Group Profiling for Understanding Social Structures

LEI TANG, Yahoo! Labs
XUFEI WANG and HUAN LIU, Arizona State University

The prolific use of participatory Web and social networking sites is reshaping the ways in which people interact with one another. It has become a vital part of human social life in both the developed and developing world. People sharing certain similarities or affiliates tend to form communities within social media. At the same time, they participate in various online activities: content sharing, tagging, posting status updates, etc. These diverse activities leave behind traces of their social life, providing clues to understand changing social structures. A large body of existing work focuses on extracting cohesive groups based on network topology. But little attention is paid to understanding the changing social structures. In order to help explain the formation of a group, we explore different group-profiling strategies to construct descriptions of a group. This research can assist network navigation, visualization, and analysis, as well as monitoring and tracking the ebbs and tides of different groups in evolving networks. By exploiting information collected from real-world social media sites, extensive experiments are conducted to evaluate group-profiling results. The pros and cons of different group-profiling strategies are analyzed with concrete examples. We also show some potential applications based on group profiling. Interesting findings with discussions are reported.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Group profiling, social structure, community, group formation, social media

ACM Reference Format:
Tang, L., Wang, X., and Liu, H. 2011. Group profiling for understanding social structures. ACM Trans. Intell. Syst. Technol. 3, 1, Article 15 (October 2011), 25 pages.
DOI = 10.1145/2036264.2036279 http://doi.acm.org/10.1145/2036264.2036279

1. INTRODUCTION

Recently, a surge of work has reported statistical patterns presented in complex networks across many domains [Chakrabarti and Faloutsos 2006; Newman 2003]. A substantial body of existing work studies global patterns present in a static or an evolving network [Kumar et al. 2006; Leskovec et al. 2007]. Microscopic patterns such as individual interaction patterns are also attracting increasing attention [Leskovec et al. 2008a]. This work, alternatively, focuses on meso-level analysis of a network. In particular, we study groups (communities) in social media. Group-level analysis plays a key role in social science. "The founders of sociology claimed that the causes of social

This research is supported in part by AFOSR.
Authors' addresses: L. Tang (corresponding author), Yahoo! Labs Silicon Valley, Santa Clara, CA 95054; email: l.tang@asu.edu; X. Wang and H. Liu, Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287-8809.

DOI 10.1145/2036264.2036279 http://doi.acm.org/10.1145/2036264.2036279

phenomena were to be found by studying groups rather than individuals" [Hechter 1988, Chapter 2, p. 15].

A *group* (*community*[1]) is a set of users who interact with each other frequently [Wasserman and Faust 1994]. It has a broad range of applications to discover groups, including network visualization, intelligence analysis [Baumes et al. 2004], network compression [Sun et al. 2007], behavioral study [Tang and Liu 2010], targeting and recommendation [Wang et al. 2010b], and collaborative filtering [Chen et al. 2009]. A variety of community detection methods (a.k.a. finding cohesive subgroups [Wasserman and Faust 1994]) have been proposed to capture such social structures in a network. With the expanded use of the Web and widely available social networks, identifying evolving groups in dynamic networks is also gaining increasing attention [Hopcroft et al. 2004; Palla et al. 2007; Sun et al. 2007; Tantipathananandh et al. 2007].

While a large body of work has been devoted to discovering groups or group evolution based on network topology, few have systematically delved into extracted groups to understand the formation of a group explicitly. Some fundamental questions remain.

— How can we understand a social structure emanated from a network?
— What is the particular aspect that binds group members together?

Some pioneering works attempt to understand group formation based on statistical structural analysis. [Backstrom et al. 2006] studied prominent online groups in the digital domain, aiming at answering some basic questions about the evolution of groups, such as: what are the *structural features* that determine which group an individual will join? They found that the number of friends in a group is the most important factor in determining whether a new actor would join the group. This provides a global level of structural analysis to help understand how communities attract new users. [Leskovec et al. 2008b] observed that spectral clustering (a popular method used for community detection) always finds tight, small-scale yet almost trivial communities, that is, the community is connected to the remaining network via one single edge. Both papers focus on a global (statistical) picture of communities. Further research is required to understand the formation of particular groups.

