Understanding Emerging Social Structures — A Group Profiling Approach

Lei Tang
Computer Science & Engineering
Arizona State University
Tempe, AZ 85287, USA
L.Tang@asu.edu

Xufei Wang
Computer Science & Engineering
Arizona State University
Tempe, AZ 85287, USA
Xufei.Wang@asu.edu

Huan Liu
Computer Science & Engineering
Arizona State University
Tempe, AZ 85287, USA
Huan.Liu@asu.edu

ABSTRACT

The prolific use of participating web and social networking sites is reshaping the way in which people interact with each other, and it has become an increasingly vital part of social life of human beings. People sharing certain similarities tend to form communities in social media. At the same time, they participate in various online activities: content sharing, tagging, twittering, etc. These diverse activities leave traces of their social life, providing clues to understand emerging social structures. Plenty of existing work focus on extracting cohesive groups based on network topology. In this work, we advance further to explore different group-profiling strategies to construct descriptions of a group, helping explain the group formation. This research and results can help network navigation, visualization and analysis, as well as monitoring and tracking the ebbs and tides of different groups in evolving networks. By exploiting the information collected from real-world social media sites, we conduct extensive experiments to evaluate group-profiling results. The pros and cons of different group-profiling strategies are analyzed with concrete examples. We then use LiveJournal as a testbed to show some potential applications based on group profiling. Interesting findings with discussions are reported.

Keywords

Group Profiling, Social Structures, Group Characterization, Group Formation, Social Media

1. INTRODUCTION

Recently, a surge of work has reported the statistical patterns presented in complex networks across many domains [23, 7]. The majority work study the global patterns presented in a static or an evolving network [15, 17]. Microscopic patterns such as the individual interaction patterns are also attracting increasing attentions [16]. This work, alternatively, focuses on meso-level analysis of a network as in Figure 1. In particular, we study groups (communities) in social media. Group-level analysis plays a key role in social science. Actually, "the founders of sociology claimed that the causes of social phenomena were to be found by studying groups rather than individuals"([13] Chapter 2, Page 15).

A *group* (or community) can be considered as a set of actors where each actor interacts with the other actors within the community more frequently than with those actors outside the community [31]. Finding out groups from network interactions has a broad range of applications, including network visualization, intelligence analysis [4], network compression [26], behavioral study [13], and collaborative filtering [9]. A variety of community detection (a.k.a. finding cohesive subgroups [31]) methods have been proposed to capture such social structures in a network. With the expanded use of Web and availability of large-scale social networks, community evolution in dynamic networks is gaining increasing attentions [14, 24, 26, 29].

Figure 1: Network Analysis at Different Levels

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 pioneering work attempt to understand the group formation based on statistical structural analysis. Backstrom et al. [3] studied prominent online groups in the digital domain, aiming at answering some basic questions about the evolution of groups, one of which is 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 actor 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 actors. Leskovec [18] observed that spectral clustering (a popular method used for community detection) always finds tight and small-scale but almost trivial communities, i.e., the community is connecting to the remaining network via one single edge. Both work above present a global (statistical) picture of communities. However, more efforts are required to understand the formation of a particular group.

In social media, people tend to interact with each other if they share certain similarity (also known as *homophily* [20]), resulting in assorted communities. There can be various reasons leading to the formation of a community. some users interact with each other because they attend the same university; some actors form a group as they are enrolled in an

event. Actors can also coalesce if they share certain political views. In this work, we attempt to understand a group from *a descriptive aspect*.

- Can we infer the latent causes of the group formation given individual attributes? Can we find out the shared group-level similarity?
- If so, what are the effective approaches?

We aim to extract group attributes that help understand a group. For the aforementioned examples, the group attributes, ideally, will be informative of the university, the event, and the political view, respectively.

Extracting descriptive attributes for a group of people is referred as *group profiling* [28]. Finding out a profile with attributes shared by individuals interpret the emergence of social structures as well as the fads in social media. To construct a group profile, we study strategies to extract attributes for a group when individual attributes are available. This is especially applicable in social media since individuals might share their profiles as well as user activities such as blog posts, status updates, comments, visited web pages, clicked ads and so on. This vast amount of individual traces pose a challenge to extract useful information to describe a group. We present three sensible methods: aggregation, differentiation, and egocentric differentiation. Another challenge is evaluation as it is subjective. It requires extensive human efforts to delve into group members' activities to figure out the shared similarity among them. Extensive experiments with concrete case studies on two social media domains demonstrate the effectiveness of our proposed (egocentric) differentiation based group profiling methods. We also enclose a discussion of potential applications based on group profiling, paving the way for in-depth network analysis at large as well as effective group search and retrieval.

2. GROUP PROFILING

Group profiling is the process of constructing descriptive profiles for a provided group. In this section, we motivate this task and formally define the problem.

2.1 Motivation

Social connections occur at a higher rate between similar people than dissimilar people. Homophily is one of the first characteristics studied by early social network researchers [2, 32, 6], and holds for a wide variety of relationships [20]. Homophily is also observed in social media [11, 30]. In this work, we study the "inverse" problem: given a group of actors, can we figure out why they got connected? What are their shared similarity?

It is impossible to answer these questions if no information other than a social network is available. Luckily, social media often provides rich information than just a network. In blogosphere, users post blogs and upload tags. For instance, in Facebook, users chat with each other, update their status, leave comments and share interesting stories. These different activities reflect users' online social life. These information can be used to answer the aforementioned questions.

Social media sites often come with a social network between users. For instance, in Twitter[1], there is a following-follower network. Some community detection methods can be applied to find out the *implicit groups* hidden beneath the interactions. Group profiling, in this case, can be used to understand the extracted communities, facilitating the network analysis.

