is an *emergent artifact of human activity that evolves over time*. Human activity is mostly social, and the social networks of human are conceivable loci for the construction of meaning. Hence, it is crucial to identify real human networks as *communities of people interacting with each other through meaningful social activities, and producing stable associations between concepts and artifacts in a coherent manner*.

Motivating applications. The discovery of human communities is not only philosophically interesting, but also has practical implications. As new concepts emerge and evolve around real human networks, community discovery can result in new knowledge and provoke advancements in information search and decision-making. Example applications include:

- Context-sensitive information search and recommendation: The discovered community around an information seeker can provide context (including objects, activities, time) that help identify most relevant information. For example, when a user is looking at a particular photo, the community structure may be used to identify peers or objects likely co-occurring with the photo.
- Content organization, tracking and monitoring: The rapid growth of content on social media sites creates several challenges. First, the content in a photo stream (either for a user or a community) is typically organized using a temporal order, making the exploration and browsing of content cumbersome. Second, sites including Flickr provide frequency based aggregate statistics including popular tags, top contributors. These aggregates do not reveal the rich temporal dynamics of community sharing and interaction. Community structure may be used to reflect the social sharing practice and facilitate the organization, tracking and monitoring of user-generated social media content.
- Behavioral prediction: Studies have shown that individual behaviors usually result from mechanisms depending on their social networks, e.g. social ebeddedness [17] and influence [13]. Community structure that accounts for inherent dependencies between individuals embedded in a social network can help understand and predict the behavioral dynamics of individuals.

Data characteristics and challenges. Large volumes of social media data are being generated from various social media platforms including blogs, FaceBook, Twitter, Digg, Flickr. The key characteristics of online social media data include:

- Voluminous: Recent technological advances allow hundreds of millions of users to create social and personal content instantly. The amount of data and the rate of data production can be enormous.
- Dynamic: Users' online actions are constantly archived with timestamps. These online activity records enable a fine-grained observation on the dynamics of human interactions and interests.
- Context-rich: Most social media platforms allow a wide array of actions for managing and sharing media objects – e.g. uploading photos, submitting and

commenting on news stories, bookmarking and tagging, posting documents, creating web-links, as well as actions with respect to other users (e.g. sharing media and links with a friend), or on media objects produced by other users. The complex social interactions among users result in multi-relational network data.

The large-scale, fine-grained, rich online interaction records pose new challenges on community discovery:

- Lack of well-defined attributes: Traditional studies of human communities are often based on fixed demographic characteristics [3,38]. Within online social networks, individuals may shift fluidly and flexibly among communities depending on their online social actions (e.g. who they recently interact with, what they recently share with each other, etc.) [4].
- Limitation in network centric analysis: Classical social network analyses mostly focus on static interpersonal relationships (e.g. self-reported friendships), with a primary interests on the graph topological properties These studies range from well-established social network analysis [38] to recent successful graph mining algorithms such as HITS [22], PageRank [5] and spectral analysis [36]. These methods are limited in discovering important aspects of online communities since the interpersonal relationships may evolve with their online interactions and may involve rich media contexts (e.g. tags, photos, time, and space).
- Scalability requirement: The Internet scale social network data requires a scalable analysis framework to support community discovery based on information latent in the multi-relational social network data [23].

Problem overview. We are interested in characterizing human communities that emerge from online interpersonal social activities. Given the challenges discussed above, we have focused on the following research problems:

- *How to identify meaningful interpersonal relationship from online social actions*? In social network literature, community discovery usually refers to detecting cohesive subgroups of individuals within networked data collected based on well-defined social relationship such as self-reported friendship [38]. Characterization of communities in online social networks deviates from traditional social network analysis because the social meaning of the networks is not definite.
- *How to identify sustained evolving communities from dynamic networks*? Online communities are temporal phenomena emerge from sustained human actions and interests, and the actions and interests may evolve over time. Traditional analysis of social networks focuses on the properties of a static graph (aggregation or snapshot of network), which overlooks the temporal characteristics of communities. Discovering communities based sustained activities and at the same time characterizing their temporal evolution is challenging.

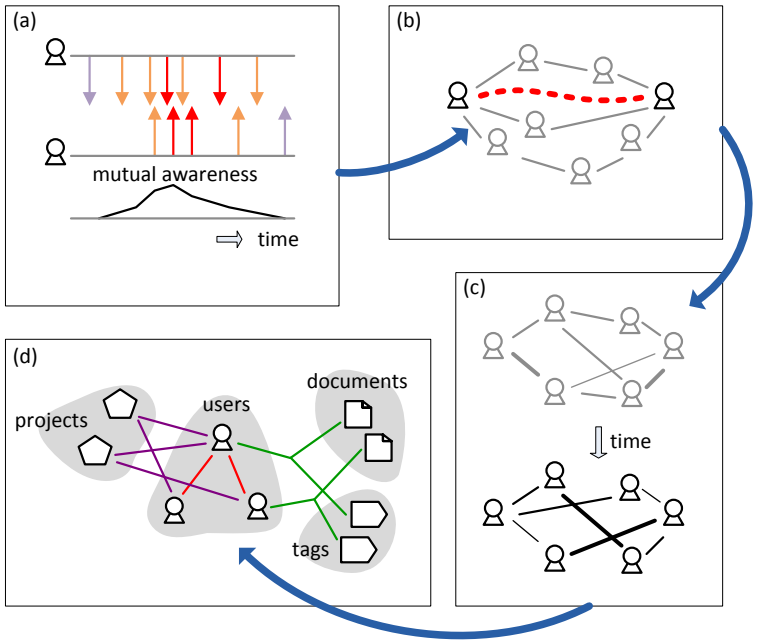


Figure 1: Our work concerns multiple aspects on community analysis: (a) Mutual awareness – a bi-directional relationship indicating how well a pair of bloggers is aware of each other, as fundamental property of a community. (b) Mutual awareness expansion – a random walk based distance measure which estimates the probability that two bloggers are aware of each other on the network. (c) FacetNet – for analyzing communities and their evolutions in a unified process. (d) MetaFac – the first graph-based multi-tensor factorization framework for analyzing the dynamics of heterogeneous social networks.

- *How to identify communities with rich interaction context*? Online social media websites (e.g. Flickr, Facebook) enable rich interaction between media and users, as well as complex social interactions among users – two users may share similar tags or read the same feeds. Discovery communities from such complex social interactions pose technical challenges that involve dealing with networked data consisting of multiple co-evolving dimensions, e.g. users, tags, feeds, comments, etc. Existing high dimensional data mining techniques are usually computational intensive and not suitable for dealing with large scale social networked data.

Our approach. Our work concerns approach to the three problems (see Figure 1 for illustrating summarization):

- *Mutual awareness*: We propose mutual awareness (ref. Figure 1(a)), a bi-directional relationship indicating how well a pair of bloggers is aware of each other, as fundamental property of a community. We provide computational definition to quantify mutual awareness and use it as a feature for community discovery. Then, we capture the amount of mutual awareness expanding on the entire network using a random walk based distance measure, commute time, which estimates the probability that two bloggers are aware of each other on the network (ref. Figure 1(b)). We propose an efficient iterative mutual awareness expansion algorithm to extract communities, which partitions the network by maximizing the commute time distance between two sets of bloggers. The experimental results for community extraction in terms of standard evaluation metrics are promising.
- *FacetNet*: We introduce the FacetNet framework to analyze communities and their evolutions in a unified process. In our framework (ref. Figure 1(c)), the community structure at a given timestep is determined both by the observed networked data and by the prior distribution given by historic community structures. Algorithmically, we propose the first probabilistic generative model for analyzing communities and their evolution. The experimental results suggest that our technique is scalable and is able to extract meaningful communities based on social media context. (e.g., dramatic change in a short time is unlikely).
- *MetaFac*: We propose MetaFac, the first graph-based tensor factorization framework for analyzing the dynamics of heterogeneous social networks (ref. Figure 1(d)). In this framework, we introduce metagraph, a novel relational hypergraph representation for modeling multi-relational and multi-dimensional social data. Then we propose an efficient multi-relational factorization algorithm for latent community extraction on a given metagraph. Extensive experiments on large-scale real-world social media data and from the enterprise data suggest that our technique is able to extract meaningful communities that are adaptive to social media context.

Organization. The rest of this article is organized as follows. Section 2 presents community discovery based on mutual awareness. Section 3 presents method for extracting sustained evolving communities. Section 4 presents method for extracting communities with rich interaction context. Section 5 concludes with future directions.

2. Actions, Networking and Community Formation

In this section we study the online blog network and propose computational approach for discovering communities in the blogosphere. Blogs (or weblogs) have become popular self-publishing social media on the Web. Although they are a type of websites, the analysis of blog communities is different from traditional Web analysis literature. The differences lie in the different semantics and structures of the hyperlinks in the context of blogosphere. A blog is typically used as a tool for communication. Driven by an event (such as a real-world news), bloggers publish entries that refer to each other. Thus links among blog entries are considered to be *interactions* between two bloggers and have significant temporal locality. On the Web, it is common that a new page refers to a relevant page that exists for a long time, such as an authoritative page [5,22]. A "community of web pages" due to the links of *relevance* is thus different from a "community of bloggers" formed due to the links of interactions. The analysis of blog network also deviates from traditional social network analysis [38] because the *social meaning* of the blog network is not as well-defined as in traditional social networks (e.g. links represent friendship). Hence, community discovery in the blogosphere requires a new analytical framework grounded in the unique characteristics of the blog media.

2.1 Mutual Awareness and Community Discovery

The notion of *virtual community*, or online community, has been discussed extensively in prior research. Rheingold [34] defined virtual communities to be "social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationship in cyberspace." Jones [20] considered four characteristics as the necessary conditions for the formation of a virtual community: interactivity, communicators, virtual common-public-place where the computer-mediated communication takes place, and sustained membership. These conditions echo Garfinkel's observation on the necessity of mutually observable actions [14]. The same idea that interactivity forms a social reality has also been discussed by Dourish [12]. According to Dourish, interaction involves presence (some way of making the actors present in the locale) and awareness (some way of being aware of the other's presence). In what Dourish called an *action community*, members share the common sense understandings through the reciprocal actions. The common aspect of the prior work is the emphasis on the significance of action and interaction in online communities. However, little work has studied the counter perspective – how to discover communities due to actions.

