Discovering Overlapping Groups in Social Media

Xufei Wang
Arizona State University
Tempe, AZ 85287, USA
Email:xufei.wang@asu.edu

Lei Tang*
Yahoo! Labs
Santa Clara, CA 95054, USA
Email:ltang@yahoo-inc.com

Huiji Gao
Arizona State University
Tempe, AZ 85287, USA
Email:huiji.gao@asu.edu

Huan Liu
Arizona State University
Tempe, AZ 85287, USA
Email:huan.liu@asu.edu

***Abstract*—The increasing popularity of social media is shortening the distance between people. Social activities, e.g., tagging in Flickr, bookmarking in Delicious, twittering in Twitter, etc. are reshaping people's social life and redefining their social roles. People with shared interests tend to form their groups in social media, and users within the same community likely exhibit similar social behavior (e.g., going for the same movies, having similar political viewpoints), which in turn reinforces the community structure. The multiple interactions in social activities entail that the community structures are often overlapping, i.e., one person is involved in several communities. We propose a novel co-clustering framework, which takes advantage of networking information between users and tags in social media, to discover these overlapping communities. In our method, users are connected via tags and tags are connected to users. This explicit representation of users and tags is useful for understanding group evolution by looking at who is interested in what. The efficacy of our method is supported by empirical evaluation in both synthetic and online social networking data.**

***Keywords*-Community Detection; Overlapping; Social Media; Co-Clustering;**

I. Introduction

The ubiquitous online social services enrich people's social activities with their families, friends and colleagues, exerting a vital impact on people's life, changing their ways of thinking and behaving. Social media sites including Facebook, Twitter, Wikipedia, Bloggers, Myspace are attracting more users than ever. In 2009, the global time spent on social media sites increased by 82%[1] than the year before. Facebook, one of the most popular social media sites, has more than 500 million active users and the number is still increasing[2]. The rapid increase in social media population suggests a dynamic social change and potential opportunities for social marketing businesses.

In social media websites, users are allowed to participate in social activities, e.g., connecting with other like-minded people, updating their status, posting blogs, uploading photos, bookmarks and tags, and so on. Besides, people can join explicit groups at different websites. For instance, fans of sports teams can join dedicated groups where they can share their opinions on team performance, comment on the newest information about player transfers. Studying individual behavior is usually difficult due to the extremely large population as well as the idiosyncrasy of human behavior. Studying statistics at website level often fail to catch sufficient detail. Group-level investigation can provide useful information with varying granuality.

A group (or community) can be considered as a set of users where each user interacts more frequently with users within the group than with users outside the group. Some social media websites (e.g., Flickr, Youtube) provide explicit groups which allow users to subscribe or join them. However, some highly dynamic sites (e.g., Twitter, Delicious) have no clear group structures, which requires quality community detection approaches to discover them. Community detection approaches are usually based on structural features (e.g., links). Since social media sites also provide metadata as well as content information, such information can also help to define the actors' social positions.

The diversity of people's interests and social interactions suggests that the community structures overlap. When there are explicit groups in social media websites, users are allowed to join more than one group based on their personal preferences and interests. When there are no explicit groups available, community detection algorithms can be used to obtain such groups. It is more reasonable to cluster users into overlapping communities. For instance, a user who is interested in football and iPad is very likely a member of two separated communities.

Social media, especially in blogosphere and bookmarking sites, provides both user information (e.g., friendship links, profiles) and user metadata (e.g., tags). These metadata contain clues to understanding communities in social media. Clustering homogeneous users and semantically close tags into communities simultaneously is a challenging but rewarding task. It is easy to obtain the common interests of a community by aggregating tags within it. This is helpful to study communities. Co-clustering is one way to obtain this kind of community structures. However, the constructed communities are disjoint which contradicts the actual social structures. Figure 1 is a toy example of two communities. Vertices $u_1 - u_5$ on the left represent users,

*This work was carried out when the author was at Arizona State University.

[1]http://blog.nielsen.com/nielsenwire/global/led-by-facebook-twitter-global-time-spent-on-social-media-sites-up-82-year-over-year/

[2]http://www.facebook.com/press/info.php?statistics

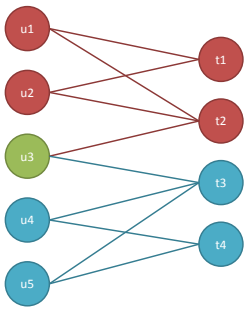

Figure 1. A 2-community toy example

$t_1 - t_4$ on the right represent tags and edges represent tag subscription relation between users and tags. According to Dhillon's [1] and Zha's [2] approaches, the singular vector corresponding to second largest singular value gives the bipartition information of the bipartite graph which is shown as follows:

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix} = \begin{bmatrix} 0.3536 \\ 0.3536 \\ 0.0000 \\ -0.3536 \\ -0.3536 \\ 0.3873 \\ 0.2582 \\ -0.2582 \\ -0.3873 \end{bmatrix} \quad (1)$$

If a clustering algorithm such as k-means is run on the singular vector, user u_3 will be assigned to either one of the two clusters if we want a bipartition. This disjoint clustering fails to uncover the real social roles of user $u3$. Based on the graph structure, it is more reasonable to have two overlapping clusters $(u_1, u_2, u_3, t_1, t_2)$ and $(u_3, u_4, u_5, t_3, t_4)$, in which the users' interests of each cluster can be summarized using t_1, t_2, and t_3, t_4, respectively.