In some other sites like LiveJournal[2], Flickr[3], YouTube[4], and Facebook[5], users are allowed to form *explicit communities*. In these sites, various explicit communities, besides the implicit groups, have cropped up. Some might suspect that for the explicit communities, the community name and description already provides enough information to peek into the group. Unfortunately, this is not necessarily true. In LiveJournal, one of the data sets we studied in the experiments, we encountered a large number of communities whose profile page provides little information of the group. For instance, the community profile of *fruits*[6] does not say much about the exact topic of the community. Group name might provide some hints, but can be misleading in certain cases. Take *fruits* as an example again. A first glimpse at the community name leads us to think that this community is composed of people who are fond of fruits. However, after we conduct group profiling[7] on this community, we obtain the following top-ranking tags for this group:

> *fruits, japan, hello kitty, sanrio lolita, fashion, Japanese street fashion.*

Except the first tag that coincides with the group name, all the other tags indicate this group is more about Japanese fashion. Though this group starts with *fruits*, some character in animes and mangas like *hello kitty*[8] are often discussed as well. Actually, hello kitty is a very popular character used in Japanese fashion.

Hence, no matter implicit communities extracted based on network topology or explicit communities formed by user subscription, group profiling can be of help. Besides understanding social structures, group profiling also helps for network visualization and navigation. With dynamic attributes (say, recent blog posts, status updates or tweets of a user), group profiling can also help to track the topic shift of a group. It has strong 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 [15]: singletons who do not interact with others, isolated communities, and a giant connected component. Isolated communities actually occupy a very stable portion of the entire network, and the likelihood of 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.

[1] http://twitter.com/

[2] http://www.livejournal.com/community/

[3] http://www.flickr.com/groups/

[4] http://www.youtube.com/groups_main

[5] http://www.facebook.com/

[6] http://community.livejournal.com/fruits/profile

[7] More details in later parts.

[8] http://www.sanrio.com/

2.2 Problem Statement

Basically, we want to build a group profile, which can help explain what a group shares about. We seek to select a list of attributes to describe a group. This *group profiling* problem can be stated formally as follows:

Given:

- A social network $G = (V, E)$ where V is the vertex (actor) set, and E is the edge set;
- A particular group $g = (V_g, E_g)$ where $V_g \subseteq V$, and $E_g \subseteq E$;
- Individual attributes $A \in \{0, 1\}^{n \times d}$ where n is the number of nodes in the network G, and d is the total number of attributes;
- The number of group attributes to pick k.

Output:

- A list of top-k descriptive attributes of group g.

Note that here we assume the attributes of individual actors are boolean. For instance, the attributes can denote whether a word occurs in an actor's status update, blog post or uploaded tags, whether he supports abortion, etc. In some real-world applications, the individual attributes might be categorical rather than boolean, e.g., a user's favorite color, location, age, etc. For these kind of attributes, we can convert them into multiple boolean features. For example, if the color attribute contains three values $\{red, yellow, green\}$, we can convert it into three boolean features A_{red}, A_{yellow}, and A_{green}. So $A_{red} = 1$ means the user likes *red*. Thereafter, we just focus on boolean attributes. For convenience, we say a node has attribute A_i if $A_i = 1$ for the node.

It is desirable if the group profiling method satisfies the following properties:

- Descriptive. The selected attributes for a group should reflect the foundation of a group and the shared interest or the associated affiliation.
- Robust. Oceans of data is produced each day in social media. These data tend to be very noisy. The group profiling method should be robust to noise.

Following the guidelines above, we now elucidate several strategies for group profiling.

3. PROFILING STRATEGIES

In this section, we present several methods for group profiling. For group profiling, we assume the groups are provided. They can be either implicit groups extracted from networks according to certain community detection methods or explicit groups of user subscriptions. Before we delve into the details, we introduce some notations for presentation convenience.

Suppose there are n nodes in a social network G, and d attributes $\{A_1, A_2, \cdots, A_d\}$. For a specified group g, we are interested in the most descriptive features to explain the group formation. We can treat the group as the positive class (denoted as "+") and some other nodes not belonging to the group as the negative class (denoted as "−"). The instances (nodes) of positive (negative) class are called positive (negative) instances, respectively.

Given a feature A, we have the following statistics as summarized in Table 1:

- true positive (tp) is the number of positive instances containing feature A.
- true negative (tn) is the number of negative instances not containing feature A.
- false positive (fp) is the number of negative instances containing feature A.
- false negative (fn) is the number of positive instances not containing feature A.

Table 1: Statistics based on group and attribute

group	+	−
$A = 1$	tp	fp
$A = 0$	fn	tn

Given these statistics above, we can compute the conditional probability of an attribute occurring in a group as follows:

- true positive rate (tpr) is the conditional probability of a feature occurring in a group. In particular,

$$tpr = P(A|+) = \frac{tp}{tp + fn} \tag{1}$$

- false positive rate (fpr) is the conditional probability that a feature associated with the nodes that are not of the group. Specifically,

$$fpr = P(A|-) = \frac{fp}{fp + tn}. \tag{2}$$

We now present the methods for group profiling (GP).

3.1 Aggregation (AGP)

Since group profiling aims to find features that are shared by the whole group, a natural and straightforward approach is to find attributes that are most likely to occur within the group, which boils down to the following problem:

$$\max_{\{A_i\}_{i=1}^k} \sum_{i=1}^{k} P(A_i|+) \tag{3}$$

We can simply aggregate individual attributes in the group and pick the top-k most-frequent features in the group. Note that this frequency-based profiling is widely used in current tagging systems. If the whole network is considered a group, *tag clouds* with the tag font size denoting the popularity of a tag is essentially aggregation-based profiling.