We introduce *mutual awareness* that is fundamental to blog community formation. By mutual awareness of action we mean that individual blogger actions must lead to bloggers becoming aware of each other's presence. The idea is in the light of Locale theory [12] that discusses how social organization of activity is supported

in different spaces. While the domains of activity must provide means for the community members to act, the space must also accord members' presence and facilitate mutual awareness.

Note that mutual awareness may be related to, but is different from, *link reciprocity*, which refers to the tendency of vertex pairs to form mutual connections between each other [15]. It is also related to *tie strength* discussed in the social network literature [16] – "the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." However, quantifying tie strength based on these elements remains challenging. In fact, mutual awareness can be considered as a mechanism for tie strength situated in particular communication media.

2.2 Extracting Communities based on Mutual Awareness Structure

We propose a computational approach for community discovery in the blogosphere. Grounded on the actions of individual bloggers, we propose to discover community based on the idea of mutual awareness. We propose a computational definition for determining mutual awareness in the blog network. Then, using mutual awareness as features, we propose methods for extracting blog communities.

2.2.1 Computable Definition for Mutual Awareness

Let us examine the actions of individual bloggers – how bloggers read and communicate ideas with other bloggers. The bloggers can act in the blogosphere, in several ways: surf / read, create entries (containing entry-to-entry links, entry to blog / web, or no link), comment or change blogroll. Some actions (e.g. surf/read) may be hidden, while others may be observable.

How a specific blogger action leads to mutual awareness may depend on (a) if the action is mutually observed, and (b) the importance of the action for the blogger who performs the action. Note that some blogger actions are not observable by other bloggers. For example, let us consider two hypothetical bloggers, Mary and John. Let us assume that Mary creates an entry with a hyperlink that points to John's blog. In this case John would be unaware of Mary's entry. On the other hand, if Mary leaves a comment on John's entry, then John is immediately aware of her presence. If Mary mostly leaves comments on other bloggers, and the importance of a comment for Mary is low – while many bloggers are aware of Mary, she may not feel that she in engaged in dialogue with them. The assessment of mutual awareness is the first step toward the discovery of blog communities.

We thus characterize mutual awareness as follows (see Figure 2 for an illustration): mutual awareness between two bloggers is affected by the type of action, the number of actions for each type, and when the action occurred. It depends on sus-

tained actions – it increases if there are follow-up actions that lead to mutual awareness and decreases if actions are not sustained over time.

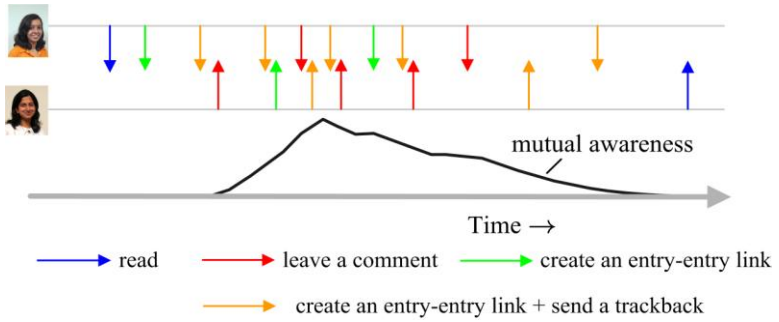


Figure 2: Mutual awareness between two bloggers is affected by the type of actions, the number of such actions, and when such actions occur. The arrow direction indicates the source and the destination blogger on whom the action is performed. A mutual awareness curve is plotted to show the action impacts.

We represent the set of bloggers in the blogspace as a weighted directed graph $G = (V, E)$, where each node $v \in V$ represents a blogger, each edge between any pair of nodes u and v represents an action performed by u with respect to v. The weight on each edge $f(u,v)$ indicates the mutual awareness between two bloggers u and v. The corresponding matrix $\mathbf{M}$ with each entry $\mathbf{M}_{uv} = f(u,v)$ is called mutual awareness matrix and is defined as follows.

Definition 1 (*Mutual awareness matrix*):

$$\mathbf{M} = \sum_k \alpha_k \min(\widehat{\mathbf{X}}_k, \widehat{\mathbf{X}}_k^{\,T}) \quad (1)$$

where the index k is used to denote a specific action (e.g. leaving a comment, or creating an entry-to-entry link) and α_k represent the importance of the actions and is usually empirically determined. $\widehat{\mathbf{X}}_k$ is aggregated action matrix for the action type k. Since mutual awareness due to earlier actions will gradually diminish, the temporal effect can be modeled as a decaying exponential function:

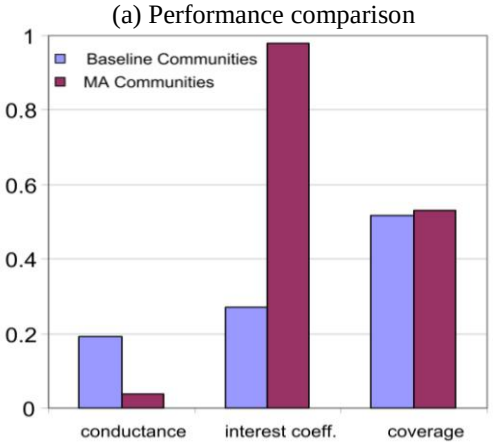
$$\widehat{\mathbf{X}}_k = \sum_{t=t_0}^{T} \mathbf{X}_{k;t} e^{-\lambda_k (T-t)} \quad (2)$$

where λ_k is the decaying factor for the action type k. Different types of actions may decay at different rate. $\mathbf{X}_{k;t}$ denotes all-pair type-k actions occurring at time t, and $\widehat{\mathbf{X}}_k$ aggregates these actions from time t_0 to time T.

Mutual awareness is a bi-directional relationship indicating how well a pair of bloggers is aware of each other. This semantics results in a symmetric mutual awareness matrix. The reciprocity condition $\min(\cdot,\cdot)$ in Eq. (1) makes the possibility of both bloggers being aware of each other to be high.

Empirical evaluation. We have studied the effectiveness of mutual awareness (MA) matrix on real-world blog datasets [26]. We use the subgroup extraction procedure described in [26] to extract subgroups, and evaluate the quality of these subgroups in terms of different metrics. Compared with the baseline adjacency matrices (with entries indicating the total number of entry-to-entry links), the subgroups extracted using mutual awareness matrices are usually of higher quality. The quality evaluation is based on several metrics, including conductance, interesting coefficient, etc. (see the definitions of the metrics in [26]). Figure 3(a) shows the performance comparison of results from the WWE 2006 Workshop Blog Dataset [26]. An example subgroup is shown in Figure 3(b) – the group is observed to have cohesive topic about mystery novels based on the top keywords from their blog contents.

(a) Performance comparison



(b) Case study

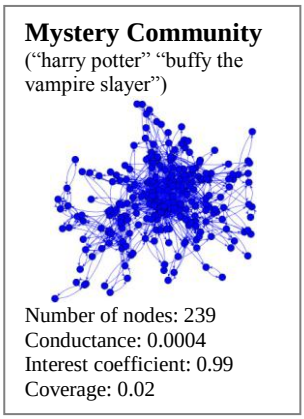


Figure 3: (a) Performance comparison between the Baseline communities and the MA communities in terms of metrics conductance, interest coefficient, and coverage. (b) An example subgroup cohesive topic about mystery novels, extracted using MA matrix.

2.2.2 Mutual Awareness Expansion

We extend the idea of mutual awareness to community extraction. Mutual awareness quantifies the relationship between two bloggers. To extract a set of bloggers having high mutual awareness, we hypothesize how mutual awareness expands in a blog network:

- Transitivity: One could become aware of a member without direct interaction since he or she can observe his or her direct peers interacting with other people. Thus awareness is transitive. (The transitivity property in social network has been first examined in Travers and Milgram's well-known small world experiment [37], which motivates our proposed algorithm [27].)

- Reciprocity: Such transitive awareness must be reciprocal. If expansion of awareness is only one directional, one might not feel belonging to the community
- Frequency: The amount of observed interaction must be sufficient for members to feel connected to each other.

We characterize such *mutual awareness expansion* process by a random walk model. The probability that two bloggers are aware of each other on the entire network is quantified using the random walk expected length between two nodes corresponding to the bloggers. We refer to this expected length as *symmetric social distance*. It is computed as follows: Given a direct graph $G = (V,E)$ and the mutual awareness matrix $\mathbf{W}$ associated with G, the random walk on G is defined to be the Markov chain with state space V and the transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with element $d_{ii} = \sum_j w_{ij}$. A random walker at a node i on G will follow the transition probability $p_{ij} = P_{ij}$ to visit the next node j. Note that by construction $w_{ii} = \sum_j w_{ij}$ (i.e. $p_{ii} = 1/2$) for $i \neq j$.

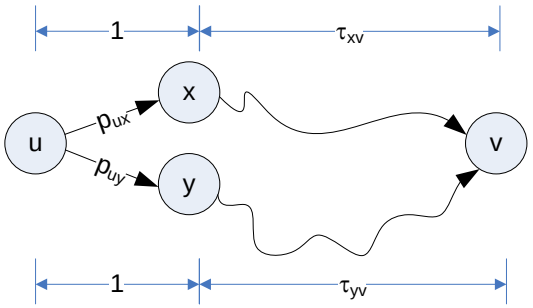


Figure 4: Transitive awareness property – the social distance from node u to v is defined by the expected number of steps before node v is visited, starting from node.

Let $\tau_{u\to v}$ denote the one-way social distance from node u to v, i.e. the expected number of steps to reach node v from node u. We define $\tau_{u\to v}$ to have the *transitive awareness* property:

Definition 2 (*Transitive awareness property*):

$$\tau_{u\to v} = \begin{cases} 1 + \sum_{(u,x)\in E} p_{ux}\,\tau_{x\to v} & if\ u \neq v \\ 0 & if\ u = v \end{cases} \qquad (3)$$

where p_{ux} is the transition probability from u to x. The equation can be illustrated in Figure 4: To reach v from u, the random walker takes one step to get to the next node x with transition probability p_{ux}, and then calculates the rest expected distance to v. The symmetric social distance is defined by $\tau_{u\leftrightarrow v} = \tau_{u\to v} + \tau_{v\to u}$.