An interesting observation in social life is that a social connection is often associated with one affiliation [3]. For instance, a person likes or dislikes a movie, he/she is or is not a member of special interest group, and so on. Instead of clustering vertices, clustering edges seems more appropriate in a sense. Clustering edges usually achieves overlapping communities. Look at the toy example shown in Figure 1, edges connecting to nodes t_1, t_2 and t_3, t_4 are clustered into two separate groups both containing user u_3. The difference between our work and traditional co-clustering of documents and words [1], [2] is that we allow cluster overlap. It is also different from fuzzy (soft) clustering because we assign discrete cluster membership. Thus our contributions are summarized as follows:

- We propose to discover overlapping communities in social media. Diverse interests and interactions that human beings can have in online social life suggest that one person often belongs more than one community.
- We use user-tag subscription information instead of user-user links. In social media, people can easily connect to thousands of like-minded users. Therefore, these links become less informative for community detection. Metadata such as tags become an important source in measuring the user-user similarity. We show that more accurate community structures can be obtained by scrutinizing tag information.
- We obtain clusters containing users and tags simultaneously. The clusters explicitly show who is interested in what, which is helpful in understanding the groups. Existing co-clustering methods cluster users/tags separately. Thus, it is not clear which user cluster corresponds to which tag cluster. But our proposed method is able to find out user/tag group structure and their correspondence.

The rest of this paper is organized as follows. Section II summarizes contemporary techniques in community detection and co-clustering. Section III defines the problem formally. A framework is presented in Section IV, followed by experimental evaluation in Section V and VI. Our work and possible future directions are summarized in Section VII.

II. Related Work

Online social networks are recognized as complex networks which are characterized by high clustering coefficient and short average distance [4]. A high clustering coefficient suggests a strong community structure in social networks. But community structure is not always explicitly available which makes community detection [22] an important component in social network analysis.

Most work in community detection attempt to discover non-overlapping communities based on different measures, objectives and statistical inference [5]. Methods based on graph partitioning is used to divide users into disjoint subgraphs such that the number of edges lying between different communities are minimized. However, the graph partition problem is usually NP-hard which is relaxed to spectral clustering [6]. Newman and Girvan [7] proposed modularity to measure the strength of community structure. Modularity of a community is defined by the number of edges within the community subtracted by the expected number of edges in this community. High modularity implies that the nodes are closely connected. Maximizing the modularity is also proven to be NP-hard and a relaxation to spectral clustering is proposed [8]. Random walk can be effective in community detection in social and biological networks [9], [10]. The basic idea is that the random walker has a higher likelihood to stay within the highly connected communities than move to another community.

Online social networks are made of highly overlapping cohesive communities. Overlapping community detection, which allows one user to be associated in several communities, attracts more attention recently. There are two

different versions of overlapping community representation. Fuzzy clustering or soft clustering is one of the ways in which each node will be assigned a membership score to a community. The probability represents the membership dedicated to a community. Yu et al. [11] propose a graph factorization framework, which approximates the original graph by constructing a node-community bipartite graph, in which each link between a node and a community represents the membership (probability) of this node to the community. Bayes inference, usually requires some observed patterns of connections between users, and builds a statistical model with a set of parameters, then these parameters are estimated by maximizing posterior [5]. Newman et al. [12] model the probabilities from users to groups via expectation-maximization in directed graphs.

The other way of overlapping community detection is discrete assignment. CFinder [13] first enumerates all k-cliques and combines them if there is a high overlapping (e.g., they share k-1 nodes) between two cliques. Cliques are fully connected sub-graphs and a node may belong to several cliques. This method can discover overlapping communities, but it is computationally expensive. EdgeCluster [3] views the graph in an edge-centric angle, i.e., edges are treated as instances and nodes are treated as features. It also shows that a user is usually involved in multiple affiliations, but an edge is usually only related to a specific group. Thus, they propose to cluster edges instead of nodes in social media. This discrete assignment of nodes in a graph gives a clear definition on the community of nodes. Evans et al. [14] proposes to partition links of a line graph to uncover the overlapping community structure. A line graph can be constructed from the original graph, i.e., each vertex in the line graph corresponds to an edge in the original graph and the links in the line graph represents the adjacency between two edges in the original graph, for instance, two vertices in line graph are connected if the corresponding edges in the original graph share a vertex. But it is difficult to scale up to large data sets because of memory requirement.

Co-Clustering is the process to cluster instances as well as their features at the same time. Dhillon et al. [1] propose to co-cluster documents and terms. At first, a bipartite graph between documents and terms is constructed, but partitioning documents and words in this graph is NP-hard, thus it is relaxed to a spectral co-clustering problem. Then top singular vectors (except the principle singular vector) of the document-word bipartite graph are clustered by k-means algorithm. The work above does not take the document-document correlation into account. Java et al. [15] advance this method by adding link structures between entities. For example, links between academic papers in terms of citation are added to the paper-word bipartite graph. The basic idea of Zha et al. [2] is close to Dhillon's work. The bipartite graph partition problem is solved by computing a partial singular vector decomposition (SVD) of the weight matrix. Furthermore, Zha et al. also show that the normalized cut problem is connected to correspondence analysis in multivariate analysis. Similar to [1], this problem is also relaxed to spectral clustering, then k-means is run on the eigenvectors to discover clusters. Compared to [1], this method requires more memory and are computationally more expensive. Information-theoretic co-clustering [16] tries to maximize mutual information between document clusters and term clusters.

III. Problem Statement

In social media, a community is a group of people who are more "similar" with people within the group than people outside this group. Homophily is one of the important reasons that people connect with others [17], which can be observed everywhere: people who come from the same city talk more frequently, people have similar political viewpoints are more likely to vote for the same candidates, and people who watch the same movies because of the commonly liked movie stars. The homophily effect suggests that like-minded people have a higher likelihood to be together.