However, this method can be sensitive to certain (dumb) features. For instance, words like "world", "good" and "2009" in blog posts or status updates can be very frequent. They do not contribute to characterizing a group. Even the crowd wisdom such as user shared tags may not help much following this aggregation strategy. Take one community named *photography*[9] in LiveJournal as an example. It is not difficult to figure out the shared interests among the group members. If we look at the top-frequent user interests associated with this group, we have the following list:

> *photography, art, music, movies, reading, writing, love, books, painting, poetry.*

[9]http://community.livejournal.com/photography/profile

Except the first two, other tags are actually not good group descriptors. This is because these tags are shared by a large number of people, thus in this group as well. Directly aggregating these tags is biased towards selecting popular tags, rather than those that are able to characterize this group.

3.2 Differentiation (DGP)

Instead of directly aggregating, we can select those which can help differentiate one group from others in the network. Hence, the group profiling problem amounts to that of feature selection [19] in a 2-class classification problem with the group being the positive class and the remaining nodes in the network as negative class. The goal is to find out those top-k *discriminative* features that are representative of a group. A difference of discriminative group profiling and feature selection is that we only care about features that are descriptive of a group (the positive class). Thus we enforce the following constraint for selected attributes:

$$tpr_{A_i} > fpr_{A_i} \tag{4}$$

In other words, feature A_i should better explain the positive class rather than the negative class.

Note that a particular group is typically fairly small compared with the whole network. For instance, the LiveJournal data set we collected has 16,444 users, the first two largest groups have around 5,000 and 1,500 members respectively. The majority (90.1%) of the groups are in the long tail, with less than 100 members. This results in a highly imbalanced class distribution [27]. With this skewed class distribution, Bi-normal separation (BNS) [12] is an effective method that outperforms other feature selection methods [12, 27]. The BNS score of an attribute is defined as

$$BNS = \left|F^{-1}(tpr) - F^{-1}(fpr)\right|, \tag{5}$$

where F^{-1} is the inverse cumulative probability function of a standard normal distribution.

Combining the BNS criterion in Eq. (5) and the constraint in Eq. (4), we have the following formulation for differentiation-based group profiling:

$$\begin{aligned} \max_{\{A_i\}_{i=1}^k} \quad & \sum_{i=1}^{k} \left|F^{-1}(tpr_{A_i}) - F^{-1}(fpr_{A_i})\right| \\ s.t. \quad & tpr_{A_i} \geq fpr_{A_i} \end{aligned}$$

Since F^{-1} is a monotonic increasing function, the objective can be reformulated as follows:

$$\max_{\{A_i\}_{i=1}^k} \quad \sum_{i=1}^{k} \left(F^{-1}(tpr_{A_i}) - F^{-1}(fpr_{A_i})\right) \tag{6}$$

3.3 Egocentric Differentiation (EDGP)

In the differentiation strategy introduced above, we deem all the nodes outside a group as belonging to negative class. However, it might be a luxury to have this global view of all the nodes in a network. Scalability can also be a concern. Most of the popular online social networks are very huge. For instance, Facebook has 250 million active users according to a recent report on July 16, 2009[10]. LiveJournal has more than 21 million accounts registered till August 10, 2009[11]. It's either time consuming or impractical to retrieve all the information of the networks. In some applications, only an egocentric view is available. In other words, we only know our friends but little knowledge about the people who are strangers to us. Is it possible to describe a group by its members and the members' network structure without knowing the global network topology?

Instead of differentiating the group from the whole network, we here differentiate the group from the neighbors of its members. Group neighbors refer to the nodes that are connected to at least one group member but not belonging to the group as shown in Figure 2. Egocentric differentiation follows a similar objective function as in Eq. (6). The key difference is that the egocentric approach treats only the group neighbors, instead of the whole network, as the negative class. Given the huge size difference of the negative classes between DGP and EDGP, one wonder if this egocentric approach suffices in finding discriminative features.

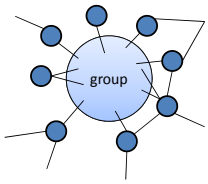

Figure 2: Group neighbors: all the shaded nodes are 1-hop away from the group.

4. EXPERIMENT SETUP

In this section, we present an evaluation strategy to compare different group profiling methods, the social media data sets and some basic properties with online groups.

4.1 Evaluation Methodology

Group profiling outputs a list of features that describe groups. The quality of the extracted profile depends on the group profiling method being used. There are several challenges to perform the comparison:

- How can we obtain the group information?
- What kind of individual attributes should we look into to extract group profiles?
- Given just a group of actors and the group description, the evaluator has to check individuals' attributes and find whether the description is consistent with the group commonality. This straightforward evaluation requires extensive human efforts, thus becomes inefficient or even impractical in huge networks.

Below, we address these concerns one by one.

- We use explicit communities in social media as the group information for experimental purposes. As we have mentioned in the introduction, in certain social media sites, users can subscribe to interested communities. These communities can serve the purpose of providing group information for our evaluation. Of course, implicit communities identified according to certain community detection method can also be used. But explicit communities come with their group names

[10]http://www.facebook.com/press/info.php?statistics
[11]http://www.livejournal.com/stats.bml

and sometimes descriptions, which can help human subjects to find out the ground truth for evaluation.

- In social media sites, users can share their profiles, upload tags, post blogs and update status. All these activities provide certain text information of the actor. Thus, we treat the words and tags occurring in user profiles or posts as attributes, and find out those key words to describe groups.