Solution. Based on property in Eq. (3), the solution for $\tau_{u\leftrightarrow v}$,denoted by $\tilde{\tau}_{u\leftrightarrow v}$, can be derived by using Green's function [10]:

$$\tilde{\tau}_{u\leftrightarrow v} = vol\sum_{i=2}^{k}\frac{1}{\lambda_i}(\phi_i(u)-\phi_i(v))^2 \quad (4)$$

where $vol=\sum_{u,v\in V} w_{uv}$, ϕ_i's and λ_i's ($0=\lambda_1<\lambda_2\leq \ldots \leq\lambda_n$) are the eigenvectors and corresponding eigenvalues of the Laplacian matrix **L**=**D**-**W**. The solution is computed by truncating after the k-th smallest eigen-pair for $k<n$. Intuitively, $\tau_{u\leftrightarrow v}$ is the random walk expected path length from u to v and back to u, which takes into account the indirect interactions between u and v derived by their interactions with other nodes over the entire network.

Community extraction. We use the symmetric social distance as a criterion to extract a community. Given a set of bloggers V, a subset S from V can be seen as a community if the symmetric social distance among members of S is short compared to those with non-members. Therefore, we iteratively split the set V into two sets S and $V\backslash S$ by maximizing the symmetric social distance as follows:

$$S = \text{argmax}_{S\subset V}\ \omega(S,V)\sum_{u\in S, v\in V\backslash S}\tau_{u\leftrightarrow v} \quad (5)$$

Where $\omega(S,V)$ is a weighted function used to obtained desirable properties (e.g. balance partition) of the set of communities. Details of the algorithm can be found in [27].

Empirical evaluation. We compare our community extraction method with well-known baseline clustering algorithms, including the kernel k-means [11], normalized cut [36] and iterative conductance cutting [21]. The results indicate that our method outperforms all baseline methods in terms of low conductance, high coverage, and low entropy [27].

2.3 Application: Query-sensitive Community Extraction

We apply the community extraction algorithm to the extraction of *query-sensitive communities*, i.e. blog communities that have a strong content related theme with respect to a given query. We summarize the idea as follows (see [27] for more details):

- Step 1: Given a query topic Q, extract query-sensitive graph G_Q to represent interactions relevant to the topic.
- Step 2: Given G_Q, extract communities as described in Section 2.2.

In the first step, we construct a weighted action matrix with respect to a query Q. Q contains query keywords that represent the given topics, e.g. "Katrina", "london bomb", etc. The weight of an interaction is determined by the relevance score of the blog content involved in the interaction. Because the query keywords in Q are relatively short, in order to further incorporate relevant blogs in G_Q, we compute the "query relevancy" by employing a web-based similarity function [27,35].

Figure 5 shows an example from our experiment. In this case, we extract communities with respect to the query keyword is "katrina", which is about a natural disaster caused by the hurricane Katrina in August 2005. In order to understand

the "meaningfulness" of the extracted communities, we employ a heuristic method [27] to examine the relationship between the topic and the communities over time: We connect those communities extracted from different time snapshots based on an interaction similarity measure[1]. In Figure 5, each node represents a detected community where the communities detected during the same week are aligned horizontally and the communities at different time snapshots are connected by arrows. The grayscale of an arrow is proportional to the interaction similarity between the two communities. The node size reflects the number of community members and the reddish shade of node is proportional to the query relevancy for the keyword "katrina". More saturated node represents more relevant community.

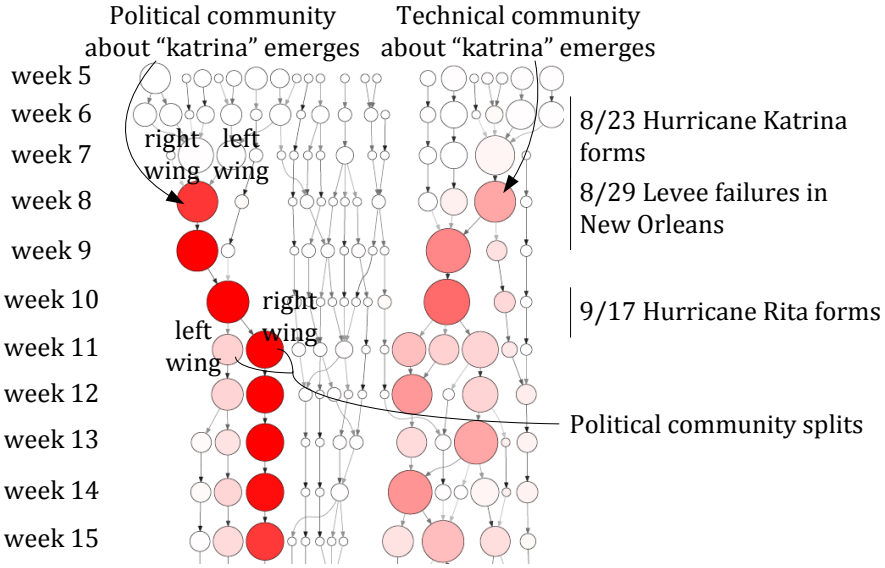


Figure 5: Communities with respect to the query "katrina." The node size reflects the number of community members and the reddish shade of node is proportional to the query relevancy for the keyword "katrina."

The interactions among the extracted communities are quite interesting. Two dominated communities, one with a focus on politics (shown on the left) and the other with a focus on technology (shown on the right), emerged and evolved due to the Katrina event. When the event Katrina occurred at week 7, we found community

[1] A more systematical solution will be presented in next section.

merged from left-wing and right-wing members, due to debates about the government response as well as cooperation of fund-raising. Later at week 11, the community split into stable communities that correspond to their political preferences. The results suggest that for queries such as “Katrina”, community extraction help identifies people with different viewpoints based on their sustained interactions.

Summary. A key idea in this work was that observable actions lead to the emergence of human communities, and awareness expansion was critical to community formation. We provide computational definition to quantify mutual awareness and use it as a feature to extract subgroups in the blogosphere. The effectiveness of mutual awareness features is verified on real-world blog datasets.

We showed how to detect blog communities based on mutual awareness expansion, given a specific query. We proposed a symmetric social distance measure that captures the expansion process and use it to detect communities. The community evolution with respect to a query reveals interesting community dynamics.

There are some open issues in this work. (1) The communities are independently extracted at consecutive timesteps and then the evolutions are characterizes to explain the difference between these communities over time. Such a two-stage approach may result in community structures with high temporal variation, and undesirable evolutionary characteristics may have to be introduced in order to explain such high variation in the community structures. A more appropriate approach is to analyze communities and their evolutions in a unified framework. (2) In order to extract communities with strong content related themes, we construct networks with query-dependent edge weights with respect to given concepts. However, the theme of a community may not be known in advance, and it may emerge and evolve over time, depending on the content and context associated with the interactions. This will require a new approach for extracting communities from rich interaction contexts. We shall discuss these directions in next two sections.

3. Analyzing Communities and Evolutions in Dynamic Network

In this section, we present the FacetNet framework that analyzes communities and their evolutions in a unified process. Traditional analysis of social networks treats the network as a static graph, where the static graph is either derived from aggregation of data over all time or taken as a snapshot of data at a particular time. These studies range from well-established social network analysis [38] to recent successful applications such as HITS [22] and PageRank [5]. However, this research omits one important feature of communities in networked data – the temporal evolution of communities.

3.1 Sustained Membership, Evolution and Community Discovery

If evolution is a nature characteristic of human communities, how are they different from a chance meeting of casual individuals? Jones [Jones 1997] argued that a virtual community is not a chance meeting of casual individuals but should involve long term, meaningful conversations among humans, and this condition suggests that there should be a minimal level of sustained membership. Lemke described *community ecology* as follows: "they have a relevant history, a trajectory of development in which each stage sets up conditions without which the next stage could not occur," "the course of their development depends in part on information laid down (or actively available) in their environments from prior (or contemporary) systems of their own kind."

Recently, there has been a growing body of analytical work on communities and their temporal evolution in dynamic networks (e.g. [2,27,32]). However, a common weakness in these studies, is that communities and their evolutions are studied separately – usually community structures are independently extracted at consecutive timesteps and then in retrospect, evolutionary characteristics are introduced to explain the difference between these community structures over time. Such a two-stage approach has two issues: (a) At each timestep, communities are extracted without considering sustained membership (temporal smoothness of clustering). (b) It may result in community structures with high temporal variation, and undesirable evolutionary characteristics may have to be introduced in order to explain such high variation in the community structures.

Sustained membership is the key to discovery time-evolving communities. We introduce the FacetNet framework to extract sustained and evolving communities from dynamic social networks.

3.2 Extracting Sustained Evolving Communities

We present the formulation of our model, and describe how to extract communities and their evolutions from the solution of our model.

3.2.1 *Problem Formulation*

We assume that edges in the networked data are associated with discrete time-steps. We use a snapshot graph $G_t=(V_t,E_t)$ to model the interactions at time t, where in G_t, each node $v_i \in V_t$ represents an individual, each edge $e_{ij} \in E_t$ represents the presence of interactions between v_i and v_j, and $w_{t;ij} = (W_t)_{ij}$ denotes the edge weight of e_{ij}. Note the edge weight can represent mutual awareness, or more generally, the frequency of interactions between nodes i and j observed at time t. Assuming G_t has n nodes, $W_t \in \Re_+^{n\times n}$ (nonnegative matrix of size $n\times n$) is the corresponding weight matrix for G_t. Over time, the interaction history is captured by a sequence of snapshot graphs $G_1, \ldots, G_t, \ldots$ indexed by time.