Various reasons can lead to the formation of a community. For example, some users may interact with each other because they attend the same university; some users form a group as they are both enrolled in the same event. Users can also coalesce if they share the same political views. In this work, we attempt to understand social group from *a descriptive* aspect, which helps explain group formation processes. We aim to extract group attributes that help understand a group. For the aforementioned examples, the group attributes, should ideally indicate the university, the event, and the political view, respectively. We investigate the following research questions.

— Given individual attributes, can we find out their group-level shared commonalities?
— If yes, what are the effective approaches?

Extracting descriptive attributes from a group of people is referred to as *group profiling* [Tang et al. 2008]. To construct a group profile, we study strategies to extract attributes for a group when individual attributes are available. This is especially applicable with online 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 large number of individual traces poses a challenge to extract useful information to describe a group. In this work, three sensible methods

[1]In this work, group and community are used interchangeably.

of group profiling are presented for comparative study: *aggregation, differentiation, and egocentric differentiation*. Another challenge is that evaluation usually requires extensive human effort to delve into group member activities to identify the shared similarities among them. We carefully designed experiments to alleviate human burden for Extensive experiments with concrete case studies on two social media domains demonstrate the effectiveness of group profiling based on (egocentric) differentiation. 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 to construct a descriptive profile for a provided group. In this section, we motivate this task and formally define the problem.

2.1. Motivation

According to the concept of homophily [McPherson et al. 2001], a connection occurs at a higher rate between similar people than dissimilar people. Homophily is one of the first characteristics studied by early social network researchers [Almack 1922; Bott 1928; Wellman 1926], and it holds for a wide variety of relationships [McPherson et al. 2001]. Homophily is also observed in social media [Fiore and Donath 2005; Lauw et al. 2010; Thelwall 2009]. In this work, we study the "inverse" problem: given a group of users, can we figure out why they are connected? Or what are their shared similarities?

It is impossible to answer these questions with social networking information alone. Luckily, social media sites often provide more information than just a network. In the blogosphere, users post and tag blog posts. On Facebook, users chat with each other, update their status, leave comments, and share interesting links. These different activities reflect online social life of users, and thus can be used to answer the aforementioned questions.

Social media sites often support interaction and networking between users. For instance, Twitter[2] has a following-follower network. There, community detection methods can be applied to find out *implicit groups* hidden beneath the interactions. Group profiling, in this case, can be used to understand the extracted communities, facilitating the network analysis and community tracking.

At other social media sites like LiveJournal[3], Flickr[4], YouTube[5], and Facebook[6], users are allowed to form their own *explicit groups*. Some might suspect that the group name and description already provide enough information to peek into an explicit group. Unfortunately, this is not necessarily true. In LiveJournal, we encountered a large number of communities whose profile page provides little information on the group. For instance, the community profile of *fruits*[7] does not say much about the exact topic of the community. The group name might provide some hints, but it can also be misleading in certain cases. Consider *fruits* as an example again. A first glimpse at the community name led us to think that this community is composed of people who

[2]http://twitter.com/
[3]http://www.livejournal.com/community/
[4]http://www.flickr.com/groups/
[5]http://www.youtube.com/groups_main
[6]http://www.facebook.com/
[7]http://community.livejournal.com/fruits/profile

are fond of fruits. However, after we applied group profiling[8] on this community, we obtained 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 characters in animes and mangas like *hello kitty*[9] are often discussed as well. It is known that *hello kitty* is a very popular character used in Japanese fashion.

Group profiling can help understand implicit communities extracted based on network topology as well as explicit communities formed by user subscriptions. Besides understanding social structures, group profiling also be helpful in network visualization and navigation, tracking the topic shift of a group, event alarming, direct marketing, and connecting the dots. 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. It is noticed that online networks can be divided into three regions [Kumar et al. 2006]: 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 a 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.

