Understanding Group Structures and Properties in Social Media

Lei Tang and Huan Liu

Computer Science and Engineering
Arizona State University
Tempe, AZ 85287-8809, USA
{L.Tang,Huan.Liu}@asu.edu

Abstract. The rapid growth of social networking sites enables people to connect to each other more conveniently than ever. With easy-to-use social media, people contribute and consume contents, leading to a new form of human interaction and the emergence of online collective behavior. In this chapter, we aim to understand group structures and properties by extracting and profiling communities in social media. We present some challenges of community detection in social media. A prominent one is that networks in social media are often heterogeneous. We introduce two types of heterogeneity presented in online social networks and elaborate corresponding community detection approaches for each type, respectively. Social media provides not only interaction information but also textual and tag data. This variety of data can be exploited to profile individual groups in understanding group formation and relationships. We also suggest some future work in understanding group structures and properties.

Key words: social media, community detection, group profiling, heterogeneous networks, multi-mode networks, multi-dimensional networks

1 Introduction

Social media such as Facebook, MySpace, Twitter, and BlogSpot facilities people of all walks of life to express their thoughts, voice their opinions, and connect to each other anytime and anywhere. For instance, popular content-sharing sites like Del.icio.us, Flickr, and YouTube allow users to upload, tag and comment different types of contents (bookmarks, photos, videos). Users registered at these sites can also become friends, a fan or a follower of others. Social media offers rich information of human interaction and collective behavior in a much larger scale (hundreds of thousands or millions of actors). It is gaining increasing attention across various disciplines including sociology, behavior science, anthropology, computer science, epidemics, economics, marketing business, to name a few.

With the expanded use of web and social media, virtual communities and online interactions have become a vital part of human experience. Members of virtual communities tend to share similar interests or topics, and connect to each

other in a community more frequently than with those outside the community. For example, there can be two groups browsing news at a website, say *digg.com*: one is interested in topics related to *Meteorology*, while the other in *Politics*; A blogger (say the owner of *http://hunch.net/*) who publishes blog posts actively on "machine learning" often has links on his/her blog site to other bloggers who concentrate on "machine learning" as well. It would be interesting to find these like-minded individuals for developing many other applications to enhance personal experience or to improve business intelligence. In this work, we focus on *communities* (or equivalently *groups*) in social media. There is a wide range of applications of discovering groups (a.k.a. *community detection*) based on the interactions among actors and capturing group properties via shared topics, including visualization [8], recommendation and classification [18, 19], influence study [1], direct marketing, group tracking and recommendation.

Community detection is a classical task in social network analysis. However, some new features presented in networks of social media entail novel solutions to handle online communities.

- *Heterogeneity.* Networks in social media tend to involve multiple types of entities or interactions. For instance, in content sharing sites like Flickr and YouTube, multiple types of entities: users, tags, comments and contents are intertwined with each other. Sometimes, users at the same social network site can interact with each other in various forms, leading to heterogeneous types of interactions between them. It is intriguing to explore whether or not heterogeneous information can help identify communities. It is also challenging to effectively fuse these heterogeneous types of information.
- *Large-Scale Networks.* Networks in social media are typically in a much larger scale than those in traditional social network analysis. Traditional social network analysis relies on circulation of questionnaires or surveys to collect interaction information of human subjects, limiting the scale of analysis to hundreds of actors mostly. Hence, scalability is seldom a focus there. Networks in social media, on the contrary, involve a much larger number of actors, which presents a challenge of scalability. In addition, large-scale networks yield similar patterns, such as power-law distribution for node degrees and small-world effect [3]. It is yet unclear how these patterns can help or guide data mining tasks.
- *Collective Intelligence.* In social media, crowd wisdom, in forms of tags and comments, is often available. Is it possible to employ collective intelligence to help understand group structures and properties? For instance, how to characterize a group? How to differentiate a group from others in social media? What are potential causes that lead some users to form a community? With abounding groups in social media, how can we understand the relationship among them?
- *Evolution.* Each day in social media, new users join the network and new connections occur between existing members, while some existing ones leave or become dormant. How can we capture the dynamics of individuals in networks? Can we find the members that act like the backbone of commu-

nities? The group interests might change as well. How can we update the group interests and relations accordingly as information evolves?

Given the features above, we will mainly discuss two research issues concerning communities in social media: (1) Identifying communities in social media via the readily-available interaction information; and (2) Profiling groups dynamically using descriptive tags and taxonomy adaptation. The two research tasks are highly related to each other. The first task identifies groups, serving as the basis for the second one; and the second task helps understand the formation of identified groups and unravel properties why users join together to form a group. In the following section, we first introduce heterogeneous networks in social media, define the problems of interest and motivations. We will then elucidate the technical details with challenges and solutions for both tasks in the subsequent sections.

2 Heterogeneous Networks in Social Media

There are two types of heterogeneous networks that demand special attention. We first illustrate the two types and then expound the necessity for considering heterogeneity in community detection.

2.1 Heterogeneous Networks

With social media, people can connect to each other more conveniently than ever. In some social networking sites, entities other than human beings can also be involved. For instance, in YouTube, a user can upload a video and another user can tag it. In other words, the users, videos, and tags are weaved into the same network. The "actors" in the network are not at all homogeneous. Furthermore, examining activities of users, we can observe different interaction networks between the same set of actors. Take YouTube again as an example. A user can become a friend of another user's; he can also subscribe to another user. The existence of different relations suggests that the interactions between actors are heterogeneous. Networks involving heterogeneous actors or interactions are referred as *heterogeneous networks*. Accordingly, heterogeneous networks can be categorized in two different types:

– *Multi-Mode Networks* [22]. A multi-mode network involves heterogeneous actors. Each mode represents one type of entity. For instance, in the YouTube example above, a 3-mode network can be constructed, with videos, tags and users each representing a mode, as seen in Figure 1. There are disparate interactions among the three types of entities: users can upload videos. They can also provide tags for some videos. Intuitively, two users contributing similar videos or tags are likely to share interests. Videos sharing similar tags or users are more likely to be related. Note that in the network, both tags, and videos are also considered as "actors", though users are probably the major mode under consideration.