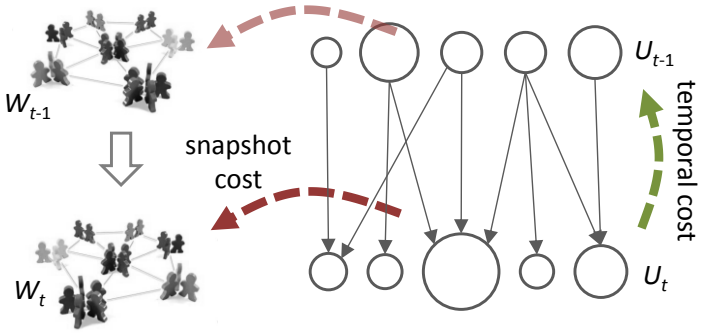


Figure 6: The community structure at time t (denoted as U_t) is determined by (1) the data observed at time t (denoted as W_t), and (2) the community structure at time t-1 (denoted as U_{t-1}).

The basic principles (as illustrated in Figure 6) behind our models are the community structure at time t is determined by (1) the data observed at time t (i.e. W, which is short for W_t), and (2) the community structure at time t-1. We propose to use the community structure at time t-1 (already extracted) to regularize the community structure at current time t (to be extracted). To incorporate such a regulation, we introduce a cost function to measure the quality of community structure at time t, where the cost consists of two parts—a *snapshot cost* and a *temporal cost*:

$$cost = \alpha \cdot \mathcal{CS} + (1-\alpha) \cdot \mathcal{CT} \qquad (6)$$

This cost function is first proposed by Chakrabarti et al. [8,9] in the context of evolutionary clustering. In this cost function, the snapshot cost $\mathcal{CS}$ measures how well a community structure fits W, the observed interactions at time t. The temporal cost $\mathcal{CT}$ measures how consistent the community structure is with respect to historic community structure (at time t-1). The parameter a is set by the user to control the level of emphasis on each part of the total cost.

A community structure at time t should fit W well, where W is the observed interaction matrix at time t. This requirement is reflected in the snapshot cost $\mathcal{CS}$ in the cost function Eq. (6). We adopt a stochastic block model first proposed in [39].

Assume that there exist m communities at time t, and that the interaction w_{ij} is a combined effect due to all the m communities. That is, we approximate w_{ij} using a mixture model $w_{ij}= \sum_{k=1}^{m} p_k \cdot p_{k\to i} \cdot p_{k\to j}$, where p_k is the prior probability that the interaction w_{ij} is due to the k-th community, $p_{k\to i}$ and $p_{k\to j}$ are the probabilities that an interaction in community k involves node v_i and v_j, respectively. Written in a matrix form, we have $W \approx X\Lambda X^T$, where $X \in \Re_+^{n\times m}$ is a non-negative matrix with $x_{ik} = p_{k\to i}$ and $\sum_i x_{ik} = 1$. In addition, Λ is an $m\times m$ non-negative diagonal matrix with $\lambda_k = p_k$, where λ_k is short for λ_{kk}. Matrices X and Λ (or equivalently, their product $X\Lambda$) fully characterize the community structure in the mixture model. Based on this model, we define the snapshot cost $\mathcal{CS}$ as the error introduced by such an approximation, i.e.,

$$\mathcal{CS} = D(W||X\Lambda X^T) \tag{7}$$

where $D(A||B) = \sum_{i,j}(a_{ij}\log\frac{a_{ij}}{b_{ij}} - a_{ij} + b_{ij})$ is the KL-divergence between A and B. The snapshot cost is high when the approximate community structure $X\Lambda X^T$ fails to fit the observed data W well.

In the cost function Eq. (6), the temporal cost $\mathcal{CT}$ is used to regularize the community structure. We propose to achieve this regularization by defining $\mathcal{CT}$ as the difference between the community structure at time t and that at time t-1. Recall that the community structure is captured by $X\Lambda$. Therefore, with $Y= X_{t-1}\Lambda_{t-1}$, the temporal cost is defined as:

$$\mathcal{CT} = D(Y||X\Lambda) \tag{8}$$

where $D(\cdot\|\cdot)$ is the KL-divergence as defined before. The temporal cost $\mathcal{CT}$ is high when there is a dramatic change of community structure from time t-1 to t.

Putting the snapshot cost $\mathcal{CS}$ and the temporal cost $\mathcal{CT}$ together, we have an optimization problem as to find the best community structure at time t, expressed by X and Λ, that minimizes the following total cost:

$$cost = \alpha \cdot D(W||X\Lambda X^T) + (1-\alpha)\cdot D(Y||X\Lambda) \tag{9}$$

subject to $X \in \Re_+^{n\times m}$, $\sum_i x_{ik} = 1$, and Λ being a $m\times m$ non-negative diagonal matrix. Solving this optimization problem is the core of our FacetNet framework.

Solution. We provide an iterative EM algorithm to find the optimal solutions for Eq. (9) as follows:

$$\begin{gathered} x_{ik} \leftarrow x_{ik}\cdot 2\alpha\sum_j \frac{w_{ij}\cdot\lambda_k\cdot x_{jk}}{(X\Lambda X^T)_{ij}} + (1-\alpha)\cdot y_{ik} \\ \text{then normalized such that } \sum_i x_{ik} = 1\ \forall k \\ \lambda_k \leftarrow \lambda_k\cdot\alpha\sum_{ij}\frac{w_{ij}\cdot x_{ik}\cdot x_{jk}}{(X\Lambda X^T)_{ij}} + (1-\alpha)\cdot\sum_i y_{ik} \\ \text{then normalized such that } \sum_k \lambda_k = 1\ . \end{gathered} \tag{10}$$

Details about the proposed model and the convergence of the solution can be found in [29]. Different from the matrix factorization formulation presented here, in [29], the problem is reformulated in terms of maximum a posteriori (MAP) es-

timation and we show a close connection between the optimization framework for solving the evolutionary clustering problem and our proposed generative probabilistic model.

3.2.2 Extracting Communities and Evolutions

Community membership. Assume we have computed the result at time *t*-1, i.e., (X_{t-1},Λ_{t-1}), and the result at time *t*, i.e., (X_t,Λ_t). We define a diagonal matrix D_t, whose diagonal elements $d_{t;ii}=\sum_{ij}(X_t\Lambda_t)_{ij}$. Then the *i*-th row of $D_t^{-1}X_t\Lambda_t$ indicates the "soft" community memberships of v_i at time *t*.

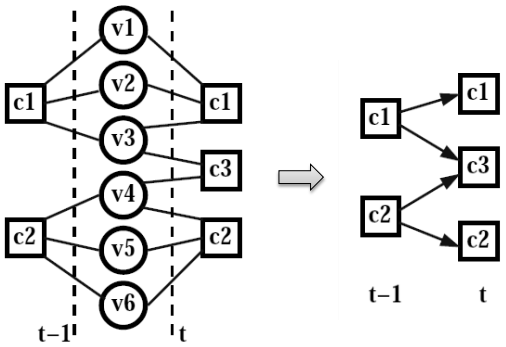


Figure 7: The evolution of communities from time *t*-1 to *t*, is obtained by merging the bipartite graphs through corresponding nodes v_i's – as the probability of starting from $c_{i;t-1}$ (community nodes at *t*-1), walking through the merged bipartite graphs, and reaching $c_{j;t}$ (community nodes at *t*).

Community evolution. To derive the community evolutions, we align the two bipartite graphs, that at time *t*-1 and that at time *t*, side by side by merging the corresponding network nodes v_i's (as illustrated in Figure 7). A natural definition of community evolution (from community $c_{i;t-1}$ at time *t*-1 to community $c_{j;t}$ at time *t*) is the probability of starting from $c_{i;t-1}$, walking through the merged bipartite graphs, and reaching $c_{j;t}$. A simple derivation shows that $P(c_{i;t-1}, c_{j;t}) = (\Lambda_{t-1}X_{t-1}^T D_t^{-1}X_t\Lambda_t)_{ij}$ and $P(c_{j;t}|c_{i;t-1}) = (X_{t-1}^T D_t^{-1}X_t\Lambda_t)_{ij}$. Each node and each edge contribute to the evolution from $c_{i;t-1}$ to $c_{j;t}$. That is, all individuals and all interactions are related to all the community evolutions, with different levels. This is more reasonable compared to traditional methods where the analysis of community evolution assumes all members having identical importance in a community.

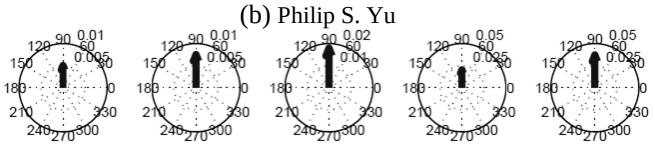
3.3 Application: Time-dependent Ranking in Communities

We apply the FacetNet algorithm on the DBLP co-authorship dataset (see [29] for more details). In Figure 8(a) we list the extracted top authors in two of the extracted communities, Data Mining (DM) and Database (DB), where the rank is determined by the value x_{ik}, i.e., , $p_{k\to i}$. Recall that $p_{k\to i}$ indicates to what level the *k*-

(a) Top members in two communities

Data Mining	**Philip S. Yu**, Jiawei Han, Jian Pei, Wei Wang, Haixun Wang, Beng Chin Ooi, Kian-Lee Tan, Charu C. Aggarwal, Jiong Yang, Hongjun Lu, Mong-Li Lee, Jeffrey Xu Yu, Tok Wang Ling, Anthony K. H. Tung, Dimitris Papadias,Wynne Hsu, Bing Liu, Ke Wang, Yufei Tao, Xifeng Yan, Wei Fan, **Laks V. S. Lakshmanan**, Sourav S. Bhowmick, Guozhu Dong, Jianyong Wang
Database	Divesh Srivastava, Nick Koudas, Divyakant Agrawal, Hans-Peter Kriegel, Surajit Chaudhuri, Amr El Abbadi, H. V. Jagadish, Rajeev Rastogi, Minos N. Garofalakis, S. Muthukrishnan, Jennifer Widom, Rakesh Agrawal, Elke A. Rundensteiner, Jeffrey F. Naughton, Rajeev Motwani, Flip Korn, Michael J. Franklin, Johannes Gehrke, Hector Garcia-Molina, Vivek R. Narasayya, Raghu Ramakrishnan, **Laks V. S. Lakshmanan**, Walid G. Aref, Christos Faloutsos, Sihem Amer-Yahia

(b) Philip S. Yu



(c) Laks V. S. Lakshmanan

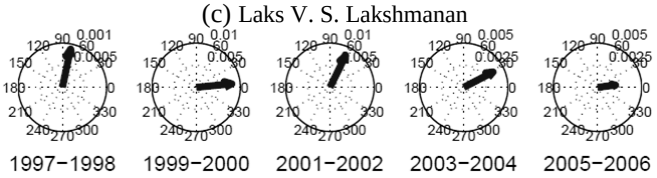


Figure 8: (a) Top members in the Data Mining (DM) and Database (DB) communities, sorted by $p_{k\to i}$. (b,c) The evolution of community memberships for two top authors. In the compasses, an arrow close to the vertical axis indicates a large value of the community membership in the DM community and to the horizontal axis indicates a large value of the community membership in the DB community. Hence, the first author consistently played an important role mainly in the DM community, whereas the second author had a varying role in both communities.

th community involves the *i*-th node, where the value is derived based on both the current and the historic community structures. So from our framework, we can directly infer who the important members in each community are. Note that the importance of a node in a community is determined by its contribution to the community structure.