2.2. Problem Statement

In order to understand an emerging structure in social media, we aim to build a group profile that illustrates the concerns of 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 the edge (connection) set;
— A particular group $g = (V_g, E_g)$ where $V_g \subseteq V$, and $E_g \subseteq V_g \times V_g$, $E_g \subseteq E$.
— 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.

Here we assume the attributes of individual users are boolean. For instance, one attribute can denote the gender of actors, or their attitude toward abortion. It can also represent whether a word occurs in an actor's status update, blog posts, or tags. In some real-world applications, individual attributes might be categorical rather than boolean, for example, 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

[8]More details in later parts.
[9]http://www.sanrio.com/

Table I. Statistics Based on Group and Attribute

	$group = +$	$group = -$
$A = 1$	true positive (tp)	false positive (fp)
$A = 0$	false negative (fn)	true negative (tn)

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 simply focus on boolean attributes. For convenience, we say a node has an attribute A_i if $A_i = 1$ for the node.

It is desirable that a group profiling method satisfies the following properties.

— *Descriptive.* The selected attributes for a group should reflect the foundation of a group, say, the shared interest or affiliation.
— *Robust.* Mountains of data are produced each day in social media. These data tend to be very noisy. The group profiling method should be robust to noise.
— *Scalable.* In social media, a network of colossal size is the norm. Typically, one network involves hundreds of thousands or millions of actors. For example, as of April 20, 2011, LiveJournal has more than 31 million registered users and around 188,194 users updated their journals in the past 24 hours[10]. Twitter has 190 million users and 65 million tweets per day[11]. Facebook has more than 500 million active users, and on average, each user creates 90 pieces of content per month[12]. Meanwhile, networks are highly dynamic. Each day, new users join a network, and new interactions occur between existing ones. Users engage in various activities, producing rich user interactions and overwhelming volume of user-generated content. This also presents a challenge for group profiling.

Following the preceding guidelines, we next present several group-profiling strategies.

3. PROFILING STRATEGIES

In this section, we present several strategies for group profiling. We assume that 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 proceed to the methods, 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 other nodes that do not belong 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 I.

— 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.

[10]http://www.livejournal.com/stats.bml
[11]http://techcrunch.com/2010/06/08/twitter-190-million-users/
[12]http://www.facebook.com/press/info.php?statistics

Given these measures, 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 is 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-based Group Profiling (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. This Aggregation-based Group Profiling (AGP) essentially solves 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. Note that this aggregation-based profiling is widely used in current tagging systems in the form of tag clouds. Tag clouds are widely used in social media to show the popularity of a tag by its font size. If we consider the whole network as a group, then a tag cloud is produced based on aggregation of the group tags.

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 wisdom of crowds such as user shared tags may not help much following this aggregation strategy. Take one community named *photography*[13] in LiveJournal as an example. It is not difficult to figure out the shared interests among the group members. If we look at those interests that occur most frequently in profiles of users within the group, we have the following list.[14]

photography, art, music, movies, reading, writing, love, books, painting

Except the first two, other tags are actually not very good group descriptors. These tags are very general, and they are shared by a large number of people, thus appear in this group as well. Directly aggregating these tags is biased towards selecting popular tags, rather than those that can characterize this group.

3.2. Differentiation-based Group Profiling (DGP)

Instead of aggregating, we can select features which differentiate one group from others in the network. Hence, the group-profiling problem amounts to feature selection [Liu and Motoda 1998] in a 2-class classification problem with the group being the positive class and the remaining nodes in the network as the negative class. The goal is to find out those top-k *discriminative* features which are representative of a group.

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

[14]More details are in the experiments in Sections 4 and 5.