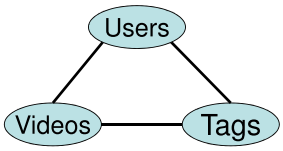

Fig. 1. A multi-mode network in YouTube

Other domains involving networks or interactions also encounter multi-mode networks. An example of multi-mode network is academic publications as shown in Figure 2. Various kinds of entities (researchers, conferences/journals, papers, words) are considered. Scientific literature connects papers by citations; papers are published at different places (conferences, journals, workshops, thesis, etc.); and researchers are connected to papers through authorship. Some might relate to each other by serving simultaneously as journal editors or on conference program committees. Moreover, each paper can focus on different topics, which are represented by words. Words are associated to each other based on semantics. At the same time, papers connect to different conferences, journals (venues for publication). In the network, there are multiple types of entities. And entities relate to others (either the same type or different types) through different links.

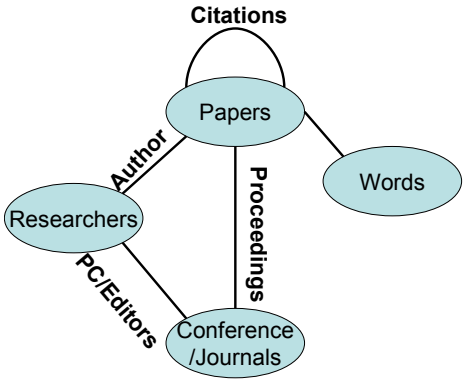

Fig. 2. A multi-mode network in academia

– *Multi-Dimensional Networks* [23, 20]. A multi-dimensional network has multiple types of interactions between the same set of users. Each dimension of the network represents one type of activity between users. For instance, in Figure 3, at popular photo and video sharing sites (e.g., Flickr and YouTube), a user can connect to his friends through email invitation or the provided "add as contacts" function; users can also tag/comment on the social contents like photos and videos; a user at YouTube can respond to another user

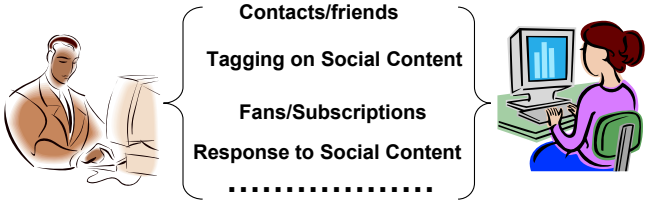

Fig. 3. An Example of Multi-Dimensional Network

by uploading a video; and a user can also become a fan of another user by subscription to the user's contributions of social contents. A network among these users can be constructed based on each form of activity, in which *each dimension represents one facet of diverse interaction.*

Actually, directed networks can be considered as a special case of multi-dimensional network. Take email communications as an example. People can play two different roles in email communications: senders and receivers. These two roles are not interchangeable. Spammers send an overwhelming number of emails to normal users but seldom receive responses from them. The sender and receiver roles essentially represent two different interaction patterns. A 2-dimensional network can be constructed to capture the roles of senders and receivers. In the first dimension, two actors are deemed related if they both send emails to the same person; in the other dimension, two actors interact if they both receive emails from another actor. A similar idea is also adopted as "hubs" and "authorities" on Web pages [10].

In this chapter, we do not use the notion of *multi-relational network*, as "multi-relational" has been used with different connotations depending on the domains. For example, multi-relational data mining [4], originating from the database field, focuses on data mining tasks with multiple relational tables. This concept can be extended to networks as well. One special case is that, each table is considered as interactions of two types of entities, leading to a multi-mode network. Meanwhile, social scientists [27] use multi-relational network for a different meaning. A multi-relational network is a network in which the connections (edges) between actors represent different type of relations, e.g., father-of, wife-of, etc. If each type of interaction in a multi-dimensional network represents one relation, the multi-dimensional network is equivalent to a multi-relational network.

Note that the two types of heterogeneous networks (multi-mode and multi-dimensional) mentioned above are not exclusive. A complicated network can be both multi-mode and multi-dimensional at the same time. As presented later, techniques to address these two types of networks can be fused together for community discovery.

2.2 Motivations to Study Network Heterogeneity

Social media offers an easily-accessible platform for diverse online social activities, but also introduces heterogeneity in networks. Thus, it calls for solutions to extract communities in heterogeneous networks, which will be covered in the next section. However, it remains unanswered why one cannot reduce a heterogeneous network to several homogeneous ones (i.e., one mode or one dimension) for investigation.

The reason is that the interaction information in one mode or one dimension might be too noisy to detect meaningful communities. For instance, in the YouTube example in Figure 1. It seems acceptable if we only consider the user mode. In other words, just study the friendship network. On the one hand, some users might not have any online friends either because they are too introvert to talk to other online users, or because they just join the network and are not ready for or not interested in connections. On the other hand, some users might abuse connections, since it is relatively easy to make connections in social media compared with in the physical world. As mentioned in [18], a user in Flickr can have thousands of friends. This can hardly be true in the real world. It might be the case that two online users get connected but they never talk to each other. Thus, these online connections of one mode or one dimension can hardly paint a true picture of what is happening.

A single type of interaction provides limited (often sparse) information about the community membership of online users. Fortunately, social media provides more than just a single friendship network. A user might engage in other forms of activities besides connecting to friends. It is helpful to utilize information from other modes or dimensions for more effective community detection. It is empirically verified that communities extracted using multi-mode or multi-dimensional information are more accurate and robust [23].

3 Community Extraction in Heterogeneous Networks

We first formulate the community detection problems for multi-mode networks and multi-dimensional networks, respectively; and then present viable solutions and their connections.

3.1 Multi-Mode Networks

Given an m-mode network with m types of actors

$$\mathbb{X}_i = \{x_1^i, x_2^i, \cdots, x_{n_i}^i\} \quad i = 1, \cdots, m$$

where n_i is the number of actors for $\mathbb{X}_i$, we aim to find community structures in each mode. Let $R_{i,j} \in \mathbb{R}^{n_i \times n_j}$ denote the interaction between two modes of actors $\mathbb{X}_i$ and $\mathbb{X}_j$, k_i and k_j denote the number of latent communities for $\mathbb{X}_i$ and

Table 1. Notations

symbol	representation
m	number of modes in a multi-mode network
n_i	number of actors in mode i
k_i	number of communities at mode i
$R_{i,j}$	interaction matrix between modes i and j
C_i	community indicator matrix of mode i
$A_{i,j}$	group interaction density between modes i and j
c^i_{st}	the (s,t)-th entry of C_i
$\mathcal{R}$	a multi-dimensional network
R_d	the dth dimension of multi-dimensional network
n	number of actors within a multi-dimensional network
d	the dimensionality of a multi-dimensional network
k	number of communities within a network
C	the community indicator matrix