We can also track the role each individual plays in a community by looking the value of $p_{k\to i}$ over time. In Figure 8(b) and (c), We demonstrate one top author (Philip S. Yu) in the DM community whose community membership remains stable over all the timesteps and another top author (Laks V. S. Lakshmanan) whose community membership varies very much over the 5 timesteps. In the figure, each

compass indicates a pair $(p_{k_1 \to i}, p_{k_2 \to i})$ where k_1 and k_2 correspond to the DB and the DM communities, respectively. So in a compass, a vertical arrow (which has a large projection on the *y*-axis) indicates a large value of the community membership in the DM community and a horizontal arrow (which has a large projection on the *x*-axis) indicates a large value of the community membership in the DB community. We can see that the first author consistently played an important role mainly in the DM community, whereas the second author had a varying role in both the two communities.

Compared to prior link analysis algorithms, such as HITS and PageRank, our FacetNet has two advantages: (a) Localized measures: Unlike most of the ranking algorithms that give global measures, In FacetNet, we obtain individual importance (in terms of his/her participation in each community) and community membership simultaneously. The importance measures are localized (per community) and can be aggregated as global measures on the entire network. (b) Temporal variation: The importance of a node, and the context in which the node is deemed important, may vary over time. A simple function for discounting the historic data is not sufficient to capture different types of variation. FacetNet allows understanding how the nodes' global and local importance change over time.

Summary. The analysis of communities and their evolutions in dynamic temporal networks is a challenging research problem with broad applications. In this work, we proposed a framework, FacetNet, that combines the task of community extraction and the task of evolution extraction in a unified process. To the best of our knowledge, our framework is the first probabilistic generative model that simultaneously analyzes communities and their evolutions. The results obtained from our model allow us to assign soft community memberships to individual nodes, to analyze the strength of ties among various communities, to study how the affiliations of an individual to different communities change over time, as well as to reveal how communities evolve over time. The experimental results on time-dependent ranking in the DBLP communities demonstrate utility of our FacetNet framework. It reveals the community membership evolution for an individual or the evolutions of the communities, and to discover many interesting insights in dynamic networks that are not directly obtainable from existing methods.

We are currently extending this framework in two directions. First, our current model only considered the link information. In many applications, the content information (e.g., the contents of blog entries and the abstracts of papers) is also very important. We are investigating how to incorporate content information into our framework. Second, so far we only use our model to explain the observed data. To extend our model to predict future behaviors of individuals in a dynamic social network is also an important research topic. We shall investigate some of these directions in next section.

4. Community Analysis on Multi-Relational Social Data

This work aims at discovering community structure in rich media social networks, through analysis of the time-varying multi-relational data. As an example scenario, let us consider the use of social media in enterprises, which have increasingly embraced social media software to promote collaboration. Such social media, including wikis, blogs, bookmark sharing, instant messaging, emails, calendar sharing, and so on, foster dynamic collaboration patterns that deviates from the formal organizational structure (e.g. cooperate departments, geographical places, etc.). People who are close in the formal organizational structure (e.g. formal collaboration network) might be far apart in the communication network (e.g. the network of instant messaging). On the other hand, users' document access patterns might be related to their corporate roles as well as personal interests. Figure 9(a) shows an example of such multi-relational social data. The complex and dynamic interplay of various social relations and interactions in an enterprise reflects the day-to-day collaboration practice – how people assemble themselves for a task or activity, how ideas are shared or propagated, through which communication means, who are considered to be expert at some tasks, or what pieces of information are relevant to a particular task, and so on.

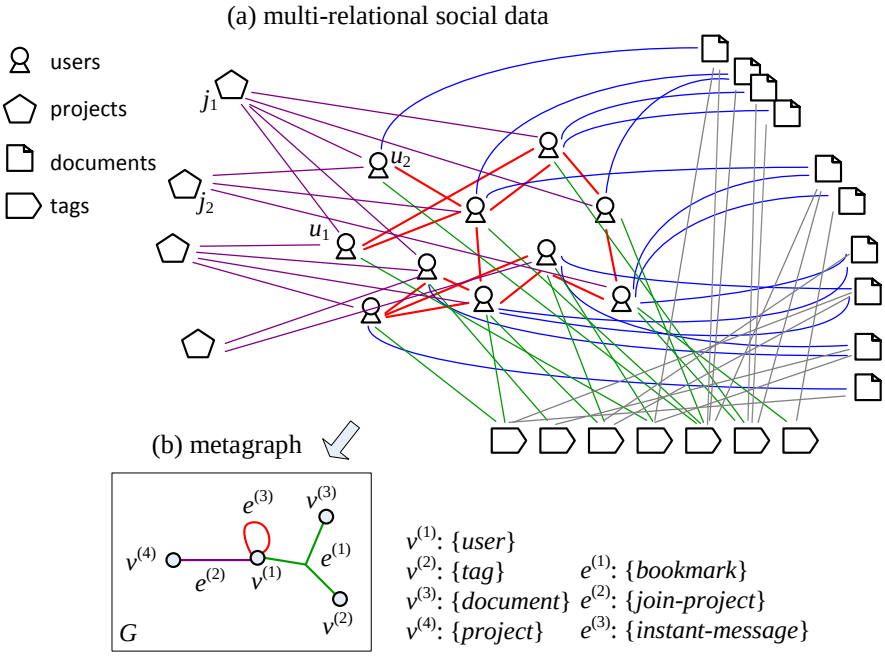


Figure 9: (a) Users and related objects in an enterprise. (c) A metagraph that represents the enterprise social context.

In this section, we introduce the first graph-based tensor factorization algorithm to analyze the dynamics of heterogeneous social networks, which can flexibly discover communities along different dimensions (membership, content, etc.), and can help predict users' potential interests.

4.1 Embeddedness, Artifacts and Community Discovery

Studies have shown that individual behaviors usually result from mechanisms depending on their social networks. Social embeddedness [17] indicates the choices of individuals depend on how they are integrated in dense clusters or multiplex relations of social networks. For example, social embeddedness in cohesive structures can lead people to make similar political contributions [31]. A similar idea has been grounded in situated cognition. According to Dewey, an individual's actions will always be interrelated to all others within certain social medium that forms the individual membership in a community. Once membership is established, the individual begins to share the same supply of knowledge that the group possesses. Accordingly, this shared experience forms an emotional tendency to motivate individual behavior in such a way that it creates purposeful activity evoking certain meaningful outcomes [1].

This work has suggested the behavioral dynamics of individuals occur under complex, social conditions that simultaneously give rise to the community structure (i.e. the "dense cluster" or "community membership"). While the conditions may be ambiguous, situated cognition theorists have suggested that "artifacts holding historic and negotiated significance within a particular context." Based on the idea, we presume the community structure latent in multiplex relations based on shared artifacts affects and is affected by individual choices. That is, community structure that accounts for inherent dependencies between individuals embedded in a multi-relational network can help understand and predict the behavioral dynamics of individuals.

4.2 Extracting Communities from Rich-context Social Networks

We focus on the multi-relational network observed from the social media. We define the problem as discovering latent community structure from the context of user actions represented by multi-relational social networks. The problem has three parts: (1) how to represent multi-relational social data, (2) how to reveal the latent communities consistently across multiple relations, and (3) how to track the communities over time.

4.2.1 Problem Formulation

We formally represent multi-relational social data through *tensor algebra* and *metagraph representation*.

Tensor algebra. A tensor is a mathematical representation of a multi-way array. The order of a tensor is the number of modes (or ways). A first-order tensor is a vector, a second-order tensor is a matrix, and a higher-order tensor has three or more modes. We use **x** to denote a vector, **X** denote a matrix, and $\boldsymbol{\mathcal{X}}$ a tensor. Each entry (*i*,*j*,*k*) in a tensor, for example, could represent the number of times the user *i* submitted an entry on topic *j* with keyword *k*.

Metagraph representation. We introduce *metagraph*, a relational hypergraph for representing multi-relational and multi-dimensional social data. We use a metagraph to configure the relational context specific to the system features – this is the key to make our community analysis adaptable to various social media contexts, e.g. an enterprise or a social media website like Digg. We shall an enterprise example to illustrate three concepts: *facet*, *relation*, and *relational hypergraph*.
As shown in Figure 9(a), assume we observe a set of users in an enterprise. These users might collaborate under different working projects, e.g. the user u_1 and u_2 work for the project j_1, and user u_2 belong to two projects j_1 and j_2 at the same time. Collaboration can occur implicitly across different social media such as instant messenger or email, e.g. user u_3 frequently IM with u_1 and u_2. We denote a set of objects or entities of the same type as a *facet*, e.g. a user facet is a set of users, a project facet is a set of projects. We denote the interactions among facets as a *relation*; a relation can involve two (i.e. binary relation) or more facets, e.g. the "join-project" relation involves two facets (user, project), and the "bookmark" relation involves three facets (user, document, tag). A facet can be implicit, depending on whether the facet entities interact with other facets, e.g. the set of bookmark object might be omitted due to no interaction with other facets. Formally, we denote the *q*-th facet as $v^{(q)}$ and the set of all facets as *V*. A set of instantiations of an *M*-way relation e on facets $v^{(1)}$, $v^{(2)}$,…, $v^{(M)}$ is a subset of the Cartesian product $v^{(1)}\times\ldots\times v^{(M)}$. We denote a particular relation by $e^{(r)}$ where *r* is the relation index. The observations of an *M*-way relation $e^{(r)}$ are represented as an *M*-way data tensor $\boldsymbol{\mathcal{X}}^{(r)}$.