In social media websites such as BlogCatalog[3] and del.icio.us[4], users are allowed to register certain resources (e.g., bookmarks, blogs). For each resource, users are asked to provide a short description in terms of tags. These tags are not randomly picked. They summarize the main topic of each resource. In this paper, the concept of community is generalized to include both users and tags. Tags of a community imply the major concern of people within it.

Let $\mathcal{U} = (u_1, u_2, \ldots, u_m)$ denote the user set, $\mathcal{T} = (t_1, t_2, \ldots, t_n)$ the tag set. A community $C_i (1 \leq i \leq k)$ is a subset of users and tags, where k is the number of communities. As mentioned above, communities usually overlap, i.e., $C_i \bigcap C_j \neq \varnothing$ (1 $\leq$ i, j $\leq$ k). On the other hand, users and their subscribed tags form a user-tag matrix M, in which each entry $M_{ij} \in \{0, 1\}$ indicates whether user u_i subscribes to tag t_j. So it is reasonable to view a user as a sparse vector of tags, and each tag as a sparse vector of users.

Given notations above, the overlapping co-clustering problem can be stated formally as follows:

Input:

- A user-tag subscription matrix $M_{N_u \times N_t}$, where N_u and N_t are the numbers of users and tags, respectively;
- The number of communities k.

Output:

- k overlapping communities which consist of both users and tags.

[3]http://www.blogcatalog.com/
[4]http://delicious.com/

IV. The Co-Clustering Framework

The observation that a user is usually involved in several affiliations but *a link is usually related to one community* enlightens us to cluster edges instead of nodes. After obtaining edge clusters, communities can be recovered by replacing each edge with its two vertices, i.e., a node is involved in a community as long as any of its connection is in the community. Then the obtained communities are often highly overlapped. This idea is similar to cluster in line graphs [14], but constructing line graph requires large amount of memory.

In a user-tag network, each edge is associated with a user vertex u_i and a tag vertex t_p. If we take an edge-centric view by treating each edge as an instance, and two vertices as features, each edge is a sparse vector. The length of vector is N_u+N_t, in which the first N_u entries correspond to users, and the other N_t entries correspond to tags. For example, the edge between u_1 and t_1 in Figure 1 can be represented as $(1,0,0,0,0,1,0,0,0)$, in which only entries for vertices u_1 and t_1 are non-zero.

Communities that aggregate similar users and tags together can be detected by maximizing intra-cluster similarity, which is shown in Eq. (2).

$$\arg\max_{C} \frac{1}{k} \sum_{i=1}^{k} \sum_{x_j \in C_i} S_c(x_j, c_i) \tag{2}$$

where k is the number of communities, C = $\{C_1, C_2, \ldots, C_k\}$, x_j represents an edge, and c_i is the centroid of community C_i. This formulation can be solved by using k-means. However, k-means is not efficient for large scale data sets. We propose to use EdgeCluster which is a k-means variant and is a scalable algorithm to extract communities for sparse social networks [3]. It treats the network in an edge-centric view. It is efficient because each centroid only compares to a small set of edges that are correlated to the centroid. It is reported to be able to cluster a sparse network with more than 1 million nodes into thousands of clusters in tens of minutes. The clustering quality is comparable to modularity maximization but the time and space reduction is significant. It should be noted that the network in [3] is 1-mode, but the user-tag network is 2-mode.

The expected density of the user-tag network is shown in Eq. (3), which guarantees an efficient solution by applying EdgeCluster (The proof is omitted due to space limitation).

$$\text{density} \approx \frac{\gamma-1}{2-\gamma} \cdot (d^{2-\gamma}-1) \cdot \frac{1}{N_u} \tag{3}$$

where d is the maximum tag degree, N_u is the number of users in this graph and γ is the exponent of the power law distribution, which usually falls between 2 and 3 in social networks [20]. The maximum degree d is usually large in a power law distribution. Thus, the density is approximately inverse to the number of users.

A key step in clustering edges is to define edge similarity (centroids can be viewed as edges as well). Given two edges $e(u_i, t_p)$ and $e'(u_j, t_q)$ in a user-tag graph, the similarity between them can be defined in Eq. (4):

$$S_e(e, e') = \alpha S_u(u_i, u_j) + (1-\alpha) S_t(t_p, t_q) \tag{4}$$

where $S_u(u_i, u_j)$ is the similarity between two users, and $S_t(t_p, t_q)$ is the similarity between two tags. This is reasonable because the edge similarity should be dependent on both user and tag similarity. And parameter α ($0 \leq \alpha \leq 1$) controls the weights of users and tags. Considering the balance between user similarity and tag similarity, α is set to 0.5 in our experiments.

In the following sections, we show that our framework can cover different similarity schemes.

A. Independent Learning

Independence assumption is a popular way to simplify the problem we want to solve. If two tags are different, their similarity can be defined as 0, and 1 if they are the same. Thus the similarity can be represented by an indicator function which can be shown by Eq. (5).

$$\delta(m, n) = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases} \tag{5}$$

The user-user similarity is also defined in a similar way. Cosine similarity is widely used in measuring the similarity between two vectors. Given two edges $e(u_i, t_p)$ and $e'(u_j, t_q)$, their cosine similarity can be rewritten in Eq. (6).

$$S_e(e, e') = \frac{1}{2} \left(\delta(u_i, u_j) + \delta(t_p, t_q) \right) \tag{6}$$

Following Eq. (4), we can define the similarity between two edges as in Eq. (6), which is essentially the cosine similarity between two edges.