- For each group, we use the three proposed approaches (AGP, DGP, EDGP) to select top k (10 in our experiments) most representative features. Since there is no ground truth about what features should be used to describe the groups, we hire people with different background to evaluate the result. We created an easy-to-use website[12] for evaluators to log in and rate. To assure the quality of evaluation, each person is asked to evaluate 10 groups in one session.On each evaluation page, we provide the group title and the corresponding link such that the evaluators can get general group information before making decisions. The profile features for each method are listed in a column from top to bottom by importance in descending order. It should be emphasized that evaluators do not know what the methods are and which column is generated by which method. To avoid the bias associated with the column position, the presentation order of methods is also randomized for each view. Suppose for one group the three columns are generated by AGP, DGP, EDGP, respectively. The next time this group or another group is chosen, the three columns might correspond to methods in a totally different order.

  Each evaluator will rate for the resultant profiles on how well they are describing this group. The rating is ranged from 0 to 3. 0 means the features are irrelevant, 1 partly related, 2 fair and 3 very good. An evaluator can also choose "No Idea" if it is difficult for him/her to decide a proper rating.

4.2 Social Media Data Sets

As mentioned above, we need data sets with groups as well as rich individual attributes. Hence, we select two social media sites for data collection: BlogCatalog[13] and LiveJournal[14]. BlogCatalog is a social blog directory where the bloggers can register their blogs under the specified categories. LiveJournal is a virtual community where the Internet users can keep a blog, journal or diary. Both websites serve as a platform for users to connect and communicate with others. On both websites, users can initiate social activities like adding friends, joining groups, commenting and so on.

On BlogCatalog, we crawled blogger's name, friends, the blogs belonging to him/her, tags, categories and most recent 6 snippets. We treat blog categories as groups. After removing the non-English blogs, we obtained 70,086 bloggers and 344 groups. The total friendship links are 1,706,145, and each blogger has 49 friends on average. On LiveJournal, we started with a seed blogger, and crawled the bloggers that are reachable in 4 hops away from this seed. We collected blogger's name, friends, posts, interests specified in his/her profile and the communities the blogger subscribes to. Each user-created community is considered a group. Finally the data set has 16,444 bloggers, more than 130 thousand pairs of friendship links and 100,441 different communities.

[12]http://149.169.226.79

[13]http://www.blogcatalog.com/

[14]http://www.livejournal.com/

Table 2: Statistics on BlogCatalog and LiveJournal

	BlogCatalog	LiveJournal
No of Bloggers	70,086	16,444
No of Links	1,706,146	131,846
Link Density	6.9×10^{-4}	9.8×10^{-4}
Average Links	49	16
Diameter	5	8
Group Title	Category Name	Community Name
Group Numbers	344	100, 441
Average Groups Joined	1.9	32.6

The statistics of these two data sets are summarized in Table 2. One key difference between these two social media websites is that LiveJournal bloggers can create communities freely but BlogCatalog users can only specify categories from a predefined list. This explains why there is a much larger number of groups in LiveJournal.

These two sites demonstrate different statistical patterns. The group size distribution for both sites are plotted in Figures 3 and 4 respectively. In both figures, the x-axis represents the group size and the y-axis the frequency. Since the number of groups is very limited in BlogCatalog, we plot the distribution in histogram instead of scatter plot. The group size on LiveJournal follows a power law distribution. However the group size distribution on BlogCatalog is more like a bell curve, possibly because of the different mechanism for creating groups as we mentioned above. On the other hand, the number of groups one blogger joins is shown in Figures 5 and 6. In BlogCatalog, most bloggers join 2 groups, but a few bloggers (0.23%) join more than 3 groups. In LiveJournal, the distribution is different, about 82.3% bloggers have joined at least 4 groups. One blogger even has joined 1,032 groups. The average numbers of groups one blogger subscribing to are 1.9 and 32.6 on these two sites, respectively.

In the experiment, we would like to test group profiling methods with different noise level and study how each method performs. Typically, the words in posts are much more noisy than tags or user interests listed in users' profile pages. Hence, we created 4 data sets: BlogCatalog based on tags (BC-Tag) or posts (BC-post), and LiveJournal based on user interests (LJ-Interest) or posts (LJ-post). We expect LiveJournal to be more noisy than BlogCatalog as the communities are user generated rather than pre-specified.

Since the evaluation involves human efforts, it is impractical to evaluate exhaustively over all groups. We select a subset of representative groups with varying group sizes and densities as listed in Tables 3 and 4. Compared to the link density of the whole network, most of the time the members inside a group are more closely connected.

5. EXPERIMENT RESULTS

In this part, we will show the experiment results with concrete examples. 52 people with assorted background (undergraduate, graduate students,university faculty and employees) participated in our evaluation. The total number of group profiling ratings collected is 2,028, of which 101 ratings are "No Idea". So only the remaining 1,927 ratings are used in our analysis. On average, each group is evalu-

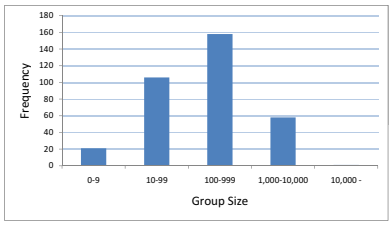

Figure 3: Group size distribution in BlogCatalog (in a bell curve). Few groups have less than 10 members, and only 1 group has a size greater than 10,000.

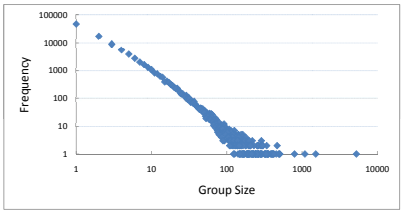

Figure 4: Group size on LiveJournal follows a power law distribution. Most of the groups have fewer than 100 members.

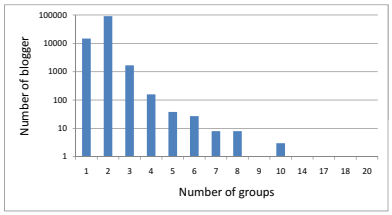

Figure 5: Group affiliation distribution of bloggers on BlogCatalog. Around 90% of the bloggers join 2 groups; the average groups one blogger joins is 1.9.