Now we introduce a *multi-relational hypergraph* (denoted as *metagraph*) to describe the combination of relations and facets in a social media context (ref. Figure 9(b)). A hypergraph is a graph where edges, called *hyperedges*, connect to any number of vertices. The idea is to use an *M*-way hyperedge to represent the interactions of *M* facets: each facet as a vertex and each relation as a hyperedge on a hypergraph. A metagraph defines a particular structure of interactions among facets, not among facet elements. Formally, for a set of facets $V=\{v^{(q)}\}$ and a set of relations $E=\{e^{(r)}\}$, we construct a metagraph $G=(V,E)$. To reduce notational complexity, *V* and *E* also represent the set of all vertex and edge indices, respectively. A hyperedge/relation $e^{(r)}$ is said to be incident to a facet/vertex $v^{(q)}$ if $v^{(q)}\in e^{(r)}$, which is represented by $v^{(q)}\sim e^{(r)}$ or $e^{(r)}\sim v^{(q)}$. E.g., in Figure 9(b), the vertex $v^{(1)}$

represents the user facet, the hyperedge $e^{(1)}=\{v^{(1)},v^{(2)},v^{(3)}\}$ represents the "book-mark" relation.

Based on the discussed in Section 4.1, we assume the interaction between any two entities (users or media objects) i and j in a community k, written as x_{ij}, can be viewed as a function of the relationships between community k with entity i, and k with j. If we consider the function to be stochastic, i.e. let $p_{k\to i}$ indicate how likely an interaction in the k-th community involves the i-th entity and p_k is the probability of an interaction in the k-th community, we can express x_{ij} by $x_{ij}\approx\sum_k p_{k\to i}\cdot p_{k\to j}\cdot p_k$ (as discussed in section 3.2). Likewise a 3-way interaction among entity i_1, i_2 and i_3 is $x_{i_1 i_2 i_3}\approx\sum_k p_k\cdot p_{k\to i_1}\cdot p_{k\to i_2}\cdot p_{k\to i_3}$. A set of such interactions among entities in facet $v^{(1)}$, $v^{(2)}$ and $v^{(3)}$ can be written by:

$$\boldsymbol{\mathcal{X}}\approx\sum_{k=1}^{K} p_k\,\mathbf{u}_k^{(1)}\circ\mathbf{u}_k^{(2)}\circ\mathbf{u}_k^{(3)}=[\mathbf{z}]\prod_{m=1}^{3}\times_m\mathbf{U}^{(m)} \qquad (11)$$

where $\boldsymbol{\mathcal{X}}\in\Re_+^{I_1\times I_2\times I_3}$, is the data tensor representing the observed three-way interactions among facet $v^{(1)}$, $v^{(2)}$ and $v^{(3)}$. $p_{k\to i_q}$ is written as an (i_q,k)-element of $\mathbf{U}^{(q)}$ for q=1,2,3. $\mathbf{U}^{(q)}$ is an $I_q\times K$ matrix, where I_q is the size of $v^{(q)}$. The probabilities of communities are elements of $\mathbf{z}$, i.e. $p_k=\mathbf{z}_k$. We use $[\mathbf{z}]$ to denote a superdiagonal tensor, where the operation $[\cdot]$ transforms a vector $\mathbf{z}$ to a superdiagonal tensor by setting tensor element $z_{k\ldots k}=\mathbf{z}_k$ and other elements as 0. The decomposition defined in eq.(11) is similar to the CP/PARAFAC tensor decomposition [7,18], except that the *core tensor* $[\mathbf{z}]$ and the *factor matrices* $\{\mathbf{U}^{(q)}\}$ are constrained to contain non-negative probability values. Under the nonnegative constraints, the 3-way tensor factorization is equivalent to the three-way aspect model in a three-dimensional co-occurrence data [33].

The nonnegative tensor decomposition can be viewed as community discovery in a single relation. The interactions in social media networks are more complex – usually involving multiple two- or multi-way relations. By using metagraphs, we represent a diverse set of relational context in the same form and define community discovery problem on a metagraph, with the following two technical issues: (a) how to extract community structure as coherent interaction latent spaces from observed social data defined on a metagraph, and (b) how to extract community structure as coherent interaction latent spaces from time evolving data given a metagraph. The problems are formally stated as follows.

Definition 3 (*Metagraph Factorization*, or *MF*): given a metagraph $G=(V,E)$ and a set of observed data tensors $\{\boldsymbol{\mathcal{X}}^{(r)}\}_{r\in E}$ defined on G, find a nonnegative core tensor $[\mathbf{z}]$ and factors $\{\mathbf{U}^{(q)}\}_{q\in V}$ for corresponding facets $V=\{v^{(q)}\}$. (Since E also represents the set of all edge indices, the notations $r\in E$ and $e^{(r)}\in E$ are exchangeable. Likewise, $q\in V$ and $v^{(q)}\in V$ are exchangeable.)

Definition 4 (*Metagraph Factorization for Time evolving data*, or *MFT*): given a metagraph $G=(V,E)$ and a sequential set of observed data tensors $\{\mathcal{X}_t^{(r)}\}_{r\in E}$ defined on G for time $t=1,2,\ldots$, find nonnegative core tensor $[\mathbf{z}_t]$ and factors $\{\mathbf{U}_t^{(q)}\}_{q\in V}$ for corresponding facets $V=\{v^{(q)}\}$.

We will present our method in two steps: (1) present a solution to MF (next section); (2) extend the solution to solve MFT (Section 4.2.3).

4.2.2 Metagraph Factorization

The MF problem can be stated in terms of optimization. Let us first consider a simple metagraph case. Assume we are given a metagraph $G=(V,E)$ with three vertices $V=\{v^{(1)}, v^{(2)}, v^{(3)}\}$ and two 2-way hyperedges $E=\{e^{(a)},e^{(b)}\}$ that describe the interactions among these three facets, as shown in Figure 10. The observed data corresponding to the hyperedges are two second-order data tensors (i.e. matrices) $\{\mathcal{X}^{(a)}, \mathcal{X}^{(b)}\}$ with facets $\{v^{(1)}, v^{(2)}\}$ and $\{v^{(2)}, v^{(3)}\}$ respectively. The facet $v^{(2)}$ is shared by both tensors.

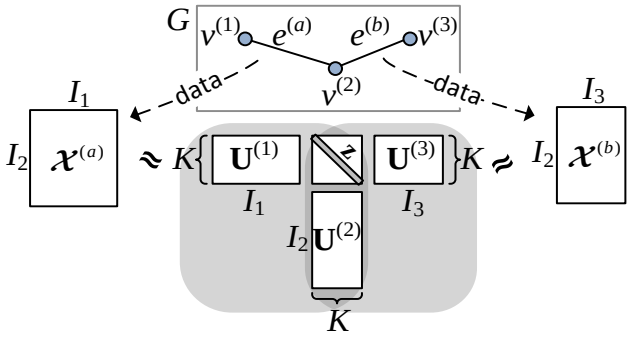


Figure 10: An example of the metagraph factorization (MF). Given observed data tensors $\{\mathcal{X}^{(a)},\mathcal{X}^{(b)}\}$ and a metagraph G that describes the interaction among facets $\{v^{(1)}, v^{(2)}, v^{(3)}\}$, find consistent community structure expressed by core tensor $[\mathbf{z}]$ and facet factors $\{\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}\}$.

The goal is to extract community structure from data tensors, through finding a nonnegative core tensor $[\mathbf{z}]$ and factors $\{\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}\}$ corresponding to the three facets. The core tensor and factors need to consistently explain the data, i.e. we can approximately express the data by $\mathcal{X}^{(a)}\approx[\mathbf{z}]\times_1\mathbf{U}^{(1)}\times_2\mathbf{U}^{(2)}$ and $\mathcal{X}^{(b)}\approx[\mathbf{z}]\times_2\mathbf{U}^{(2)}\times_3\mathbf{U}^{(3)}$, as illustrated in Figure 10. The core tensor $[\mathbf{z}]$ and facet $\mathbf{U}^{(2)}$ are shared by the two approximations, and the length of $\mathbf{z}$ is determined by the number of latent spaces (communities) to be extracted. Since both the left- and the right-hand side of the approximation are probability distributions, it is natural to use the KL-divergence (denoted as $D(\cdot\|\cdot)$) as a measure of approximation cost.

We can generalize Figure 10 to any metagraph G, as: given a metagraph $G=(V,E)$, the objective is to factorize all data tensors such that all tensors can be approximated by a common nonnegative core tensor $[\mathbf{z}]$ and a shared set of nonnegative factors $\{\mathbf{U}^{(q)}\}$, i.e. to minimize the following cost function:

$$\begin{aligned} &J(G) = \min_{\mathbf{z},\{\mathbf{U}^{(q)}\}} \sum_{r\in E} D(\boldsymbol{\mathcal{X}}^{(r)} || [\mathbf{z}] \prod_{m:v^{(m)}\sim e^{(r)}} \times_m \mathbf{U}^{(m)}) \\ &\text{s.t. } \mathbf{z} \in \Re_+^{1\times K}, \mathbf{U}^{(q)} \in \Re_+^{I_q \times K} \ \forall q, \textstyle\sum_i \mathbf{U}_{ik}^{(q)} = 1 \ \forall q \forall k \end{aligned} \tag{12}$$

where K is the number of communities, and $D(\cdot\|\cdot)$ is the KL-divergence as described above. The constraint that each column of $\{\mathbf{U}^{(q)}\}$ must sum to one is added due to the modeling assumption that the probability of an occurrence of a relation on an entity is independent of other entities in a community. This equation can be easily extended to incorporate weights on relations.