B. Normalized Learning

In online social networks, the tag usage behavior differs one user to another. For example the tag usage distribution follows a power law: some tags are shared by a small group of people, which might suggest a higher likelihood that they form a community. On the other hand, popular tags may not be discriminative in inferring group structures. Thus there is a need to differentiate the importance of different users and tags.

Let d_{u_i} denote the degree of the user u_i, and d_{t_p} represent the degree of tag t_p in a user-tag network. After applying normalization, edge $e(u_i, t_p)$ can be represented by $(0, \ldots, 0, \frac{1}{d_{u_i}}, 0, \ldots, 0, \frac{1}{d_{t_p}}, 0, \ldots, 0)$. Given two edges $e(u_i, t_p)$ and $e(u_j, t_q)$, the cosine similarity after normalization between them can be written in Eq. (7).

$$S_e(e, e') = \frac{d_{t_p} d_{t_q} \delta(u_i, u_j) + d_{u_i} d_{u_j} \delta(t_p, t_q)}{\sqrt{d_{u_i}^2 + d_{t_p}^2} \sqrt{d_{u_j}^2 + d_{t_q}^2}} \tag{7}$$

Setting α to 0.5, $S_u(u_i, u_j)$ and $S_t(t_p, t_q)$ given by Eq. (8), we can derive Eq. (7) from Eq. (4). Thus normalized edge similarity is consistent with the proposed framework.

$$\begin{aligned} S_u(u_i, u_j) &= \frac{2d_{t_p}d_{t_q}\delta(u_i, u_j)}{\sqrt{d_{u_i}^2 + d_{t_p}^2}\sqrt{d_{u_j}^2 + d_{t_q}^2}} \\ S_t(t_p, t_q) &= \frac{2d_{u_i}d_{u_j}\delta(t_p, t_q)}{\sqrt{d_{u_i}^2 + d_{t_p}^2}\sqrt{d_{u_j}^2 + d_{t_q}^2}} \end{aligned} \quad (8)$$

It is noticed that the similarity between two users is not only related to users, but also the tags they are associated with. Eq. (6) and Eq. (7) both assume tags (users) are independent, which is not true in real applications. We next propose a similarity measurement based on correlation.

C. Correlational Learning

Users often use more than one tag to describe the main topic of a bookmark. Grouped tags indicate their correlation. For instance, the tags *car information*, *auto info* and *online cars info*, are used to describe a blog[5] registered on BlogCatalog, are different, but semantically close.

In a user-tag network, a user can be viewed as a vector by treating tags as features. On the other hand, a tag can also be viewed as a vector by treating users as features. Representing users in a latent semantic space captures the correlation between tags, for example, mapping several semantically close tags to a common latent dimension. Let $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m$ be the orthogonal basis of a latent semantic sub-space for tags, user vectors in the original space can be mapped to new vectors in the latent space, which is shown in Eq. 9.

$$\tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m) = \mathcal{M}(u_i(t_1, t_2, \ldots, t_n)) \quad (9)$$

where $\mathcal{M}$ is a linear mapping from the original space to the latent sub-space. Singular Value Decomposition (SVD) is one of the ways to obtain the set of orthogonal basis. The singular value decomposition of user-tag network M is given by $M = U\Sigma V^T$, where columns of U and V are the left and right singular vectors and Σ is the diagonal matrix whose elements are singular values. User vectors in the latent space can be formulated in Eq. (10).

$$\begin{aligned} & u_i(t_1, t_2, \ldots, t_n) = \{U\Sigma\}_i V^T \\ \Leftrightarrow \quad & u_i(t_1, t_2, \ldots, t_n) = \tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m)V^T \\ \Leftrightarrow \quad & \tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m) = u_i(t_1, t_2, \ldots, t_n)V \end{aligned} \quad (10)$$

where $u_i(t_1, t_2, \ldots, t_n)$ and $\tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m)$ are the user vectors in the original and latent space, respectively.

However, only a small set of right singular vectors $V' = (v_2, v_3, \ldots, v_m)$ are necessary to be computed. Dhillon [1] suggests that it be $\lceil \log_2 k \rceil + 1$. Recent experimental evaluation in text corpus suggests the dimension between 50 and 1,000 depending on the corpus size and the problem being studied [18]. Another reason of taking a relatively small m is to reduce noise in the data. The user vectors in the latent space can be represented by pluging V' into Eq. (10). We set m to 10 for synthetic data sets and to 300 for social media data sets. The user similarity and tag similarity are then defined by the corresponding vectors in the latent space.

$$\begin{aligned} S_u(u_i, u_j) &= \frac{\tilde{u_i} \cdot \tilde{u_j}}{\|\tilde{u_i}\| \, \|\tilde{u_j}\|} \\ S_t(t_i, t_j) &= \frac{\tilde{t_i} \cdot \tilde{t_j}}{\|\tilde{t_i}\| \, \|\tilde{t_j}\|} \end{aligned} \quad (11)$$

This can be interpreted from the graph partition point of view. Graph partition based on ratio-cut or normalized-cut can be relaxed to spectral clustering problem [6].

$$Lz = \lambda W z \quad (12)$$

where z solves the generalized eigenvectors of above equation, L is the laplacian matrix and W is the adjacency matrix, their definitions are shown in Eq. (13) in which D_1 and D_2 are diagonal matrix whose non-zero entries are user degrees and tag degrees, respectively.

$$\begin{aligned} L &= \begin{bmatrix} D_1 & -M \\ -M^T & D_2 \end{bmatrix} \\ W &= \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix} \end{aligned} \quad (13)$$

Let $Z = \begin{bmatrix} U \\ V \end{bmatrix}$ denote the eigenvectors of Eq. (12). The generalized eigenvector problem can be rewritten by:

$$\begin{bmatrix} D_1 & -M \\ -M^T & D_2 \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} = \lambda \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} \quad (14)$$

After simple algebraic manipulation, we obtain

$$\begin{aligned} M &= (1-\lambda)V^T D_1 U \\ M^T &= (1-\lambda)U^T D_2 V \end{aligned} \quad (15)$$

Thus eigenvectors Z are actually the right and left singular vectors of adjacency matrix M. Thus top singular vectors (except the principle singular vector) of the adjacency matrices contain partition information [1], [2], [6]. Since the user-tag graph studied in this paper is connected, the principle singular vector is discarded.