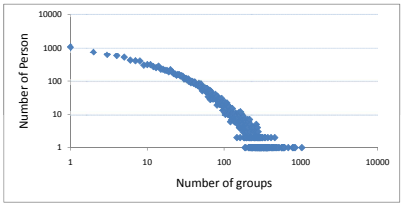

Figure 6: Group subscriptions distribution of bloggers on LiveJournal (in a power law distribution). On average, a blogger joins around 32 groups.

Table 3: Selected Groups on BlogCatalog.

Group	Size	Density	Group	Size	Density
Personal	11478	1.3×10^{-3}	Dogs	173	8.0×10^{-3}
Blogging	7727	2.7×10^{-3}	Adult Edu.	139	1.3×10^{-3}
Entert.	4671	1.9×10^{-3}	Buddhism	96	1.1×10^{-2}
Health	3877	2.4×10^{-3}	Hunting	86	4.1×10^{-2}
Shopping	2687	2.1×10^{-3}	Sailing	71	8.9×10^{-3}
Sports	2529	2.0×10^{-3}	Lawn&Garden	55	8.9×10^{-3}
Computers	1934	2.4×10^{-3}	Music Industry	47	6.1×10^{-3}
Animals	1357	5.6×10^{-3}	Natural	41	1.0×10^{-2}
Investing	906	3.8×10^{-3}	City Guides	40	3.2×10^{-2}
Science	826	2.4×10^{-3}	Anarchism	29	3.4×10^{-2}
Home Cook.	564	3.7×10^{-3}	Auto Repair	23	4.3×10^{-3}
Hardware	424	1.2×10^{-3}	Earth Science	22	1.6×10^{-2}
Pop	254	2.5×10^{-3}	Aquarium Fish	19	1.7×10^{-2}
Stock&Bond	245	7.1×10^{-3}	Choreography	13	2.6×10^{-2}
Cultural	229	4.5×10^{-3}	Extinct Birds	3	0

Table 4: Selected Groups on LiveJournal.

Group	Size	Density	Group	Size	Density
photography	320	1.3×10^{-2}	ontd_startrek	139	1.2×10^{-2}
sextips	297	1.8×10^{-3}	behind_the_lens	134	1.6×10^{-2}
mp3_share	288	2.1×10^{-3}	tvshare	132	5.2×10^{-3}
art_nude	232	3.3×10^{-2}	ru_portrait	131	7.6×10^{-2}
ourbedrooms	216	1.2×10^{-2}	knitting	124	2.3×10^{-3}
houseepisode	211	6.2×10^{-3}	girl_gamers	121	3.6×10^{-3}
fruits	205	1.6×10^{-2}	wow_ladies	115	2.0×10^{-3}
free_manga	205	9.1×10^{-3}	art_links	113	5.0×10^{-2}
ucdavis	189	3.9×10^{-2}	weddingplans	110	4.7×10^{-3}
photographie	188	1.2×10^{-2}	doctorwho_eps	109	2.5×10^{-2}
cooking	181	2.3×10^{-3}	ru_travel	108	2.0×10^{-2}
hot_fashion	161	2.5×10^{-2}	blythedoll	108	1.1×10^{-1}
naturalliving	157	3.8×10^{-3}	rural_ruin	105	1.4×10^{-2}
topmodel	155	2.8×10^{-3}	supernatural_tv	103	1.5×10^{-2}
photocontest	147	1.5×10^{-3}	animeicons	102	5.0×10^{-3}
cheaptrip	142	2.9×10^{-2}	gossipgirltv	101	8.1×10^{-3}

ated 32 times and the average ratings are reported in next subsection.

5.1 Comparative Study

The average ratings for each method on different data sets are shown in Table 5. On BC-Tag, three methods are comparable, whereas aggregation deteriorates when we use words in blog posts as features. A similar pattern is observed on LiveJournal, though the ratings drop dramatically. On both data sets, DGP and EDGP consistently outperform aggregation. This is most observable when the individual features are noisy (say, adopting blog posts as attributes).

This result is more visible in Figure 7, where we plot the chance of each group profiling method being the winner. The ratio is computed as the frequency of one method winning over the total number of evaluations. One method wins when it receives the highest rating among the three. It is noticed that ties often occur during evaluation. For example, if the ratings for AGP, DGP and EDGP are 2, 3, 3, then we consider both DGP and EDGP win. On BC-Tag, all three methods show similar chances. But on the other data sets, DGP and EDGP are consistently better than AGP, and the difference between the former and latter increases as the noise level multiplies (LiveJournal is more noisy than BlogCatalog as communities are not pre-specified, and posts are more noisy than tags or user-specified interests).

The performance of DGP and EDGP are pretty comparable, with the former slightly better. This demonstrates that little information is lost if we only compare a group with its adjacent neighbors, rather than with all users. With only an egocentric view, the computation cost of profiling a particular group can dramatically drop because of a much smaller number of involved bloggers. On BlogCatalog, the number of 1-hop away bloggers averaged on the selected groups is 8,274, or around 11.8% of the whole network. On LiveJournal, we only select the groups with a size greater than 50 due to the extremely large number of groups and around half of groups has only 1 member. The average number of 1-hop away bloggers of selected groups is 1,016, or around 6.2% of all the bloggers. This egocentric differentiation method is favorable in dynamic and evolving huge networks because updating the features is easy since only the local information is required instead of the whole network.

Table 5: Ratings averaged over all groups.