Note that a particular group is fairly small compared to the whole network. For instance, the LiveJournal dataset that we collected has 16,444 users, and the top two largest groups have around 5,000 and 1,500 members respectively. The majority (90.1%) of the groups are in the long tail, each with less than 100 members. This results in a highly unbalanced class distribution [Tang and Liu 2005]. With this skewed class distribution, Bi-Normal Separation (BNS) [Forman 2003] is an effective method that outperforms other feature selection methods [Forman 2003; Tang and Liu 2005] such as information gain and χ^2 statistic. The BNS score of an attribute is defined as

$$BNS = \left| F^{-1}(tpr) - F^{-1}(fpr) \right|, \tag{4}$$

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

$$tpr_{A_i} > fpr_{A_i} \tag{5}$$

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

Combining the BNS criterion in Eq. (4) and the constraint in Eq. (5), we have the following formulation for Differentiation-based Group Profiling (DGP).

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

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

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

Essentially, we select those features that appear frequently in one group but rarely outside the group.

3.3. Egocentric Differentiation-based Group Profiling (EDGP)

In the previous differentiation strategy, all the nodes outside a group are deemed as belonging to the 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 popular online social networks are huge, with hundreds of millions of nodes. It is either time consuming or impractical to retrieve all the information of a real-world network. In some applications, only an egocentric view is available. In other words, we know our friends, but have 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 a group from the whole network, we propose to differentiate the group from the neighbors of group members, that is, group profiling based on the Egocentric View (EDGP). Group neighbors refer to nodes outside a group that are connected to at least one group member as in Figure 1. Egocentric differentiation follows the same 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

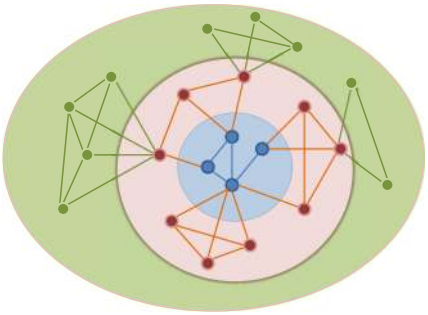

Fig. 1. Blue nodes in the center form a group. Red nodes in pink area are the group neighbors.

and EDGP, one wonders if this egocentric approach suffices in finding discriminative features.

4. EXPERIMENT SETUP

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

4.1. Evaluation Methodology

Group profiling outputs a list of features to describe groups. The quality of the extracted profile depends on the group-profiling method being used. There are several challenges in the comparison. We will address them one by one.

(1) How can we obtain group information? For evaluation purpose, we use explicit communities in social media as the group information. As we mentioned in the Introduction, in certain social media sites, users can subscribe to one or more interest groups. Explicit communities come with their group names and sometimes descriptions as well. These information can help human subjects find out the ground truth for evaluation. Of course, this evaluation strategy does not prevent the group-profiling approach from being applied to implicit groups extracted from a network. As shown later, most explicit online groups also demonstrate a much higher link density than expected.
(2) In order to extract group profiles, what kind of individual attributes should we look into? In social media sites, users can share their profiles, upload tags, post blogs, and update status. All these activities provide some signals. For experiment purpose, we treat user interests in profiles or words and tags occurring in their posts as attributes, and find the key attributes to describe groups.
(3) How to evaluate the quality of extracted group profiles? Since there is no ground-truth information available, we invite people with different backgrounds to evaluate the results.

We launched a Web site with a user-friendly interface for evaluators to log in and rate. A screenshot of the Web site after a user log in is shown in Figure 2. For each group, we use the three proposed approaches (AGP, DGP, EDGP) to select top k (k = 10 in our experiments) most representative features. On each evaluation page, the profile features extracted according to one method were listed in a column in order of descending importance. The three approaches are denoted as methods 1, 2, and 3, respectively in the screenshot. It should be emphasized that evaluators do not know what the group profiling methods are, or which column corresponds to which method. To avoid the bias associated with the column position, the presentation order of group

Fig. 2. Screenshot of the evaluation system.