$\mathbb{X}_j$, respectively. The interactions between actors can be approximated by the interactions between groups in the following form [12]:

$$R_{i,j} \approx C_i \; A_{i,j} \; C_j^T$$

where $C_i \in \{0,1\}^{n_i \times k_i}$ denotes some latent cluster membership for $\mathbb{X}_i$, $A_{i,j}$ the group interaction, and T the transpose of a matrix. In other words, the group identity determines how two actors interact, essentially making a similar assumption as that of block models [17]. The difference is that block models deal with the problem from a probabilistic aspect and concentrate on one-mode or two-mode networks. Here we try to identify the block structure of multi-mode networks via matrix approximation:

$$\min \sum_{1 \le i < j \le m} w_{ij} \| R_{i,j} - C_i A_{i,j} C_j^T \|_F^2 \tag{1}$$

$$s.t.\ C_i \in \{0,1\}^{n_i \times k_i} \quad i = 1, 2, \ldots, m \tag{2}$$

$$\sum_{t=1}^{k_i} c^i_{st} = 1, \quad s = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, m, \tag{3}$$

where w_{ij} are the weights associated with different interactions and c^i_{st} the (s,t)th entry of C_i.

The constraints in Eq. (3) force each row of the indicator matrix to have only one entry being 1. That is, each actor belongs to only one community. Unfortunately, the discreteness of the constraints in Eq. (2) makes the problem NP-hard. A strategy that has been well studied in spectral clustering is to allow the cluster indicator matrix to be continuous and relax the hard clustering constraint as follows:

$$C_i^T C_i = I_{k_i}, \quad i = 1, 2, \ldots, m \tag{4}$$

This continuous approximation of C_i can be considered as a low-dimensional embedding such that the community structure is more prominent in these dimensions. Consequently, the problem can be reformulated as

$$\min_{C,\ A} \sum_{1\le i<j\le m} w_{ij}\|R_{i,j} - C_i\ A_{i,j}\ C_j^T\|_F^2 \tag{5}$$

$$s.t. \quad C_i^T C_i = I_{k_i}, \quad i = 1, 2, \ldots, m \tag{6}$$

Since the solution of C_i of the above formulation is continuous, a post-processing step is required to obtain the disjoint partition of actors. A commonly used technique is to treat each column of C_i as features and then conduct k-means clustering to obtain discrete assignment of clusters [13]. Below, we briefly describe the computation of $A_{i,j}$ and C_i in Eq. (5).

Note that the problem in Eq. (5) is too complicated to derive a closed-form solution. However, it can be solved iteratively. First, we show that $A_{i,j}$ has a closed-form solution when C_i is fixed. Then, we plug in the optimal $A_{i,j}$ and compute C_i via alternating optimization. Basically, fix the community indicator at all other modes while computing the community indicator C_i at mode i. We only include the key proof here due to the space limit. Please refer to [12, 22] for details.

Theorem 1. *Given C_i and C_j, the optimal group interaction matrix $A_{i,j}$ can be calculated as*

$$A_{i,j} = C_i^T R_{i,j} C_j \tag{7}$$

Proof. Since $A_{i,j}$ appears only in a single term, we can focus on the term to optimize $A_{i,j}$.

$$\begin{aligned}
&\|R_{i,j} - C_i A_{i,j} C_j^T\|_F^2 \\
&= tr\left[(R_{i,j} - C_i A_{i,j} C_j^T)(R_{i,j} - C_i A_{i,j} C_j^T)^T\right] \\
&= tr\left[R_{i,j} R_{i,j}^T - 2 C_i A_{i,j} C_j^T R_{i,j}^T + A_{i,j} A_{i,j}^T\right]
\end{aligned}$$

The second equation is obtained based on the property that $tr(AB) = tr(BA)$ and column orthogonality of C_i and C_j. Setting the derivative with respect to $A_{i,j}$ to zero, we have $A_{i,j} = C_i^T R_{i,j} C_j$. The proof is completed. □

Given the optimal $A_{i,j}$ as in Eq. (7), it can be verified that

$$\|R_{i,j} - C_i A_{i,j} C_j^T\|_F^2 = \|R_{i,j}\|_F^2 - \|C_i^T R_{i,j} C_j\|_F^2 \tag{8}$$

Since $\|R_{i,j}^t\|_F^2$ in (8) are constants, we can transform the formulation in Eq. (5) into the following objective:

$$\max \sum_{1\le i<j\le m}^{m} w_{i,j}\|C_i^T R_{i,j} C_j\|_F^2 \tag{9}$$

$$s.t.\ C_i^T C_i = I_{k_i}, \quad i = 1, 2, \ldots, m \tag{10}$$

Note that C_i is interrelated with C_j $(j \neq i)$. There is no closed-form solution in general. However, given C_j $(j \neq i)$, the optimal C_i can be computed as follows:

Theorem 2. *Given C_j $(j \neq i)$, C_i can be computed as the top left singular vectors of the matrix P_i concatenated by the following matrices in column-wise:*

$$P_i = \left[\left\{ \sqrt{w_{ij}}\ R_{i,j} C_j \right\}_{i<j}, \quad \left\{ \sqrt{w_{ki}}\ R_{k,i}^T C_k \right\}_{k<i} \right] \tag{11}$$

Proof. We only focus on those terms in the objective involving C_i.

$$\begin{aligned} L &= \sum_{i<j} w_{ij} \|C_i^T R_{i,j} C_j\|_F^2 + \sum_{k<i} w_{ki} \|C_k^T R_{k,i} C_i\|_F^2 \\ &= \sum_{i<j} w_{ij}\ tr\left(C_i^T R_{i,j} C_j C_j^T R_{i,j}^T C_i\right) + \sum_{k<i} w_{ki}\ tr\left(C_i^T R_{k,i}^T C_k C_k^T R_{k,i} C_i\right) \\ &= tr\left[C_i^T \left(\sum_{i<j} w_{ij} R_{i,j} C_j C_j^T R_{i,j}^T + \sum_{k<i} w_{ki} R_{k,i}^T C_k C_k^T R_{k,i} \right) C_i \right] \\ &= tr\left(C_i^T M_i C_i\right) \end{aligned}$$

where M_i is defined as

$$M_i = \sum_{i<j} w_{ij} R_{i,j} C_j C_j^T R_{i,j}^T + \sum_{k<i} w_{ki} R_{k,i}^T C_k C_k^T R_{k,i} \tag{12}$$

So the problem boils down to a well-defined max-trace problem with orthogonality constraints. The community indicator matrix C_i has a closed-form solution, which corresponds to the subspace spanned by the top k_i eigenvectors of M_i. Note that M_i is normally a dense $n_i \times n_i$ matrix. Direct calculation of M_i and its eigenvectors is expensive if n_i is huge (which is typically true in social media). However, M_i can be written as

$$M_i = P_i P_i^T \tag{13}$$

where P_i is defined as in Eq. (11). Thus the optimal C_i, which corresponds to the top eigenvectors of M_i can be computed as the top left singular vectors of P_i. Note that the ordering of columns in P_i does not affect the final solution. □

As can be seen in Eq. (11), the clustering results of interacted entities, essentially form weighted features for clustering of the ith mode. The matrix M_i, being the outer product of P_i, acts like a similarity matrix for clustering. Based on Theorem 2, we can update the cluster indicator matrix iteratively based on the "attributes" obtained from the clustering results of related entities.