Data set	AGP	DGP	EDGP
BC-Tag	2.55	**2.62**	**2.62**
BC-Post	1.92	**2.35**	2.26
LJ-Interest	1.53	1.91	**2.00**
LJ-Post	0.54	**1.42**	1.35

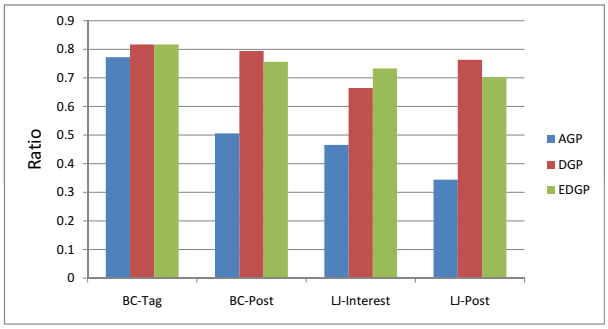

Figure 7: The ratio one method receives highest ratings. We treat all methods have highest ratings if all the three methods have the same rating score. The performance of AGP deteriorates as the noise level increases. DGP and EDGP are consistently better than AGP on all data sets.

5.2 Case Study

To have a tangible understanding of the outcome of different methods, here we show two concrete examples: *Health* group in BlogCatalog and *Blythedoll* group in LiveJournal. Health group has 2,607 members. Table 6 is the features extracted to describe group Health on BC-Tag as well as on BC-Post. The topics covered in this group are *medicine, diet, weight loss, men's and woman's health,* and so on. The features are sorted by importance in descending order. In BC-Tag data set, features extracted by all the three methods are related to Health. Only the order of some keywords are different. However, the result from AGP on BC-Post data set is not as descriptive as that on BC-Tag. Some features like *world*, *long*, *find*, *back* and *important*, are irrelevant to Health. By looking at the features generated by DGP and EDGP, it is not difficult to figure out that they are about Health. Also there is no big difference between these two.

Table 7 shows the features extracted from Blythedoll group on LiveJournal data set. Blythedoll was first created in 1972 by U.S. toy company Kenner, later it was spread out to the world. Takara, a Japanese company, is one of the most famous producers. On LJ-Interest, some of the features extracted by aggregation method are very common, e.g., photography, art and music, and we can hardly connect them to Blythedoll. However, on LJ-Post, the AGP result is almost no relation to Blythedoll group. The other two methods, DGP and EDGP, perform consistently better than simple aggregation. This example demonstrates the superiority of DGP and EDGP even if the data set is very noisy.

Table 6: Profiles for Health group in BlogCatalog. All methods work pretty good based on tags. Methods DGP and EDGP perform much better than AGP on posts. Two abbreviations: "mental h." for "mental health" and "impt." for "important".

BC-Tag			BC-Post		
AGP	DGP	EDGP	AGP	DGP	EDGP
health	health	health	people	health	health
fitness	fitness	fitness	health	people	people
diet	diet	diet	body	body	body
weight loss	weight loss	weight loss	life	life	weight
nutrition	nutrition	nutrition	world	weight	life
exercise	exercise	exercise	weight	disease	disease
beauty	cancer	cancer	long	diet	diet
medicine	medicine	medicine	find	food	treatment
cancer	beauty	mental h.	back	healthy	food
mental h.	mental h.	wellness	impt.	treatment	healthy

Table 7: Profiles for Blythedoll group in LiveJournal. AGP performs poorly on LJ-Post since all the features are not explicitly related to Blythedoll. DGP and EDGP are consistently better than AGP.

LJ-Interest			LJ-Post		
AGP	DGP	EDGP	AGP	DGP	EDGP
blythe	blythe	blythe	love	blythe	blythe
photography	dolls	dolls	back	doll	doll
sewing	sewing	sewing	ll	flickr	dolly
japan	japan	blythe dolls	people	ebay	dolls
dolls	blythe dolls	super dollfie	work	dolls	ebay
cats	super dollfie	japan	things	photos	sewing
art	hello kitty	hello kitty	thing	dolly	flickr
music	knitting	toys	feel	outfit	blythes
reading	toys	knitting	life	sell	outfit
fashion	junko mizuno	re-ment	pretty	vintage	dollies

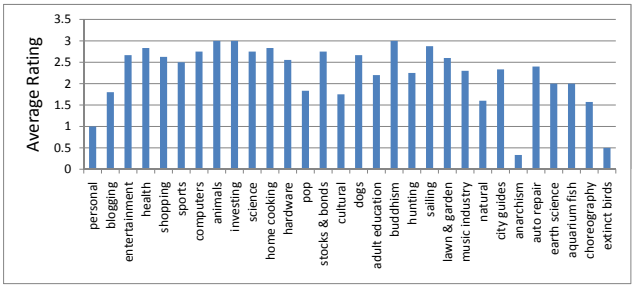

Figure 8: The rating of individual groups based on EDGP on BC-Post. The groups, from left to the right, are sorted by the group size. No significant correlation between ratings and group sizes is found. The generality (please see text) of the groups (topics) affects more to the ratings than the group sizes.

5.3 Further Analysis

We noticed that different groups receive quite distinctive ratings even for the same group profiling method. What might be the reason leading to this difference? Is there any connection between the group size and the ratings? Figure 8 plots the group sizes and ratings by EDGP on BC-Post. A conclusion we can draw is that there is no direct connection between the group sizes and average ratings. Large groups such as "personal" can receive low ratings, and small groups like "auto repair" can have high ratings. We observe similar patterns on other data sets with different methods (They are not included due to the space limitation).

One interesting finding is that the more general a group is, the relatively lower average rating it has. For instance, the largest group "personal" contains 11,478 members but has an average rating of 1, however, group "auto repair" only has 23 members and has an average rating of 2.4. This result is consistent with intuition: it is difficult to describe general concepts, but easy to describe a specific one.