V. Synthetic Data and Findings

Clustering evaluation is difficult when there is no ground truth. Synthetic data, which is controlled by various parameters, facilitates a comparative study between the uncovered and actual clusters. We first introduce the synthetic data and how they are generated, then the clustering quality measurement Normalized Mutual Information (NMI). Finally, the NMI of different clustering methods are reported.

[5] http://www.blogcatalog.com/blogs/online-cars-info-auto-info-car-news.html

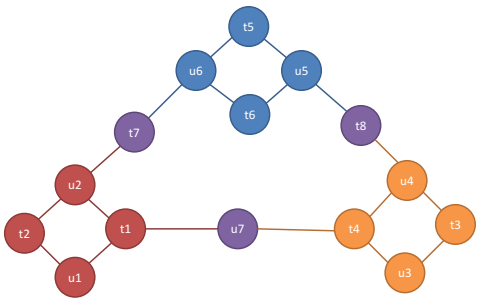

Figure 2. A (toy) synthetic graph with three clusters

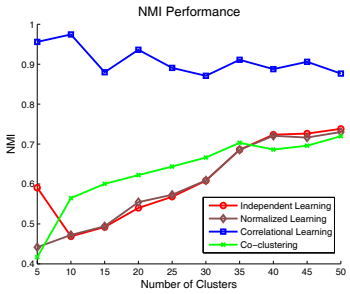

Figure 3. NMI Performance w.r.t Number of Clusters

A. Synthetic Data Generation

We develop a synthetic data generator that allows input of the numbers of clusters, users and tags. First users and tags are split evenly into each cluster. Then, in each cluster users and tags are randomly connected with a specified density (e.g., 0.8). Links between clusters which account for 1% of the total number of links are randomly assigned to two users or two tags belonging to distinct clusters. For such between cluster links, additional user nodes or tag nodes are added such that users are connected to tags and tags are connected to users. Figure 2, shows a toy example of the synthetic user-tag graph in which users are labeled as $u1-u7$ and tags $t1-t8$. Three overlapping clusters are highlighted with different colors. Nodes labeled as t7, t8 and u7 are shared by two of the clusters. As shown in the toy example, links within clusters are dense, and links between clusters are sparse, thus the link structure implies a separation of clusters which will be served as an *approximate* ground truth.

B. NMI Evaluation in Synthetic Data

The advantage of a synthetic study is that the ground truth is under control. Thus, it is possible to measure the clustering performance by comparing with the ground truth. The Normalized Mutual Information (NMI) is commonly used to measure the clustering quality. Since we are studying overlapping clustering, the NMI definition given by Lancichinetti et al. [19] will be used in the following evaluations. It is an extension of NMI for non-overlapping clustering. Given two clusterings X and Y, the NMI is defined below.

$$
\begin{aligned}
NMI(X,Y) &= 1 - \frac{1}{2}\left(H(X|Y)_{norm} + H(Y|X)_{norm}\right) \\
H(X|Y)_{norm} &= \frac{1}{|C_X|}\sum_k \frac{\min_{l\in\{1,2,\ldots,|C_Y|\}} H(X_k|Y_l)}{H(X_k)} \\
H(Y|X)_{norm} &= \frac{1}{|C_Y|}\sum_k \frac{\min_{l\in\{1,2,\ldots,|C_X|\}} H(Y_k|X_l)}{H(Y_k)}
\end{aligned} \tag{16}
$$

where $H(X|Y)$ and $H(Y|X)$ are conditional entropy, $|C_X|$ and $|C_Y|$ are the number of clusters in X and Y, respectively. The NMI is computed in two steps. First, find the pairs of clusters that are most close to each other in two clusterings. Second, average the mutual information between those pairs of clusters. The higher the NMI value is, the more similar between two clusterings. If two clusterings X and Y are exactly the same, the NMI value is 1.

C. NMI and Number of Clusters

We generate another data set with 1,000 users and 1,000 tags and with different number of clusters which range from 5 to 50 and cluster density is set to 1 such that all users connect to all tags within each cluster. The latent dimension m is set to 20 in the synthetic evaluations. Since our proposed algorithms are basically k-means variants, we run our methods 100 times and report the averaged NMI. In each run, we set the same seed for Independent Learning, Normalized Learning and Correlational Learning. Dhillon's co-clustering method is also included for the comparative study. The results are summarized in Figure 3.

We can see that the method considering tag correlation performs much better than the other two. This indicates that correlation helps to aggregate users and tags that are semantically close. It is interesting to note that the Normalized Learning is inferior to the counterpart without normalization. Co-clustering fails to uncover overlapping structure, and has a similar performance as that of Independent Learning.

D. NMI and Link Density

We also study how intra-cluster link density affects clustering in synthetic data sets. We created synthetic data sets (50 clusters, 1,000 users and 1,000 tags) with different intra-cluster densities that range from 0.1 to 1. The data set is sparse when the link density is low and users and tags are fully connected when the link density is 1. The NMI results for different methods are shown in Figure 4. When the intra-cluster link density is greater than or equal to 0.2, the averaged NMI for correlational learning is above 0.8 which suggests the overlapping structures are well recovered. A high NMI value suggests the robustness of the proposed framework to work well even when the intra-cluster link density is low. Interestingly, co-clustering does not work well when the link density is low, e.g., NMI values are below 0.3 when the intra-cluster density is smaller than 0.5.