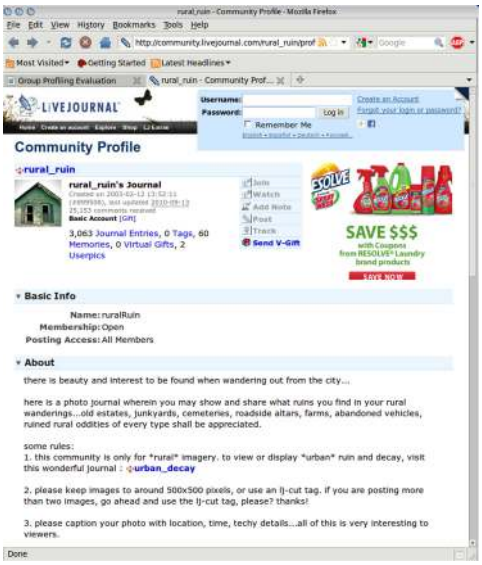

Fig. 3. Group profile page for reference.

profiles is randomized for each page. 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.

We also highlighted the title of the studied group and provided a link to the particular online group profile page, so that evaluators are encouraged to get general group information before making a decision. For instance, by clicking on the link at the top of the screenshot in Figure 2, one will be directed to the group page as in Figure 3. The

Table II. Statistics on BlogCatalog and LiveJournal

	BlogCatalog	LiveJournal
# Bloggers	70,086	16,444
# 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

profile page contains some description, as well as links to the activities and journal posts within the group. Hopefully this can help a subject make the right decision.

Each evaluator will rate the resultant profiles on how well they are describing this group. The rating ranges from 0 to 3, respectively representing "irrelevant", "partly related", "reasonable", and "very good". An evaluator can also decline to give a rating (by choosing a "no idea" option) if he is not sure. As we notice in one pilot study, subjects tend to assign random ratings if the task takes too long. To assure the quality of evaluation, each person was asked to evaluate only 10 group profiles in one session, which can be finished in roughly 2–5 minutes.

4.2. Social Media Data

As mentioned before, we need datasets with groups as well as rich individual attributes. Hence, we select two social media sites for data collection: BlogCatalog[15] and LiveJournal[16]. BlogCatalog is a social blog directory where bloggers can register their blogs under specified categories. LiveJournal is a virtual community where users can keep a blog, journal, or diary. Both Web sites serve as a platform for users to connect and communicate with others. At both sites, users can engage in social activities like adding friends, joining groups, commenting, tagging, and so on.

On BlogCatalog, we crawled each blogger's name, his friends, his blog sites, tags, categories, and the most recent post snippets. We treat blog categories as groups. After removing the non-English blogs, we obtained $70,086$ bloggers and 344 groups. There are in total $1,706,145$ friendship links. Each blogger has 49 friends on average. On LiveJournal, we started with a popular blogger *just_ducky*, and crawled bloggers who are reachable within 4 hops from this seed user by following their friendship connections. We collected each 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 dataset has $16,444$ bloggers, more than $130K$ friendship links and $100,441$ different communities. The statistics of these two datasets are summarized in Table II. One key difference between these two social media Web sites is that LiveJournal bloggers can create communities freely. BlogCatalog users, however, 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 distributions at both sites are plotted in Figures 4 and 5 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 a histogram instead

[15]http://www.blogcatalog.com/
[16]http://www.livejournal.com/

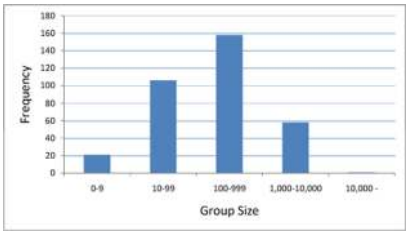

Fig. 4. Group size distribution in BlogCatalog (in a bell curve). Few groups have less than 10 members, and only 1 group has a size greater than 10,000.

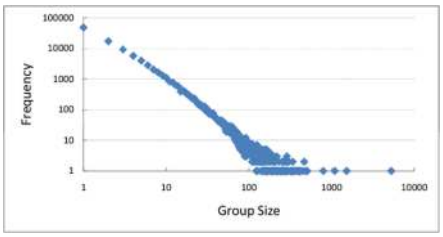

Fig. 5. Group size on LiveJournal follows a power law distribution. Most groups have less than 100 members.