Here, we further analyze the user evaluation behavior. We show the average ratings for each group in descending order

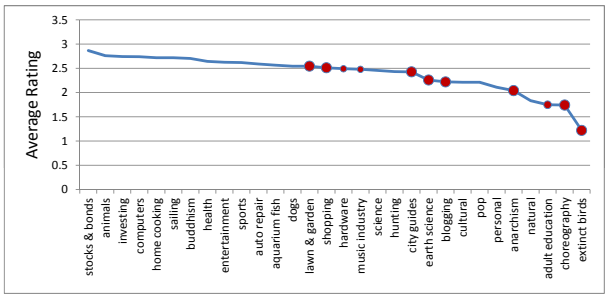

Figure 9: The average ratings on BlogCatalog. The red circles with varied size represent the relative probability of the group receiving "No Idea" ratings. The groups with "No Idea" ratings have a larger chance to show up at the tail.

in Figure 9 on BlogCatalog. The red circles in the curve highlight the groups receiving "No Idea" during evaluation, with their sizes indicating the relative probability. It is noticed that the markers tend to reside at the tail of the curve, i.e., when the rating for the group profile is relatively low. This basically shows that when it is difficult for human to judge what a particular group is about, it is not surprising the performance of group profiling decreases as well.

6. POTENTIAL APPLICATIONS OF GROUP PROFILING

Group profiling can help describe groups. Below we show two potential benefits of group profiling on social network sites and search engines.

6.1 Word Clouds

In current content sharing sites like del.icio.us and flickr, tag cloud has been used to describe the global trend of a social network. Some social network sites (e.g., Twitter, Facebook) do not provide tagging service but status updates. Instead of tag cloud, word cloud can be created from status messages, tweats. Quality word cloud on status updates helps users to understand the major trend as shown in a site in just a glimpse, considering that the status updates and tweats are typically much difficult to scrutinize than the summarization. It is also intriguing to compare the trend divergence demonstrated in a word cloud based on blog posts with a tag cloud. Creating reasonably good word cloud is challenging considering the noisy nature of blog posts.

As a proof of concept, we again use LiveJournal as a test bed to construct word clouds. One way, following current prevalent scheme of tag clouds, is to aggregate words and present the top-frequent ones. The cloud of the top 100 words generated by Wordle [15] is shown in Figure 10. Alternatively, we apply one effective group-profiling method EDGP to extract words for different groups in the site and then aggregate them at the group level. The corresponding word cloud is shown in Figure 11 with word font size denoting the relative importance.

From Figure 10, which is created by aggregation, we cannot find much useful information, though the words are not necessarily stop words. Comparatively, the word cloud

[15]http://www.wordle.net

Figure 10: Word cloud based on frequency on LiveJournal. Most features are commonly used words and convey very limited information.

Figure 11: Word cloud based on group profiling on LiveJournal. We find interesting topics such as sports, entertainments, popular names and so on.

based on group profiling (Figure 11) is more informative, with some interesting patterns. Features like leonard, patrick, michael, jackie, bob, alex and jackson are some frequently used names in reality. Words like derby, club, tickets, team are related to sports. Words like rip, cd, video, mtv, actor, lyrics, image, arts are highly relevant to entertainment. We also find more topics including belief, education, daily life, fashion, internet and so on. It seems to cover most of the hot topics on LiveJournal. Another difference is that the word font sizes in Figure 11 are more uniform than those in Figure 10 which suggests the "equal" importance of topics.

6.2 Group Search and Retrieval

On online social network sites, users usually do not know groups they want to join and are more likely to choose from the output of keyword (i.e., interest) searching. The above facts and observations suggest that group profiling, by providing an expanded and discriminative description of groups, can be used to build a better group recommendation system. Here we show one simple example how to retrieve and rank related groups to a query based on the result of group profiling. A query can have multiple words $q = \{w_1, w_2, \cdots, w_\ell\}$. Given a group profile, i.e., the ranked list of top-k features, we deem a group relevant if at least one word appears in the list. We determine each word's ranking score $r(w_i)$ by its position in the group profile. That is, if a word w_i appears in the m-th position of the profile, $r(w_i) = m$. If the word does not appear in the profile, we enforce a penalty by setting $r(w_i) = k + 1$. Then, we can compute the proximity of

the query and the group:

$$P(q, g) = \sum_{i=1}^{\ell} r(w_i)$$

Those groups with lower proximity can be returned. For instance, in LiveJournal data set, when we search "street fashion", we obtain the following top-ranking groups:

> photo_loli, fott, flammable_live, the cutters, fashion_fucks, books_and_knits, neon_haul, thriftybusiness, alt_boutique, print_project, ru_york, girl_style, egl_glamour, pansy_club, purple_hair, the_chic

Most are reasonable by looking at the group name. Some like *thriftybusiness*[16] seem irrelevant at first glimpse. But once we look at the pictures uploaded by its members, the majority of the uploaded pictures are indeed about clothes and accessories, which is relevant to the query. This example showcases the power of group profiling. We notice that in LiveJournal site, the returned groups are sorted by the last timestamp of one group being active. Instead, group profiling strategy can output groups based on the relevance. Other variants of the ranking algorithm such as a hybrid criterion of group activeness and group-query relevance can be explored. But that is beyond the scope of this work.

7. RELATED WORK

Group profiling describes the shared characteristics of a group of people. It can be applied for policy-making, direct marketing, trend analysis, group search and tracking. Tang et al. [28] present the group profiling problem in terms of topics shared by the group. They propose to classify the online documents associated with groups and aggregate them to represent the shared group interests. To capture the latent semantic relationship between different groups, the topics are organized in a hierarchical manner, represented as a taxonomy. As the semantics of different topics can vary in an evolving online environment, they propose to adapt the taxonomy accordingly when new contents arrive. Note that the work [28] concentrates on topic taxonomy adaptation. The group profiles constructed by aggregation as introduced as one of the baseline methods. Different from previous work, we systematically study different approaches for effective group profiling in this work.