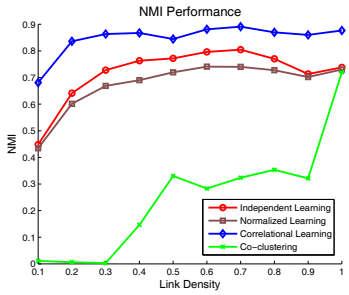

Figure 4. NMI Performance w.r.t Intra-cluster Link Density

In summary, Correlational Learning is more effective than the other two methods in recovering overlapping clusters in terms of NMI. It works well even when the intra-cluster link density is low. Co-clustering performs poorly because it only finds non-overlapping clusters.

VI. Social Media Data and Findings

BlogCatalog is a social blog directory where the bloggers can register their blogs under predefined categories. We crawled user names, user ids, their friends, blogs, the associated tags and blog categories. For each blog, users are allowed to specify several tags as a short description. These tags are usually correlated with each other. We crawled more than 10,000 users. Users who have no tags are removed from the data set, and tags that were used by less than two persons were removed as well. Finally, we obtained a data set with 8,797 users and 7,418 tags.

Delicious is a social bookmarking website, which allows users to tag, manage, and share online resources (e.g., articles). For each resource, users are asked to provide several tags to summarize its main topic. We crawled 11,285 users whose information include user name, user id, their friends and fans, their subscribed resources and tags for each resource. The top 10 most frequent tags of each person are kept, which is 13,592 in total. In contrast to BlogCatalog, two kinds of links are formed in Delicious. Fans are the connections from other people (in-links) and friends are the links point to others (out-links). Thus, the connections are directional in Delicious.

The statistics of both data sets are summarized in Table I. The most important difference between the two data sets is that BlogCatalog has category information which can be served as a ground truth for clustering distribution.

A. Interplay between Link Connection and Tag Sharing

There exist explicit and implicit relations between users. Examples of explicit relations are friends or fans people choose to be. Examples of implicit relations are tag sharing, i.e., people who use the same tags. Are there any correlation between the two different relations? What drives people connect to others? Is it a random operation? We conducted statistical analysis between user-user links and tag sharing.

Table I
Statistics of BlogCatalog and Delicious

	BlogCatalog	Delicious
# of users	8,797	11,285
# of unique tags	7,418	13,592
# of links	69,045	112,850
density	1.1×10^{-3}	7.3×10^{-4}
maximum tag usage	165	10
minimum tag usage	1	10
average tag usage	7.8	10

In the first study, we fix users who have or have no connection with others, then show the tag sharing probabilities. Figure 5 shows the tag sharing probabilities in BlogCatalog and Delicious data sets. For Delicious data, the friends network and fans network are evaluated separately. All three graphs show a similar pattern that the tag sharing probability is higher among users who are connected than users who are not. This can be explained by the homophily principle that people tend to connect with those who are like-minded.

Figures 6 and 7 are the probability that two users being connected if they share tags in BlogCatalog and Delicious, respectively. In Figure 6, the probability of a link between two users increases with respect to the number of tags they share. In Delicious, similar pattern is observed. It is also intriguing to show the probability that two users are connected is higher in fans network than that in friends network, which implies users are more *similar* to their fans than their friends.

B. Clustering Evaluation

The clustering evaluation consists of three studies. First, cross-validation is performed to demonstrate the effectiveness of different clustering algorithms in BlogCatalog data set. Then we study the correlation between user connectivity and co-occurrence in extracted communities. Finally, concrete examples illustrate what clusters are about.

1) Comparative Study: In BlogCatalog, categories for each blog are selected by the blog owner from a predefined list. A category is treated as a community or group which suggests the common interest of people within the group. For example, category "Blog Resources" is related to the gadgets used to manage blogs or to communicate with other social media sites. Around 90% of bloggers had joined two categories, and few bloggers had more than 4 categories.

With category information, certain procedures such as cross validation (e.g., treating categories as class labels, cluster memberships as features) can be used to show the clustering quality. Linear SVM [21] is adopted in our experiments since it scales well to large data sets. As recommended by Tang et al. [3], 1,000 communities are used in our experiments. We vary the fraction of training data from 10% to 90% and use the rest as test data. The training data are randomly selected. This experiment is repeated for

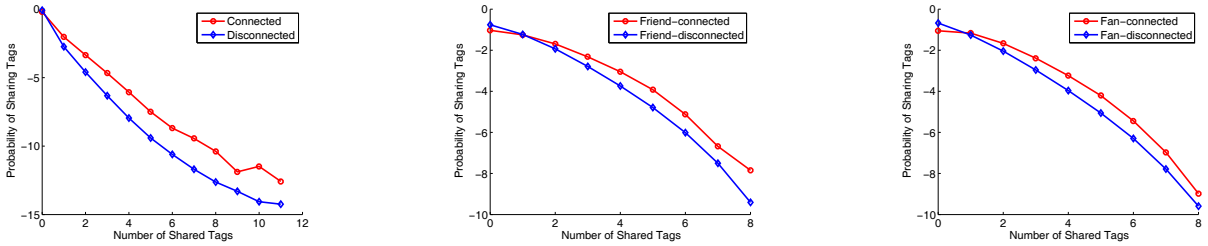

Figure 5. X-axis represents the number of tags that two users share. Y-axis in log plot is the probability that two users share tags. Left graph shows the tag sharing probability in BlogCatalog data set by fixing the users we want to study. Center and Right graphs are the corresponding probabilities in Delicious data set. Center graph is summarized in friends network and Right graph is in fans network. The red curves represent the probability that users are connected, and the blue curves represent there are no links between these users.