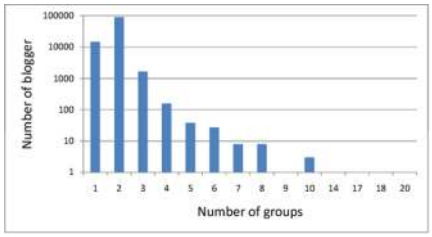

Fig. 6. Group affiliation distribution of bloggers on BlogCatalog. Around 90% of the bloggers join 2 groups; the average groups one blogger joins is 1.9.

of a scatter plot. The group size distribution in BlogCatalog is more like a bell curve, possibly because of the different mechanism for creating groups as we mentioned before. On the contrary, group size in LiveJournal follows a power law distribution as observed in many large-scale networks.

The number of groups one blogger joins is shown in Figures 6 and 7. In BlogCatalog, most bloggers join two groups, but a few bloggers (0.23%) join more than three groups. In LiveJournal, the distribution is different, with 82.3% bloggers joining at least four groups. One blogger has joined 1, 032 groups. On average, a blogger subscribes to 1.9 and 32.6 groups respectively on these two sites.

In the experiment, we would like to test group-profiling methods with different noise levels and investigate how each method performs. Typically, words in blog posts are much noisier than tags or user interests listed in users' profile pages. Hence, we created four datasets: BlogCatalog based on tags (BC-Tag) or blog posts (BC-post), and LiveJournal based on user interests (LJ-Interest) or journal posts (LJ-post). We expect LiveJournal to be noisier than BlogCatalog as the communities there are user-generated rather than prespecified.

Since the evaluation involves human effort, it is impractical to exhaustively evaluate all groups. We select a subset of representative groups with varying sizes and

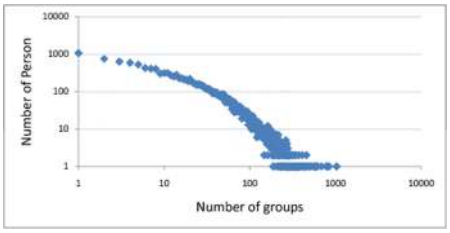

Fig. 7. Group subscriptions distribution of bloggers on LiveJournal (in a power law distribution).

Table III. Selected Groups in BlogCatalog

Group	Size	Density	Group	Size	Density
personal	11478	1.3‰	dogs	173	8.0‰
blogging	7727	2.7‰	adult education	139	1.3‰
entertainment	4671	1.9‰	buddhism	96	11.0‰
health	3877	2.4‰	hunting	86	41.0‰
shopping	2687	2.1‰	sailing	71	8.9‰
sports	2529	2.0‰	lawn&garden	55	8.9‰
computers	1934	2.4‰	music industry	47	6.1‰
animals	1357	5.6‰	natural	41	10.0‰
investing	906	3.8‰	city guides	40	32.0‰
science	826	2.4‰	anarchism	29	34.0‰
home cooking	564	3.7‰	auto repair	23	4.3‰
hardware	424	1.2‰	earth science	22	16.0‰
pop	254	2.5‰	aqua. fish	19	17.0‰
stock&bond	245	7.1‰	choreography	13	26.0‰
cultural	229	4.5‰	extinct birds	3	0.0‰

densities as listed in Tables III and IV. In particular, we select 30 groups from BlogCatalog and 32 groups from LiveJournal. For evaluation purpose, here we use explicit groups, that is, groups in which the membership is determined by subscription. But we would like to point out that the density of most groups is much higher than the network density, suggesting frequent within-group interactions. Their neighborhood size versus the group size is also plotted in Figure 8. Because each node has a plurality of connections, the neighborhood size is typically much larger and increasing with respect to the group size.