Once the approximate cluster indicator matrix C_i is computed, k-means can be applied to obtain the discrete assignment of communities for actors at each mode. The overall description of the algorithm is presented in Figure 4. In the algorithm, we specify the objective to be calculated via Eq. (9), as the direct calculation of the original formation in Eq. (5) usually requires computation of dense matrices, which is not applicable for large-scale multi-mode networks.

Input: $\{R_{i,j}\}$, $\{k_i\}$, $\{w_{ij}\}$
Output: $\{idx_i\}$, $\{C_i\}$, $\{A_{i,j}\}$
1. Generate initial cluster indicator matrix $\{C_i\}$.
2. **Repeat**
3. **For** $i = 1, 2, \ldots, m$
4. construct P_i as in Eq. (11);
5. update C_i as top k_i left singular vectors of P_i;
6. **Until** the relative change of the objective in Eq. (9) $\leq \epsilon$.
7. calculate $A_{i,j}$ as in Eq. (7)
8. calculate the cluster idx_i with k-means on C_i

Fig. 4. Algorithm for Community Extraction in Multi-Mode Networks

3.2 Multi-Dimensional Networks

In a multi-dimensional network, there are multiple dimensions of interactions between the same set of users. A latent community structure in social media exists among these actors, indicating various interactions along different dimensions. The goal of community extraction in a multi-dimensional network is to infer the shared community structure. A d-dimensional network is represented as

$$\mathcal{R} = \{R_1, R_2, \ldots, R_d\}$$

R_i represents the interactions among actors in the ith dimension. For simplicity, we assume the interaction matrix R_i is symmetric. We use $C \in \{0, 1\}^{n \times k}$ to denote the community membership of each actor.

Since the goal of community extraction in multi-dimensional networks is to identify a shared community structure that explains the interaction in each dimension, one straightforward approach is to average the interaction in each dimension, and treat it as a normal single-dimensional network. Then, any community extraction methods proposed for networks or graphs can be applied. This simple averaging approach becomes problematic if the interaction in each dimension is not directly comparable. For example, it can be the case that users interact with each other frequently in one dimension (say, leave some comments on friend's photos), whereas talk to each other less frequently in another dimension (say, sending emails in Facebook). Averaging the two types of interaction might misrepresent the hidden community information beneath the latter dimension with less frequent interactions. One way to alleviate this problem is to assign different weights for each dimension. Unfortunately, it is not an easy task to assign appropriate weights for effective community extraction.

Another variant is to optimize certain averaged clustering criteria. Let $Q_i(C)$ denote the cost of community structure C on the ith dimension of interaction R_i. We list some representative criteria in existing body of literature as follows:

– Block model approximation [28] minimizes the divergence of the interaction matrix and block model approximation:

$$\min Q = \ell(R; C^T \Lambda C) \tag{14}$$

where ℓ is a loss function to measure the difference of two matrices, and Λ a diagonal matrix roughly representing the within-group interaction density.
- Spectral Clustering [13] minimizes the following cost function

$$\min Q = tr(C^T LC) \tag{15}$$

where L is the graph Laplacian.
- Modularity maximization [15] maximizes the modularity of a community assignment:

$$\max Q = tr(C^T BC) \tag{16}$$

where B is a modularity matrix.

Given a multi-dimensional network, we can optimize the following cost function,

$$\min_C \sum_{i=1}^{d} w_i Q_i(C) \tag{17}$$

The weighted optimization criterion with graph Laplacian and random walk interpretation are presented in [29]. Weighted modularity maximization is explored in [23] as a baseline approach.

The drawback of the aforementioned two approaches (averaging network interactions or minimize average cost) is that they can be sensitive to noisy interactions. Assigning proper weights can help alleviate the problem, but it is equally, if not more, difficult to choose a good heuristic of weighting scheme. Instead, an alternative paradigm based on structural features is proposed in [23] to overcome these disadvantages. The basic idea is that the community structure extracted from each dimension of the network should be similar. Hence, we can extract the "rough" community structure at each dimension, and then integrate them all to find out the shared community structure. Thus, the paradigm consists of two phases: (i) *structural feature extraction from each dimension* and (ii) *cross-dimension integration.*

- **Phase I: Structural Feature Extraction** Structural features, which are indicative of some community structure, are extracted based on network connectivity. Any methods finding out a community assignment can be used to extract structural features. Note that finding out a discrete assignment of clusters with respect to the criteria in Eq. (14) - Eq. (16) is NP-complete. Commonly used algorithms are very sensitive to network topology [7] and suffer from local optima. In practice, some approximation scheme of the discrete assignment is often exploited.
  One widely used relaxation, as we have done in the previous section, is to allow C to be continuous while satisfying certain orthogonal constraints (i.e., $C^T C = I_k$). This relaxation results in an approximation of C which can be considered as a lower-dimensional embedding that captures the community structure. The optimal C typically corresponds to the top eigenvectors of a certain matrix. This relaxation is adopted in [15] for modularity maximization and many spectral clustering approaches [13]. Note that after relaxation,

the obtained community indicator matrix C is typically globally optimal with respect to certain criteria. This avoids the randomness of a discrete assignment due to the noise in network connections or algorithm initialization. Hence, structural feature extraction based on relaxed community indicator is a more favorable solution. Networks in social media are very noisy. Extracting some prominent structural features indeed helps remove the noise, enabling more accurate community identification in the second stage.

– **Phase II: Cross-Dimension Integration** Assuming a latent community structure is shared across dimensions in a multi-dimensional network, we expect that the extracted structural features to be "similar". However, dissimilar structural feature-values do not necessarily indicate that the corresponding community structures are different as an orthogonal transformation or reordering of columns in C can be "equivalent" solutions [23]. Instead, we expect the structural features of different dimensions to be highly correlated after certain transformation. Thus, the integration boils down to finding transformations that can be applied to the extracted structural features to maximize the correlation.