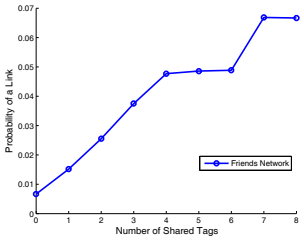

Figure 6. Link probability w.r.t tag sharing in BlogCatalog

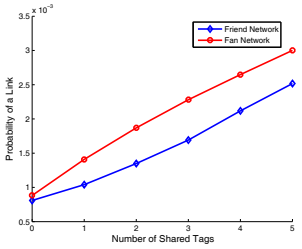

Figure 7. Link probability w.r.t tag sharing in Delicious

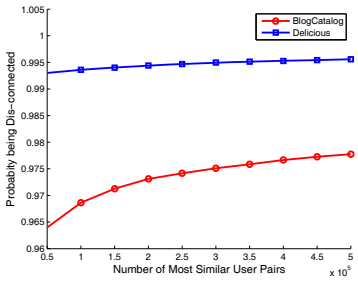

Figure 8. Probability being Dis-connected between Top Similar Users

10 times and the average Micro-F1 and Macro-F1 measures are reported.

Table II shows five different clustering methods and their prediction performance. In this table, the fourth algorithm EdgeCluster [3] uses user-user network rather than the user-tag network. Dhillon's co-clustering algorithm is based on Singular Value Decomposition (SVD) of the normalized user-tag matrix. As shown in Table II, Correlational Learning consistently performs better, especially when the training set is small. According to Table II, normalization does not improve performance. This suggests normalization should be taken cautiously. Dhillon's co-clustering method which can only deal with non-overlapping clustering does not perform well compared to other methods.

It is also interesting to notice that clustering based on user-tag is significantly better than user-user connection which suggests that meta data (e.g., tags) rather than connection is more accurate in measuring the homophily between users. The clustering difference between meta data and links also reveals promising applications of the framework in link prediction systems. Next, we try to interpret clustering results.

2) Connectivity Study: We study the correlation between user co-occurrence in extracted communities and the actual social connections between them. We also study the connectivity between users who are in the top similar list. 1,000 overlapping communities are extracted by Correlational Learning.

In Table III, first row represents the number of communities two users co-occur, and each entry in this table is the probability that two users have a connection established in actual social networks. The last column lists the probability if two users are connected randomly. Higher probability than randomness suggests that users within communities are similar to each other. As observed in Table III, frequent co-occurrence of users in different communities implies that they are more likely to be connected. Therefore, it is reasonable to state that higher co-occurrence frequency suggests that two users are more similar. Similar patterns are observed in the other two methods.

We compute pairwise cosine similarity between users (in the latent space) and sort them in descending order, then study the dis-connectivity between users who are most similar. Figure 8 shows that the probability of being disconnected is higher than 96% and 99% in BlogCatalog and Delicious, respectively, which means that the majority

Table II
CROSS VALIDATION PERFORMANCE IN BLOGCATALOG DATA SET

Proportion of Labeled Nodes		10%	20%	30%	40%	50%	60%	70%	80%	90%
Micro-F1(%)	Correlational Learning	**38.45**	**37.75**	**40.53**	**38.84**	**41.92**	**41.30**	**43.77**	**43.15**	**44.88**
	Independent Learning	33.96	36.15	35.07	34.72	35.36	37.32	42.12	41.83	43.09
	Normalized Learning	23.89	28.10	29.22	32.14	34.52	35.19	35.79	35.74	37.62
	EdgeCluster(user-user)	24.85	25.55	26.27	25.18	25.28	24.80	24.11	23.94	22.22
	Co-clustering	23.18	24.18	24.11	24.30	24.34	24.23	24.18	24.15	23.97
Macro-F1(%)	Correlational Learning	**28.85**	**26.83**	**27.68**	**28.52**	**28.18**	**29.69**	28.60	**30.16**	**29.96**
	Independent Learning	23.84	25.32	24.34	23.81	25.06	26.28	**29.05**	27.27	26.84
	Normalized Learning	14.76	17.61	16.85	18.78	21.66	21.80	22.07	22.39	24.20
	EdgeCluster(user-user)	14.24	15.16	16.43	15.75	15.96	16.08	15.42	15.78	14.99
	Co-clustering	4.95	5.06	5.11	5.19	5.07	5.18	5.17	5.23	4.66

Table III
CO-OCCURRENCE VS. CONNECTIVITY

# of Co-occurrence	1	2	3	4	5	Random
BlogCatalog($\times 10^{-2}$)	1.64	2.78	4.27	4.43	4.48	0.74
Delicious($\times 10^{-3}$)	2.52	3.83	3.94	3.97	3.45	0.35

of homogeneous users are not connected in actual social networks. For example, users *marama*[6] and *ameer157*[7] both are interested in the online game "World of Warcraft". Their tags highly overlap, but there is no connection between them. In online social networks, most users are scattered in the long tail, and are usually unreachable by following their and their friends' links. But it is possible to recommend links to connect them with our Correlational Learning.

3) Illustrative Examples: Health is the second largest category (the largest is *personal*) in BlogCatalog, a hot topic that attracts lots of cares. To visualize communities, we create tag clouds using Wordle[8]. In a tag cloud, size of a tag is representative of its frequency or importance in a set of tags or phrases. Figure 9 shows the tag cloud for Category Health (category-health) including all tags of this category. The most frequent 5 tags, *health*, *weight loss*, *diet*, *fitness* and *nutrition*, are all about health.