Some work try to extract the annotations from relational data with text. For instance, Roy et al. [25] construct a hierarchical structure as well as corresponding annotations based on a complicated generative process. The model complexity and scalability hinder its application to group profiling in large-scale networks. Chang et al. [8] propose NUBBI (Networks Uncovered By Bayesian Inference) to infer descriptions of its entities and of relationships between these entities from a text corpora. The probabilistic topic model assumes the words are generated based on the topics associated with an entity or the topics of the pairwise relationship of entities. NUBBI annotates connections rather than groups as we do in this work.

Some other work extracts groups based on relation and text information together. Here, each topic represents a distribution of words and can be considered as the words associated with a group. Link-LDA [10] treats the citations the same way as normal words, i.e., the citation is generated based on a multinomial distribution over the documents. Pairwise Link-LDA [22] essentially combines the topic model [1] and the mixed membership stochastic block model [5] via forcing the latent mixture of communities to be the same for both word topics and relation topics. Link-PLSA-LDA [22] extends the model link-LDA one step further by modeling the citation as a mixture of latent topics instead of a multinomial distribution. Mei et al. [21] treats the connections between documents in a different fashion. It enforces the connected documents to share similar topics and use the network information as regularization while extracting the topics of texts. All those methods extract the topics associated with text instead of capturing the corresponding text associated with a given group.

[16] http://community.livejournal.com/thriftybusiness

8. CONCLUSIONS

In social media with dynamic large-scale social networks, users form implicit communities or subscribe to explicit communities. It is intriguing to understand the emergence of these social structures. In this work, we adopt a group-profiling approach to extract descriptive features for a given group. We studied different group profiling strategies to find the most descriptive features for a group. A simple aggregation strategy over individual attributes is workable only in a relatively noise-free environment. Differentiation based methods, which differentiate a group from the global network or only its neighbors, consistently outperform aggregation. More interestingly, taking an egocentric view for group profiling works as well as a global view. This suggests we can examine those actors 1-hop away from the group to understand a particular group.

As we have discussed, group profiles can be used to construct word clouds to perform network analysis or trend monitoring. They can also enhance the group search and retrieval results. More potential applications of group profiling include network visualization and navigation, tracking topic shifts, event alarming, directed marketing, group tracking, etc. Many extensions following group profiling can be explored. One possible future direction on group profiling is to study the internal structure of groups and how they affect the group profiling performance. Another interesting future task is to study the group interaction and how the group profiling can be useful in link prediction. It's also intriguing to use group profiling methods to monitor topic shifting and study group evolution. This work can also help apply community evolution and group profiling together to understanding some highly-dynamic online social networks such as Twitter.

9. ACKNOWLEDGMENTS

This research is, in part, sponsored by the Air Force Office of Scientific Research Grant FA95500810132 and FA95500910261. We thank Reza Zafarani for crawling the LiveJournal Data.

10. REFERENCES

[1] E. Airodi, D. Blei, S. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, 2008.

[2] J. Almack. The influence of intelligence on the selection of associates. *School and Society*, 16:529–530, 1922.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. 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*, pages 44–54, New York, NY, USA, 2006. ACM.

[4] J. Baumes, M. Goldberg, M. Magdon-Ismail, and W. Wallace. Discovering hidden groups in communication networks. In *2nd NSF/NIJ Symposium on intelligence and Security Informatics*, 2004.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] H. Bott. Observation of play activities in a nursery school. *Genetic Psychology Monographs*, 4:44–88, 1928.

[7] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[8] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178, New York, NY, USA, 2009. ACM.

[9] W.-Y. Chen, J.-C. Chu, J. Luan, H. Bai, Y. Wang, and E. Y. Chang. Collaborative filtering for orkut communities: discovery of user latent behavior. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 681–690, New York, NY, USA, 2009. ACM.

[10] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(90001):5220–5227, 2004.

[11] A. T. Fiore and J. S. Donath. Homophily in online dating: when do you like someone like yourself? In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1371–1374, New York, NY, USA, 2005. ACM.

[12] G. Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, 2003.

[13] M. Hechter. *Principles of Group Solidarity*. University of California Press, 1988.

[14] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5249–5253, 2004.

[15] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.

[17] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[18] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 695–704, New York, NY, USA, 2008. ACM.

[19] H. Liu and H. Motoda. *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers., 1998.

[20] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[21] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 101–110, New York, NY, USA, 2008. ACM.

[22] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550, New York, NY, USA, 2008. ACM.

[23] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[24] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, April 2007.

[25] D. M. Roy, C. Kemp, V. K. Mansinghka, and J. B. Tenenbaum. Learning annotated hierarchies from relational data. In *NIPS*, pages 1185–1192, 2006.

[26] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 687–696, New York, NY, USA, 2007. ACM.

[27] L. Tang and H. Liu. Bias analysis in text classification for highly skewed data. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 781–784, Washington, DC, USA, 2005. IEEE Computer Society.

[28] L. Tang, H. Liu, J. Zhang, N. Agarwal, and J. J. Salerno. Topic taxonomy adaptation for group profiling. *ACM Trans. Knowl. Discov. Data*, 1(4):1–28, 2008.

[29] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *KDD*, pages 717–726, New York, NY, USA, 2007. ACM.

[30] M. Thelwall. Homophily in myspace. *J. Am. Soc. Inf. Sci. Technol.*, 60(2):219–231, 2009.

[31] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[32] B. Wellman. The school child's choice of companions. *Journal of Educational Research*, 14:126–132, 1926.