  To capture the correlations between multiple sets of variables, (generalized) canonical correlation analysis (CCA) [9] is a standard statistical technique. CCA attempts to find a linear transformation for each set of variables such that the pairwise correlations are maximized. It has been widely used to integrate information from multiple different sources or views [16, 6]. Here we briefly illustrate one scheme of generalized CCA that turns out to equal to principal component analysis (PCA) under certain constraints.

  Let $C_i \in R^{n\times \ell_i}$ denote the ℓ_i structural features extracted from the ith dimension of the network, and $w_i \in \mathbb{R}^{\ell_i}$ be the linear transformation applied to the structural features of network dimension i. The correlation between two dimensions after transformation is

$$(C_i w_i)^T (C_j w_j) = w_i^T (C_i^T C_j) w_j = w_i^T O_{ij} w_j$$

  with $O_{ij} = C_i^T C_j$ representing the covariance between the structural features of the ith and the jth dimensions. One scheme of generalized CCA attempts to maximize the summation of pairwise correlations in the following form:

$$\max \sum_{i=1}^{d} \sum_{j=1}^{d} w_i^T O_{ij} w_j \tag{18}$$

$$s.t. \sum_{i=1}^{d} w_i^T O_{ii} w_i = 1 \tag{19}$$

  Here, the objective in Eq. (18) is to maximize the pairwise correlations; and the constraints in Eq. (19) confine the scale of transformation. Using standard Lagrange multiplier and setting the derivatives respect to w_i to

Input: $\mathcal{R} = \{R_1, R_2, \ldots, R_d\}$, k , ℓ
Output: idx, C_i $(i = 1, 2, \ldots, d)$

1. Compute top ℓ structural features C_i based certain criteria as in Eq. (14)-(16);
2. Construct $X = [C_1, C_2, \ldots, C_d]$;
3. Compute slim SVD of $X = UDV^T$;
4. Obtain lower-dimensional embedding $\widetilde{U} = U(:, 1 : k-1)$;
5. Calculate the cluster idx with k-means on $\widetilde{U}$.

Fig. 5. Algorithm for Community Extraction in Multi-Dimensional Networks

zero, we obtain the following (where λ is a Lagrange multiplier):

$$\begin{bmatrix} O_{11} & O_{12} & \cdots & O_{1d} \\ O_{21} & O_{22} & \cdots & O_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ O_{d1} & O_{d2} & \cdots & O_{dd} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \lambda \begin{bmatrix} O_{11} & 0 & \cdots & 0 \\ 0 & O_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & O_{dd} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \tag{20}$$

Recall that our structural features extracted from each dimension satisfy $C_i^T C_i = I$. Thus, the matrix $diag(O_{11}, O_{22}, \ldots, O_{dd})$ in Eq. (20) becomes an identity matrix. Hence $\mathbf{w} = [w_1, w_2, \cdots, w_d]^T$ corresponds the top eigenvector of the full covariance matrix on the left-hand side in Eq. (20). This essentially equals to PCA applied to data of the following form:

$$X = [C_1, C_2, \ldots, C_d] \tag{21}$$

After the transformation w to the structural feature sets, the corresponding community at each dimension get aligned with each other. In order to partition the actors into k disjoint communities, we can extract the top $k-1$ dimensions such that the community structure is most prominent. Let $X = UDV^T$ be the SVD of X. It follows that the top $(k-1)$ vectors of U are the lower-dimensional embedding.

In summary, to handle multiple dimensions of network interaction, we can first extract structural features from each dimension. Then, we concatenate all the structural features and perform PCA to find out the low-dimensional embedding. Based on the embedding, k-means can be applied to find out the discrete community assignment. The detailed algorithm is summarized in Figure 5.

Different from the two alternatives (average interaction or average criteria to optimize), the proposed approach is more robust to noisy interactions in multi-dimensional networks [23]. Moreover, this scheme does not require any weighting scheme for real-world deployment.

3.3 Connections between Multi-Mode and Multi-Dimensional Networks

Comparing the algorithms for multi-mode networks and multi-dimensional networks, we can find a common component: *extract structural features and con-*

catenate them to form a feature-based data set of actors, and then apply SVD to obtain the lower-dimensional embedding (Steps 4-5 in Figure 4 and Steps 2-4 in Figure 5). The basic scheme is to *convert the network interactions into features*. This scheme can work not only for community identification, but also for relational learning and behavior prediction [18].

A social media network can be both multi-mode and multi-dimensional. One can combine the two algorithms to handle multi-mode and multi-dimensional challenges. The combination is straightforward: if there are within-mode interactions that are multi-dimensional, we can simply append to P_i in Eq. (11) with some structural features that are indicative of the community structure. That is,

$$P_i = \left[\left\{ \sqrt{w_{ij}}\ R_{i,j} C_j \right\}_{i<j}, \quad \left\{ \sqrt{w_{ki}}\ R^T_{k,i} C_k \right\}_{k<i}, \quad \{C^d_i\} \right] \tag{22}$$

where C^d_i denotes the structural features extracted from dth dimension of interaction in the ith mode. In this way the presented algorithm is able to handle diverse heterogeneous networks.

4 Understanding Groups

In earlier sections, we concentrate on group structures. That is, how to extract groups from network topology. Extracting groups is the first step for further analysis to answer questions such as *why are these people connected to each other?* and *what is the relationship between different groups?* In this section, we seek to capture group profiles in terms of topics or interests they share [24]. This helps understand group formation as well as other group related task analysis. As the total number of groups' interests can be huge and might change over time, a static group profile cannot keep pace with an evolving environment. Therefore, *online group profiling based on topic taxonomy* [21] is proposed to serve the need.

4.1 Group Profiling

While a large body of work has been devoted to discover groups based on network topology, few systematically delve into the extracted groups to understand the formation of a group. Some fundamental questions remain unaddressed:

> What is the particular reason that binds the group members together? How to *interpret* and *understand* a social structure emanated from a network?

Some work attempts to understand the group formation based on statistical structural analysis. Backstrom et al. [2] studied prominent online groups in the digital domain, aiming at answering some basic questions about the evolution of groups, like *what are the structural features that influence whether individuals will join communities*. They found that the number of friends in a group is the most important factor to determine whether a new user would join the group. This result is interesting, though not surprising. It provides a global level of

structural analysis to help understand how communities attract new members. However, more efforts are required to understand the formation of a particular group.