Solution. By employing the concavity of the log function (in the KL-divergence), we derive a local minima solution to Eq. (12). The solution can be found by the following updating algorithm:

$$\begin{aligned} &\mathbf{z}_k \leftarrow \textstyle\sum_{r\in E} \sum_{i_1\cdots i_{M_r}} \boldsymbol{\mathcal{X}}^{(r)}_{i_1\cdots i_{M_r}} \mu^{(r)}_{i_1\cdots i_{M_r} k} \\ &U^{(q)}_{i_q k} \leftarrow \textstyle\sum_{l:e^{(l)}\sim v^{(q)}} \sum_{i_1\cdots i_{q-1} i_{q+1}\cdots i_{M_l}} \boldsymbol{\mathcal{X}}^{(l)}_{i_1\cdots i_{M_l}} \mu^{(l)}_{i_1\cdots i_{M_l} k} \\ &\text{where } \mu^{(r)}_{i_1\cdots i_{M_r} k} \leftarrow \frac{\mathbf{z}_k \prod_{m:v^{(m)}\sim e^{(r)}} \times_m \mathbf{U}^{(m)}_{i_m k}}{([\mathbf{z}] \prod_{m:v^{(m)}\sim e^{(r)}} \times_m \mathbf{U}^{(m)})_{i_1\cdots i_{M_r}}} \end{aligned} \tag{13}$$

where $\mathbf{z}$ is a length K vector, $L=|E|$ denotes the total number of hyperedges on G. After updates, each column of $\mathbf{U}^{(q)}$ and the vector $\mathbf{z}$ are normalized to sum to one. Because of this normalization step, we have omitted the scaling constant for updating $\mathbf{z}$ and $\mathbf{U}^{(q)}$. This iterative update algorithm is a generalization of the algorithm proposed by Lee et al. [24] for solving the single nonnegative matrix factorization problem. In metagraph factorization, the update for core tensor $[\mathbf{z}]$ depends on all hyperedges on the metagraph, and the update for each facet factor $\mathbf{U}^{(q)}$ depends on the hyperedges incident to the facet. The details of this algorithm can be found in [30].

4.2.3 Time Evolving Extension

In the MFT problem, the relational data is constantly changing as evolving tensor sequences. We propose an online version of MF to handle dynamic data. Since historic information is contained in the community model extracted based on previously observed data, the new community structure to be extracted should be consistent with previous community model and new observations, which is similar to evolutionary clustering discussed in Section 3. To achieve this, we extend the objective in Eq. (12) as follows.

A community model for a particular time t is defined uniquely by the factors $\{\mathbf{U}_t^{(q)}\}$ and core tensor $[\mathbf{z}_t]$. (To avoid notation clutter, we omit the time indices for t.) For each time t, the objective is to factorize the observed data into the nonnega-

tive factors $\{\mathbf{U}^{(q)}\}$ and core tensor $[\mathbf{z}]$ which are close to the prior community model, $[\mathbf{z}_{t\text{-}1}]$ and $\{\mathbf{U}_{t\text{-}1}{}^{(q)}\}$. We introduce a cost l_{prior} to indicate how the new community structure deviates from the previous structure in terms of the KL-divergence. The new objective is defined as follows:

$$\begin{aligned} &J_2(G) = \min_{\mathbf{z},\{\mathbf{U}^{(q)}\}}(1-\alpha)\sum_{r\in E} D(\boldsymbol{\mathcal{X}}^{(r)}||[\mathbf{z}]\prod_{m:v^{(m)}\sim e^{(r)}}\times_m \mathbf{U}^{(m)}) + \\ &\alpha l_{prior} \\ &\text{with } l_{prior} = D(\mathbf{z}_{t-1}||\mathbf{z}) + \sum_q D(\mathbf{U}_{t-1}^{(q)}||\mathbf{U}^{(q)}) \\ &\text{s.t. } \mathbf{z} \in \Re_+^{1\times K}, \mathbf{U}^{(q)} \in \Re_+^{I_q\times K}\ \forall q, \sum_i \mathbf{U}_{ik}^{(q)} = 1\ \forall q \forall k \end{aligned} \tag{14}$$

where α is a real positive number between 0 and 1 to specify how much the prior community model contributes to the new community structure. l_{prior} is a regularizer used to find similar pair of core tensors and pairs of facet factors for consecutive time. The new community structure will be a solution incrementally updated based on a prior community model.

Solution. Based on a derivation similar to the discussion in Section 4.2.2, we provide a solution to Eq. (14) as follows:

$$\begin{aligned} &\mathbf{z}_k \leftarrow (1-\alpha)\sum_{r\in E}\sum_{i_1\cdots i_{M_r}} \boldsymbol{\mathcal{X}}^{(r)}_{i_1\cdots i_{M_r}}\mu^{(r)}_{i_1\cdots i_{M_r}k} + \alpha\mathbf{z}_{k;t-1} \\ &U^{(q)}_{i_q k} \leftarrow (1-\alpha)\sum_{l:e^{(l)}\sim v^{(q)}}\sum_{i_1\cdots i_{q-1}i_{q+1}\cdots i_{M_l}} \boldsymbol{\mathcal{X}}^{(l)}_{i_1\cdots i_{M_l}}\mu^{(l)}_{i_1\cdots i_{M_l}k} + \alpha U^{(q)}_{i_q k;t-1} \end{aligned} \tag{15}$$

where $\mathbf{z}$ is a length K vector, $\mu^{(l)}_{i_1\cdots i_{M_l}k}$ is defined as in eq. (13). After updates, each column of $\mathbf{U}^{(q)}$ and the vector $\mathbf{z}$ are normalized to sum to one. Because of this normalization step, we have dropped the scaling constant for updating $\mathbf{z}$ and $\mathbf{U}^{(q)}$. It can be shown that the parameters in the previous model ($\mathbf{z}_{t\text{-}1}$ and $\{\mathbf{U}_{t\text{-}1}{}^{(q)}\}$) act as Dirichlet prior distribution to inform the solution search (ref. [28]), thus the solution is consistent with previous community structure.

4.3 Application: Context-sensitive Prediction in Enterprise

We design a prediction task to illustrate how our community tracking algorithm can be utilized to predict users' future interests based on the multi-relational social data. Specifically, given data D_t at time t, we extract communities to predict users' future use of tags, and compare the prediction with the ground truth in data D_{t+1}. We collected collaboration relationships from the employee profiles and social media (e.g. bookmarks, wiki, etc.) in an enterprise. We then construct multiple relations from the different data sources.

In our experiment, the time interval is one month. The overall prediction performance is obtained by taking average prediction performance over 10-month data. We compare our method with two baseline methods: (1) *recurring interests* – predicting future tags (at t+1) as the tags mostly frequently used by the user at t. (2) *collective interests* (pLSA) – predicting future tags by using a well-known collec-

tive filtering method (probabilistic latent semantic analysis [19] or pLSA) on the user-tag matrix.

We generate predictions base on the community structure extracted by our method, denoted by MF and MFT. The MF algorithm outputs community structure from relational data of each time slot t-1. The MFT algorithm uses the same data as MF, with an aid of prior community model extracted for time t-2 as an informative prior. Hence MFT gives results incrementally. From an extracted community model we obtain the probability of a community k, i.e. $p(k)$, and the probability of a user u and a tag q, given community k, i.e. $p(u|k)$ and $p(q|k)$. Then a prediction is made based on the condition probability $p(q|u) \propto p(u,q) \approx \Sigma_{k'} p(k)p(u|k')p(q|k')$. The detailed experiment setting can be found in [30]. Our method can also be applied to a *cold-start* setting by incorporating a folding-in technique (ref. e.g. [33]) to overcome the situation where $p(q|k)$ may not be directly available from the model parameters (e.g. q is a new tag which has not been used before t).

Figure 11 shows the relevant improvement compared with the first baseline method, i.e. the recurring interests. The results indicate the prediction given by our community tracking algorithms outperform the baseline methods by 36-250% on the average, which suggest that our method can better capture the cohesive structures of the contexts around users' interests.

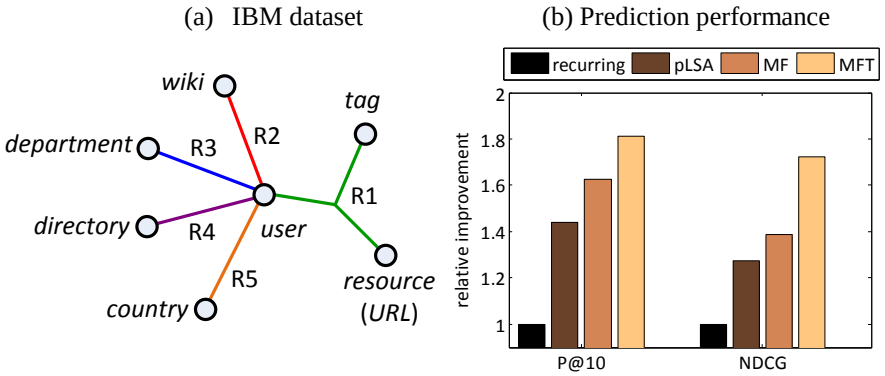


Figure 11: (a) IBM dataset: R1, ..., R5 are different relations among the 7 facets, e.g. bookmark (R1), join-wiki (R2), etc. The sizes of these relational data from R1~R5 are: 3K×12K×61K, 3K×1K, 3K×2K, 3K×90 and 3K×42. (b) Prediction performance: Our framework improves the prediction of users' future tag use.

Summary. We proposed the MetaFac framework to extract community structures from various social contexts and interactions. There were three key ideas: (1) me-

tagraph, a relational hypergraph for representing multi-relational social data; (2) MF algorithm, an efficient non-negative multi-tensor factorization method for community extraction on a given metagraph; (3) MFT, an on-line factorization method to handle time-varying relations. To illustrate the utility of our method, we design a tag prediction task in an enterprise context. We generated the predictions based on the extracted community models and compare results with baselines. Our method outperformed baselines up to an order of magnitude. We show significant improvement of our method due to (a) incorporating a historic model and (b) leveraging diverse relations through a metagraph.