5. EXPERIMENT RESULTS

52 people with assorted backgrounds (undergraduate, graduate students,university faculty, and employees) participated in our evaluation. In total, 2, 028 ratings were collected, of which 101 ratings were "no idea". The remaining 1, 927 ratings were used in our analysis. Each group was evaluated 32 times, and the average ratings were reported.

5.1. Comparative Study

The average ratings for each method on different datasets are shown in Table V. On BC-Tag, three methods are comparable, however, the aggregation-based approach deteriorates when we use words in the blog posts as features. A similar pattern is observed on LiveJournal, though the ratings drop sharply. On both datasets, DGP and EDGP consistently outperform AGP. This is most observable when individual

Table IV. Selected Groups in LiveJournal

Group	Size	Density	Group	Size	Density
photography	320	13.0‰	ontd_startrek	139	12.0‰
sextips	297	1.8‰	behind_the_lens	134	16.0‰
mp3_share	288	2.1‰	tvshare	132	5.2‰
art_nude	232	33.0‰	ru_portrait	131	76.0‰
ourbedrooms	216	12.0‰	knitting	124	2.3‰
houseepisode	211	6.2‰	girl_gamers	121	3.6‰
fruits	205	16.0‰	wow_ladies	115	2.0‰
free_manga	205	9.1‰	art_links	113	50.0‰
ucdavis	189	39.0‰	weddingplans	110	4.7‰
photographie	188	12.0‰	doctorwho_eps	109	25.0‰
cooking	181	2.3‰	ru_travel	108	20.0‰
hot_fashion	161	25.0‰	blythedoll	108	110.0‰
naturalliving	157	3.8‰	rural_ruin	105	14.0‰
topmodel	155	2.8‰	supernatural_tv	103	15.0‰
photocontest	147	1.5‰	animeicons	102	5.0‰
cheaptrip	142	29.0‰	gossipgirltv	101	8.1‰

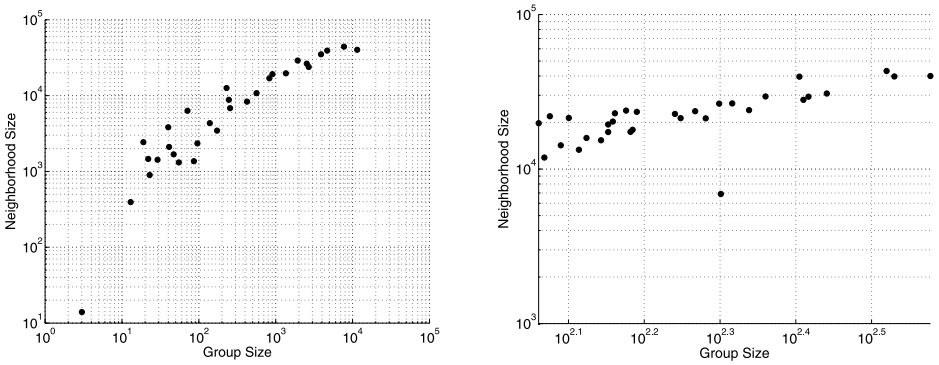

Fig. 8. Selected group size versus neighborhood size on BlogCatalog and LiveJournal.

attributes are noisy. That is, a large number of attributes are associated with individuals, among which only a few are relevant to the group topic (say, when words appearing in blog posts are used as attributes).

This result is clearer in Figure 9, where we plot the probability of each group profiling method being the winner. It 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 yield a similar performance. But on the other datasets, DGP and EDGP are consistently better than AGP, and the difference between the former and the latter increases as the noise level increases (LiveJournal is noisier than BlogCatalog as communities are not prespecified, and posts are noisier than tags or user-specified interests).