According to the concept of Homophily [14], a connection occurs at a higher rate between similar people than dissimilar people. Homophily is one of the first characteristics studied by early social science researchers and holds for a wide variety of relationships [14]. Homophily is also observed in social media [5, 26]. In order to understand the formation of a group, the *inverse* problem can be investigated: Given a group of users, can we figure out why they are connected? What are their shared similarities? ***Group Profiling*** [24], by extracting shared attributes of group members, is one approach proposed to address the problem.

Besides understanding social structures, group profiling also helps for network visualization and navigation. It has potential applications for event alarming, direct marketing, or group tracking. As for direct marketing, it is possible that the online consumers of products naturally form several groups, and each group posts different comments and opinions on the product. If a profile can be constructed for each group, the company can design new products accordingly based on the feedback of various groups. Group profiles can be also used to connect dots on the Web. It is noticed that an online network (e.g., blogosphere) can be divided into three regions: singletons who do not interact with others, isolated communities, and a giant connected component [11]. Isolated communities actually occupy a very stable portion of the entire network, and the likelihood for two isolated communities to merge is very low as the network evolves. If group profiles are available, it is possible for one group or a singleton to find other similar groups and make connections of segregated groups of similar interests.

A set of topics can be used to describe a group. Since a group consists of people with shared interests, one intuitive way of group profiling is to clip a group with some topics shared by most members in the group. Luckily, social media provides not only network connectivity, but also textual information. For instance, in blogosphere, bloggers upload blog posts; in content sharing sites like Digg.com and Del.icio.us, users post news or bookmarks. These content information essentially represents the latent interests of individuals. Moreover, users also provide tags on the shared content. These tags can serve as topics.

In order to achieve effective group profiling, one straightforward approach is *aggregation*. For instance, if a tag is commonly used by the majority of group members, then the tags with highest frequency can be used to describe the group. This technique is widely used to construct tag clouds to capture the topic trend of a social media site. However, as pointed out in [24], aggregation can lead to selection of irrelevant tags for a group, especially those popular tags. This is even worse if the topics are extracted from raw text such as blog posts, comments, and status updates. Instead, to find out the description of a group, *differentiation*-based method can be exploited. That is, we can treat the group as a positive class, and the remaining actors in the network as a negative class. Then, only those features that occur frequently in the group while rarely outside the group are selected. More interestingly, it is empirically shown that by

Table 2. Profiles constructed by various strategies for Blythedoll group in LiveJournal. Each profile consists of the top 10 selected features. The first block show the profiles constructed based on individual interests on user profiles, and the second block based on group members' blog posts.

Profiles based on Individual Interests		
Aggregation	Differentiation	Ego-Differentiation
blythe	blythe	blythe
photography	dolls	dolls
sewing	sewing	sewing
japan	japan	blythe dolls
dolls	blythe dolls	super dollfie
cats	super dollfie	japan
art	hello kitty	hello kitty
music	knitting	toys
reading	toys	knitting
fashion	junko mizuno	re-ment

Profiles based on Blog Posts		
Aggregation	Differentiation	Ego-Differentiation
love	blythe	blythe
back	doll	doll
ll	flickr	dolly
people	ebay	dolls
work	dolls	ebay
things	photos	sewing
thing	dolly	flickr
feel	outfit	blythes
life	sell	outfit
pretty	vintage	dollies

comparing the group with its neighboring actors (those actors outside the group but connecting to at least one member in the group), the extracted features are equivalently informative. Essentially, we can consider the group as a unit and take an egocentric view. The group profiles can be extracted by differentiate the group from their friends (denoted as *ego-differentiation*).

Table 2 shows one example of profiles extracted based on different strategies on Blythedoll group[1] of over 2000 members in a popular blog site LiveJournal[2]. Blythedoll was first created in 1972 by U.S. toy company Kenner, later it spread out to the world. Takara, a Japanese company, is one of the most famous producers. As seen in the table, the aggregation-based method tends to select some popular interests such as *music*, *photography*, *reading* and *cats*. On the contrary, differentiation based methods select interests that are more descriptive. This pattern is more observable when the profiles are constructed from individual blog posts. Aggregation reports a profile that is hardly meaningful, while differ-

[1] `http://community.livejournal.com/blythedoll/profile`
[2] http://www.livejournal.com/

entiation still works reasonably well. Even if we take an egocentric-view for the differentiation-based method, a similar result is observed.

4.2 Topic Taxonomy Adaptation

In social media, there are hundreds of thousands of online groups with diverse interests. The topics associated with different groups can be inordinate, and the total number of topics can be huge. Moreover, the selected topics in group profiles can be highly correlated as different users use tags or words at different granularity. Facing a large number of topics, we need to find a more suitable representation to understand the relationship between different groups.

Organizing the topics into a tree-structured taxonomy or hierarchy is a natural solution, as it provides more contextual information with refined granularity compared with a flat list. The left tree in Figure 6 shows one simple example of a topic taxonomy. Basically, each group is associated with a list of topics. Each topic can be either a non-leaf (internal) node like *Meteorology* or *Politics*, or a leaf node like *Hurricane*. Different groups can have shared topics. Given a topic taxonomy, it is easy to find related or similar topics via parent, sibling, or child nodes. Taxonomies also facilitate the visualization of relationships between different groups and the detection of related or similar groups.

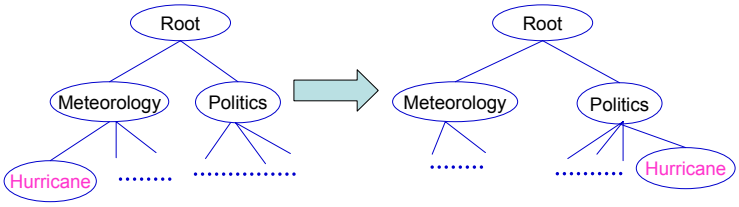

Fig. 6. "Hurricane" Example

A topic taxonomy can be provided by human beings based on topic semantics or abridged from a very large taxonomy like Yahoo! or Google directory. It is a relatively stable description. However, group interests develop and change. Let us look at an example about "Hurricane" [25]. As shown in Figure 6, in a conventional topic taxonomy, the topic *Hurricane* is likely to locate under *Meteorology*, and not related to *Politics*. Suppose we have two groups: one is interested in *Meteorology* and the other in *Politics*. The two groups have their own interests. One would not expect that "Hurricane" is one of the key topics under *Politics*. However, in a period of time in 2005, there was a surge of documents/discussions on "Hurricane" under *Politics*. Before we delve into why this happened, this example suggests *the change of group interests and the need for corresponding change of the taxonomy*. This reason for this shift is that, a good number of online documents in topic *Hurricane* are more about *Politics* because

Hurricanes 'Katrina' and 'Rita' in the United States in 2005 caused unprecedented damages to life and properties; and some of the damages might be due to the faults of federal emergency management agency in preparation for and responding to the disasters.