The largest cluster about Health obtained by Correlational Learning is cluster-health with 127 users and 102 tags. The cluster that has the maximum user overlapping with cluster-health is cluster-nutrition with 83 users and 25 tags. Their tag clouds are shown in Figures 10 and 11. Between the two clusters, there are 18 users and 3 tags *health*, *nutrition* and *weight loss* in common. Both clusters are related to health but the first has an emphasis on physical health, highlighted by tags *arthritis*, *drugs*, *food*, *dentist*, and the second is more about *nutrition*. We study the tag overlapping between category-health and cluster-health, and between category-health and cluster-nutrition. The top 102 tags of category-health are compared to the tags of cluster-health and the top 25 tags of category-health to those of cluster-nutrition. The numbers of shared tags are 16 for cluster-health and 9 for

[6]http://www.blogcatalog.com/user/marama
[7]http://www.blogcatalog.com/user/ameer157
[8]http://www.wordle.net/

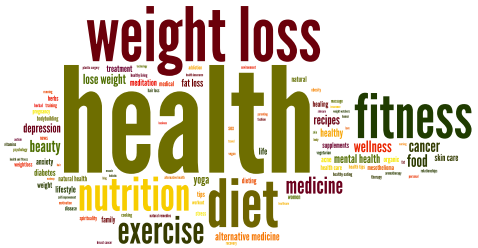

Figure 9. Tag cloud for category-health in BlogCatalog

Figure 10. Tag cloud for cluster-health in BlogCatalog

cluster-nutrition. The overlapping analysis indicates that tags of the two clusters differ (with only 3 tags in common), the tags of the two clusters are not the same as those of category-health, and each cluster represents a new concept (or a sub-topic of health) that is buried in the tags of category-health.

In addition, we aggregate tags of the users in cluster-health and present the most frequent 102 tags in Figure 12. Comparing these tags with those of cluster-health, 40 tags are in common. Many tags such as *environment*, *humor*, *jokes* are not present in the tag cloud of cluster-health, which suggests that these users actually have other interests besides health. A similar pattern is observed for cluster-nutrition. The proposed approach clusters users and tags simultaneously can find clusters with more semantically similar tags.

VII. CONCLUSIONS AND FUTURE WORK

Multiple interests and diverse interactions a person has in his real social life suggests that community structures in social media are often overlapping in nature. Rich metadata available in online social media provides new opportunity to discover communities by the content users produce. We

Figure 11. Tag cloud for cluster-nutrition in BlogCatalog

Figure 12. Tag cloud for users from cluster-health

proposed a framework to study the overlapping clustering of users and tags in online social media which helps to understand the major concerns within the groups. Experimental results in synthetic data reveal that Correlational Learning is very effective in recovering the overlapping cluster structures even when the inner cluster density is low. We reported several interesting findings in BlogCatalog and Delicious data sets. For instance, learning from the metadata is more accurate than the link information, people are more *similar* to their fans, and so on.

This study suggests more interesting problems that are worth further exploring. Formulating the co-clustering problem into an objective function and maximizing it is one direction to work on. With the large scale online social media data, the computational cost poses a serious challenge, which suggests that we develop more scalable algorithms to efficiently obtain co-clusters. Link prediction is another line of research in which the Correlational Learning framework can help.

VIII. Acknowledgments

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

References

[1] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *KDD '01*, NY, USA.

[2] H. Z. Xiaofeng, X. He, C. Ding, H. Simon, and M. Gu, "Bipartite graph partitioning and data clustering," in *CIKM'01*.

[3] L. Tang and H. Liu, "Scalable learning of collective behavior based on sparse social dimensions," in *CIKM'09*, NY, USA.

[4] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[5] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010.

[6] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[7] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, p. 026113, Feb 2004.

[8] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *SIAM'05*, April 2005.

[9] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. of Graph Alg. and App. bf*, vol. 10, pp. 284–293, 2004.

[10] H. Zhou, "Distance, dissimilarity index, and network community structure," *Phys. Rev. E*, vol. 67, p. 061901, 2003.

[11] K. Yu, S. Yu, and V. Tresp, "Soft clustering on graphs," in *NIPS*, p. 05, 2005.

[12] M. E. J. Newman and Leicht, "Mixture models and exploratory analysis in networks." *PNAS'07*, vol.104, p.9564.

[13] G. Palla, I. Dernyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature'05*, vol.435, no.7043, p.814.

[14] T. S. Evans and R. Lambiotte, "Line graphs, link partitions and overlapping communities," *Phy. Rev. E*, vol.80, p.016105,2009.

[15] A. Java, A. Joshi, and T. Finin, "Detecting commmunities via simultaneous clustering of graphs and folksonomies," in *WebKDD*, 2008.

[16] I. Dhillon, S. Mallela, and D. Modha, "Information-theoretic co-clustering," in *SIGKDD'03*. New York, NY, USA.

[17] L. S.-L. Miller McPherson and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[18] T. K. Landauer and S. T. Dumais, "Latent semantic analysis," *Scholarpedia*, vol. 3, no. 11, p. 4356, 2008.

[19] A. Lancichinetti, S. Fortunato, and J. Kertesz, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, Mar 2009.

[20] M. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46(5), p. 323–352, 2005.

[21] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[22] L. Tang and H. Liu, "Community Detection and Mining in Social Media," Morgan & Claypool Publishers, Synthesis Lectures on Data Mining and Knowledge Discovery, 2010.