There are some open issues in this work. For example, there are different aspects of community evolution, including change in the community size, change in the number of communities and change in the community content or features (what the community is about). To study the evolution within communities, our method has assumed the number of communities does not change across time (i.e. we do not consider the second aspect). Learning and comprehending several evolution aspects in a unified process is a challenging issue.

Nevertheless, our work can lead to several interesting directions. (1) As our algorithm does not tie to a specific data schema, it can be easily extended to deal with schema changes. (2) By combining various social relations of data, it can be used to identify effective social relations based on model selection approaches. As a potential extension of this framework, we are interested in utilizing the relational hypergraph to study the correlation between networks and the behavioral dynamics of individuals.

5. Conclusions and Future Directions

In this article, we have discussed our current work on community analysis in dynamic, multi-relational social networks. Our work includes several key ideas: (1) We introduce mutual awareness, a fundamental property of communities in online social media, which is computationally defined based on observable individual actions within the social media context. The effectiveness of mutual awareness features is empirically verified on large-scale real-world blog datasets. We propose an efficient iterative mutual awareness expansion algorithm for community extraction using a random walk based distance measure that quantifying the amount of mutual awareness expanding on the entire network. The community evolution with respect to a query reveals interesting community dynamics. (2) We propose FacetNet framework, the first probabilistic generative model that simultaneously analyzes communities and their evolutions. The results obtained from our method allow us to assign soft community memberships to individual nodes, to analyze the strength of ties among various communities, to study how the affiliations of an individual to different communities change over time, as well as to reveal how communities evolve over time. Extensive experimental studies demonstrated that

by using our FacetNet framework, we are able to discover many interesting insights in dynamic networks that are not directly obtainable from existing methods, such as the evolution of individuals' contribution to different communities. (3) We propose MetaFac, the first graph-based tensor factorization framework for analyzing the dynamics of heterogeneous social networks. We introduce metagraph for modeling multi-relational and multi-dimensional social data. Then we propose an efficient non-negative multi-tensor factorization method for community extraction on a given metagraph. In addition, we provide an on-line extension of this method to handle time-varying multi-relations. Extensive experiments on enterprise and large-scale social media data suggest that our technique is scalable and can help predict users' future interests based on the cohesive structure of contexts extracted by our method.

Our current work has led to several interesting research directions. In social media like Facebook, users often experience overload of online social connections, shared interests and information, as well as the interplay of people and subject matter. For people who are interested in certain topics, it is difficult to understand how the topics (or media items) are, and have been, shared and discussed by different people. This leads to important technical questions on how to disentangle and display the complex dynamic relationships between people and subjects over time. Our methods can be used to support interactive visualization that allows users to explore and query multiple aspects of community activities, such as relevant topics, representative users and artifacts (e.g. tweets, photos) of the communities, as well as their relationships and evolutions.

Our work can also contribute to research in social science. For example, anthropologists are interested in material cultural transition based on artifacts and their association with time and space. Our work on community analysis has identified several relevant structural elements from large scale observable interpersonal activities in social media, including community awareness (mutual and transitive awareness), community composition (degree of individuals' participation in communities), community inter-structure (the relationship among communities and how it changes over time), community context (e.g. time and location associated with community activities) and community artifacts (e.g. tags and photos generated by communities). We have used these structural elements to discover interesting cultural patterns among communities, e.g. the interaction of right-wing and left-wing communities in blog data. We plan to extend the current approach to discover the structural changes of a community as well as the context where the changes occur, to support the search and detection of emergent or transitional cultural patterns. Ethnographic investigation is needed in order to reveal finer-grained patterns of human interactions as well as qualitative understanding of the extracted structures. Extracting the evolution of inter-structure among different communities

will help understand the condition and impact of social media as a new communicative practice.

6. Acknowledgement

This material is based upon work supported in part by NEC Labs America, an IBM Ph.D. Fellowship and a Kauffman Entrepreneur Scholarship. We are pleased to acknowledge Yun Chi, Shenghuo Zhu, Belle Tseng, Jun Tatemura and Koji Hino, from NEC Labs America, for providing the invaluable advices on community discovery and the NEC Blog dataset. We are indebted to Jimeng Sun, Paul Castro and Ravi Konuru, from IBM T.J. Watson Research Center, for providing advices on tensor analysis and the IBM enterprise data.

References

[1] A. AGOSTINO (1999). *The relevance of media as artifact: Technology situated in context.* Educational Technology & Society **2**(4): 46-52.

[2] S. ASUR, S. PARTHASARATHY, et al. (2007). *An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs*, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining,

[3] L. BACKSTROM, D. HUTTENLOCHER, et al. (2006). *Group formation in large social networks: membership, growth, and evolution*, SIGKDD, 44-54, 2006.

[4] L. BACKSTROM, R. KUMAR, et al. (2008). *Preferential behavior in online groups.* Proceedings of the international conference on Web search and web data mining: 117-128.

[5] S. BRIN and L. PAGE (1998). *The anatomy of a large-scale hypertextual Web search engine.* Computer Networks and ISDN Systems **30**(1-7): 107-117.

[6] J. BROWN, A. COLLINS, et al. (1989). *Situated cognition and the culture of learning.* Educational researcher **18**(1): 32.

[7] J. CARROLL and J. CHANG (1970). *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition.* Psychometrika **35**(3): 283-319.

[8] D. CHAKRABARTI, R. KUMAR, et al. (2006). *Evolutionary clustering*, SIGKDD, 554-560,

[9] Y. CHI, X. SONG, et al. (2007). *Evolutionary Spectral Clustering by Incorporating Temporal Smoothness*, SIGKDD, 2007.

[10] F. CHUNG and S. YAU (2000). *Discrete Green's functions.* J. of Combinatorial Theory (A) **91**(1-2): 191-214.

[11] I. DHILLON, Y. GUAN, et al. (2005). *A unified view of kernel k-means, spectral clustering and graph partitioning*. Technical report, University of Texas at Austin, 2005.

[12] P. DOURISH (2001). Where the Action Is:: the Foundations of Embodied Interaction, Mit Pr.

[13] N. FRIEDKIN and E. JOHNSEN (1999). *Social influence networks and opinion change.* Advances in Group Processes **16**: 1-29.

[14] H. GARFINKEL (1984). Studies in ethnomethodology, Polity.

[15] D. GARLASCHELLI and M. LOFFREDO (2004). *Patterns of link reciprocity in directed networks.* Physical Review Letters **93**(26): 268701.

[16] M. GRANOVETTER (1973). *The strength of weak ties.* American journal of sociology **78**(6): 1360.

[17] M. GRANOVETTER (1985). *Economic action and social structure: A theory of embeddedness.* American journal of sociology **91**(3): 481-510.

[18] R. HARSHMAN (1970). *Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis.* UCLA working papers in phonetics **16**(1): 84.

[19] T. HOFMANN (1999). *Probabilistic latent semantic indexing*, SIGIR, 1999.

[20] Q. JONES (1997). *Virtual-communities, virtual settlements & cyber-archaeology: A theoretical outline.* Journal of Computer Mediated Communication **3**(3): 35-49.

[21] R. KANNAN, S. VEMPALA, et al. (2004). *On Clusterings: Good, Bad and Spectral.* J. of the ACM **51**(3): 497-515.

[22] J. M. KLEINBERG (1999). *Authoritative sources in a hyperlinked environment.* J. ACM **46**(5): 604-632.

[23] T. KOLDA and J. SUN (2008). *Scalable Tensor Decompositions for Multi-aspect Data Mining*, ICDM, 2008.

[24] D. LEE and H. SEUNG (2001). *Algorithms for non-negative matrix factorization*, NIPS, 556–562, 2001.

[25] J. LEMKE (1997). *Cognition, context, and learning: A social semiotic perspective.* Situated cognition: Social, semiotic, and psychological perspectives: 37–56.

[26] Y.-R. LIN, H. SUNDARAM, et al. (2006). *Discovery of Blog Communities based on Mutual Awareness*, the 3rd Annual Workshop on the Weblogging Ecosystems: Aggregation, Analysis and Dynamics,

[27] Y.-R. LIN, H. SUNDARAM, et al. (2007). *Blog Community Discovery and Evolution Based on Mutual Awareness Expansion*, 2007 IEEE/WIC/ACM International Conference on Web Intelligence, 2007.

[28] Y.-R. LIN, Y. CHI, et al. (2008). *FaceNet: A Framework for Analyzing Communities and Their Evolutions in Dynamics Networks*, WWW, 2008.

[29] Y.-R. LIN, Y. CHI, et al. (2009). *Analyzing Communities and Their Evolutions in Dynamics Networks.* Transactions on Knowledge Discovery from Data (TKDD) **3**(2).

[30] Y.-R. LIN, J. SUN, et al. (2009). *MetaFac: Community Discovery via Relational Hypergraph Factorization*, SIGKDD, 2009.

[31] J. MOODY and D. WHITE (2003). *Structural cohesion and embeddedness: A hierarchical concept of social groups.* American Sociological Review: 103-127.

[32] G. PALLA, A. BARABASI, et al. (2007). *Quantifying social group evolution.* eprint arXiv: 0704.0744.

[33] A. POPESCUL, L. H. UNGAR, et al. (2001). *Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments*, UAI 2001, 437-444,

[34] H. RHEINGOLD (1899). Virtual community: Homesteading on the electronic frontier, Mit Press.
[35] M. SAHAMI and T. HEILMAN (2006). *A web-based kernel function for measuring the similarity of short text snippets*, Proc. of the 15th international conference on World Wide Web, 377-386, 2006.
[36] J. SHI and J. MALIK (2000). *Normalized Cuts and Image Segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8): 888-905.
[37] J. TRAVERS and S. MILGRAM (1969). *An Experimental Study of the Small World Problem.* Sociometry **32**(4): 425-443.
[38] S. WASSERMAN and K. FAUST (1994). Social Network Analysis: methods and applications, Cambridge University Press.
[39] K. YU, S. YU, et al. (2005). *Soft Clustering on Graphs.* NIPS'05.

Index terms (alphabetically):

Community discovery
Community dynamics
Dynamic networks
FacetNet
Graph mining
MetaFac
Multi-relational mining
Mutual awareness
Social media
Social network analysis