This example above demonstrates some inconsistency between a stagnant taxonomy and changing interests of an online group. Group interests might shift and the semantics of a topic could be changed due to a recent event. To enable a topic taxonomy to profile the changing group interest, we need to allow the topic taxonomy to adapt accordingly and reflect the change. The dynamic changes of semantics are reflected in documents under each topic, just like in the *hurricane example*. This observation motivates us to adjust a given topic taxonomy in a data-driven fashion.

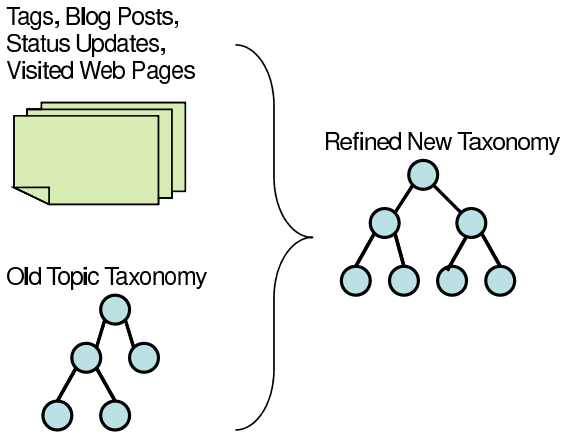

Fig. 7. Topic Taxonomy Adaptation

Figure 7 illustrates a typical process of topic taxonomy adaption. By observing the difference between the original taxonomy and the newly generated taxonomy, we notice that topics can emerge and disappear for various groups. Given recent text data (e.g., tags, blog posts, visited web pages, submitted search queries) extracted from social media and a given topic taxonomy, we aim to automatically find a revised taxonomy of topics (tags) that is consistent with the data and captures dynamic group interests.

One fundamental question is how to measure the discrepancy between the semantics reflected in texutal contents and a topic taxonomy. While it is a thorny challenge to quantify the discrepancy, a surrogate measure, the classification performance based on the topic taxonomy, can be calibrated. In order to obtain the classification performance, we can exploit the content and tag information from social media. *The tags provide topic information while the shared contents act like data*. With a robust hierarchical classifier built from some collected data and an existent taxonomy, new documents can be labeled automatically by the

classifier. If the label are consistent with the associated tags, the taxonomy, in a sense, captures the relationship of tags. So the corresponding classification performance based on a taxonomy is one effective way of indirectly measuring how good a topic taxonomy is to represent relationships of different topics. In other words, the quality of a topic taxonomy reduces to the classification performance (e.g. recall, precision, ROC, etc.) based on the taxonomy.

We can change the topic taxonomy via classification learning as shown in Figure 8. Suppose a topic taxonomy is constructed based on text information from before. The taxonomy is then adapted to maximize the classification performance on the newly arrived texts. The basic idea is, given a predefined taxonomy, a different hierarchy can be obtained by performing certain operations. Then, the newly generated hierarchy is evaluated on collected shared contents with tag information. If the taxonomy change results in a performance improvement, it is kept; otherwise, alternative change to the original taxonomy is explored. This process is repeated until no more change can lead to performance improvement, ending up with a new taxonomy which acclimatizes the taxonomic semantics according to the contents.

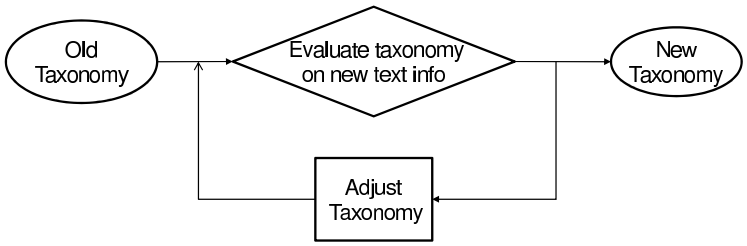

Fig. 8. Taxonomy Adaptation via Classification Learning

Since a topic taxonomy does not change considerably in a short time period, we expect only a small portion of tags change their positions in the taxonomy. Tang et al. [25, 21] propose to adapt a provided taxonomy *locally* according the classification performance on novel data. Three elementary operations are defined to change a taxonomy locally as shown in Figure 9:

- Promote: roll up one node to upper level;
- Demote: push down one node to its sibling;
- Merge: merge two sibling nodes to form a super node;

Since the defined local changes can be applied to any node in a taxonomy, the total number of operations can be abundant. Generally, the nodes at higher level plays a more important role for classification. Hence, it is proposed to follow a top-down traversal of a hierarchy to search for applicable operations [25]. It is empirically shown that two iterations of the traversal are often sufficient to achieve a robust taxonomy that captures the dynamic relationship between different groups.

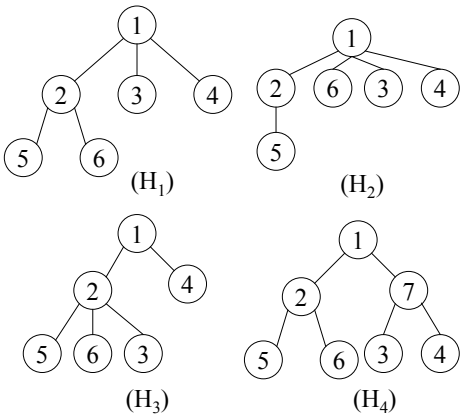

Fig. 9. Elementary Operations. H_1 is the original hierarchy. H_2, H_3 and H_4 are obtained by performing different elementary operations. H_2: Promote node 6; H_3: Demote node 3 under node 2; H_4: Merge node 3 and node 4.

5 Summary and Future Work

Social media is replete with diverse and unique information. It provides heterogeneous network data as well as collective wisdom in forms of user-generated contents and tags. In this chapter, we present important research tasks and intriguing challenges with social media, and elaborate issues related to the understanding of online group structures and properties. In particular, we discuss two aspects of the problem: (1) how to extract communities given multi-mode and multi-dimensional data, and (2) how to dynamically capture group profiles and relationships.

The social-media networks are heterogeneous: their idiosyncratic entities and various interactions within the same network result in multi-mode and multi-dimensional networks, respectively. Although abundant, the information can be sparse, noisy, and partial. Therefore, special care is required to understand group structures and properties. We present some feasible solutions to extract reliable community structures in both types of networks. We also show that the two algorithms share a common component to extract "structural features" from each mode or dimension and then concatenate them to find some lower-dimensional embedding which is indicative of some community structure. This simple scheme has been shown effective in community extraction in social media. Another task equally important to community extraction is to capture group interests based on textual and tag information. We describe strategies to perform effective group profiling, as well as topic taxonomy adaptation to capture dynamic group relationship using noisy and time-sensitive tag and content information.

This chapter has only addressed a couple of essential issues. Many research directions are worthy pursuing in our endeavor to understand group structures and properties in social media. We propose the following for further research:

- How can one determine the number of communities in heterogeneous networks? In the current models, we assume the number of communities at each mode or dimension is fixed. Some parameter-free process will be very useful to automatically determine the number of communities.
- It is interesting to study communities at different degrees of granularity in heterogeneous networks. One possibility is to handle heterogeneity with hierarchical clustering.
- To deal with multi-dimensional networks, our current solution is to integrate different dimensions of interactions globally. Since it is more likely that some groups are more involved in one dimension than in other dimensions, can we integrate the interactions in different dimensions differently depending on dimensional intensities? It is a challenge to simultaneously discover a common community structure as well as the integration scheme for each group.
- Extracting communities in dynamic heterogeneous networks demands for effective solutions. Social media is evolving continuously, newcomers joining the network, extant members generating new connections or becoming dormant. It is imperative to efficiently update the acquired community structure. It is also interesting to consider the temporal change of individuals for community detection.
- The work of group profiling only employs descriptive tags and contents to profile groups. More can be attempted for group profiling. For example, How to integrate the differentiation-based profiling into a taxonomy? Though the current taxonomy representation of topics does not allow one topic to have multiple parent nodes (topics), tags (especially those words with multiple meanings) can relate to different parent nodes depending on the context.
- The current scheme of group profiling is separated from group detection. If the associated tags and contents could be considered as one mode, it may be possible to exploit the methods developed for multi-mode networks to handle joint group detection and profiling.

In a nutshell, social media is a rich data source of large quantity and high variety. It is a fruitful field with many great challenges for data mining. In achieving the understanding of group structures and properties in social media, we genuinely expect that this line of research will help identify many novel problems as well as new solutions in understanding social media.

Acknowledgments This work is, in part, supported by ONR and AFOSR.

References

1. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: WSDM '08: Proceedings of the international conference on Web search and web data mining. pp. 207–218. ACM, New York, NY, USA (2008)

2. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 44–54. ACM, New York, NY, USA (2006)
3. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Comput. Surv. 38(1), 2 (2006)
4. Džeroski, S.: Multi-relational data mining: an introduction. SIGKDD Explor. Newsl. 5(1), 1–16 (2003)
5. Fiore, A.T., Donath, J.S.: Homophily in online dating: when do you like someone like yourself? In: CHI '05: CHI '05 extended abstracts on Human factors in computing systems. pp. 1371–1374. ACM, New York, NY, USA (2005)
6. Hardoon, D.R., Szedmak, S.R., Shawe-taylor, J.R.: Canonical correlation analysis: An overview with application to learning methods. Neural Comput. 16(12), 2639–2664 (2004)
7. Hopcroft, J., Khan, O., Kulis, B., Selman, B.: Natural communities in large linked networks. In: KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 541–546. ACM, New York, NY, USA (2003)
8. Kang, H., Getoor, L., Singh, L.: Visual analysis of dynamic group membership in temporal social networks. SIGKDD Explorations, Special Issue on Visual Analytics 9(2), 13–21 (dec 2007)
9. Kettenring, J.: Canonical analysis of several sets of variables. Biometrika 58, 433–451 (1971)
10. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)
11. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 611–617. ACM, New York, NY, USA (2006)
12. Long, B., Zhang, Z.M., Wú, X., Yu, P.S.: Spectral clustering for multi-type relational data. In: ICML '06: Proceedings of the 23rd international conference on Machine learning. pp. 585–592. ACM, New York, NY, USA (2006)
13. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416 (2007)
14. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444 (2001)
15. Newman, M.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 74(3) (2006), http://dx.doi.org/10.1103/PhysRevE.74.036104
16. Nielsen, A.: Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. Image Processing, IEEE Transactions on 11(3), 293–305 (Mar 2002)
17. Nowicki, K., Snijders, T.A.B.: Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96(455), 1077–1087 (2001)
18. Tang, L., Liu, H.: Relational learning via latent social dimensions. In: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 817–826. ACM, New York, NY, USA (2009)
19. Tang, L., Liu, H.: Scalable learning of collective behavior based on sparse social dimensions. In: CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management. pp. 1107–1116. ACM, New York, NY, USA (2009)

20. Tang, L., Liu, H.: Uncovering cross-dimension group structures in multi-dimensional networks. In: SDM workshop on Analysis of Dynamic Networks (2009)
21. Tang, L., Liu, H., Zhang, J., Agarwal, N., Salerno, J.J.: Topic taxonomy adaptation for group profiling. ACM Trans. Knowl. Discov. Data 1(4), 1–28 (2008)
22. Tang, L., Liu, H., Zhang, J., Nazeri, Z.: Community evolution in dynamic multi-mode networks. In: KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 677–685. ACM, New York, NY, USA (2008)
23. Tang, L., Wang, X., Liu, H.: Uncovering groups via heterogeneous interaction analysis. In: Proceeding of IEEE International Conference on Data Mining. pp. 503–512 (2009)
24. Tang, L., Wang, X., Liu, H.: Understanding emerging social strucutres: A group-profiling approach. Tech. rep., Arizona State University (2010)
25. Tang, L., Zhang, J., Liu, H.: Acclimatizing taxonomic semantics for hierarchical content classification. In: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 384–393. ACM, New York, NY, USA (2006)
26. Thelwall, M.: Homophily in myspace. J. Am. Soc. Inf. Sci. Technol. 60(2), 219–231 (2009)
27. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
28. Yu, K., Yu, S., Tresp, V.: Soft clsutering on graphs. In: NIPS (2005)
29. Zhou, D., Burges, C.J.C.: Spectral clustering and transductive learning with multiple views. In: ICML '07: Proceedings of the 24th international conference on Machine learning. pp. 1159–1166. ACM, New York, NY, USA (2007)