The performance of DGP and EDGP are 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,

Table V. 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

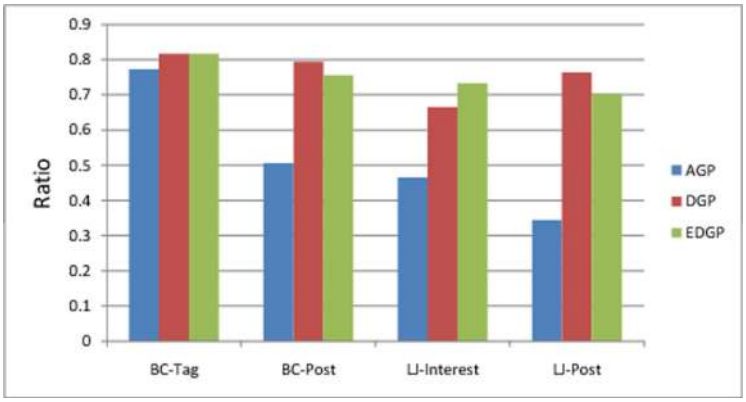

Fig. 9. The probability one method receives highest ratings. We treat all methods have highest ratings if all the three methods have the same rating score. The performance of AGP deteriorates as the noise level increases. DGP and EDGP are consistently better than AGP on all datasets.

the computation cost of profiling a particular group can dramatically drop because of a much smaller number of involved bloggers. In 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, for groups whose sizes are larger than 50, the average number of 1-hop away bloggers is 1,016, or around 6.2% of all the bloggers. The egocentric differentiation method is favorable in dynamic and evolving huge networks, because updating features is straightforward. Only the local information, instead of the whole network, is required.

5.2. Case Studies

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. The topics covered in this group are *medicine, diet, weight loss, men's and woman's health,* and so on. Table VI presents profiles extracted to describe the group based on tags and posts, respectively. The features are sorted by importance in descending order. In BC-Tag, features extracted by all the three methods are related to health. Only the order of some keywords are different. In BC-post, the result of AGP becomes worse. Some features like *world*, *long*, *find*, and *important* seem 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. These two methods demonstrate subtle differences, only the order of some features differs.

Table VII shows profiles for *blythedoll* group on LiveJournal. Blythedoll was first created in 1972 by U.S. toy company Kenner. Later it spread out to the world. In LJ-Interest, some of the features extracted by the AGP method are very frequently used words, for example, photography, art, and music, and we can hardly connect them to

Table VI. Profiles for *Health* Group in BlogCatalog

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 health	back	healthy	food
mental health	mental health	wellness	important	treatment	healthy

All methods based on tags are comparable. But for blog posts, methods DGP and EDGP perform much better than AGP.

Table VII. Profiles for *Blythedoll* Group in LiveJournal

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

AGP performs poorly in LJ-Post since all the features are not explicitly related to blythedoll. DGP and EDGP are consistently better than AGP.

blythedoll. In LJ-Post, the AGP result is even worse: there is almost no connection to the blythedoll group. The other two methods, DGP and EDGP, perform consistently better than simple aggregation. This example demonstrates the superiority of DGP and EDGP with noisy data.

5.3. Similarity between Profiles of Different Methods

In previous experiments, we have shown that (egocentric) differentiation-based group profiling tends to outperform the aggregation-based method. In this subsection, we systematically examine the similarity of the profiles produced by the three methods. We notice that DGP and EDGP receive similar ratings as reported in Section 5.1. Is this due to the fact that they often select similar features to construct group profiles?

As each method outputs a ranked list of attributes, we use Kendall's Tau(τ) rank correlation coefficient [Kendall 1938] to measure the difference of the ordering. Kendall Tau Coefficient measures the agreement between two ranked lists. In our experiments, only ten terms are selected for each group, and we first construct two ranked lists by assigning a rank for each term. Given two rankings R_1 and R_2 concerning the same set of elements, let x_1 and x_2 denote the rank of element x in R_1 and R_2 respectively. Two elements x and y are a *concordant pair* when the ranks for both elements agree,

Table VIII. Average Kendall's Tau Rank Coefficient between Different Methods

	BC-Tag	BC-Post	LJ-Interest	LJ-post
AGP / DGP	0.48	0.18	0.10	0.14
AGP / EDGP	0.42	0.08	**0.11**	0.11
DGP / EDGP	**0.60**	**0.31**	0.10	**0.15**

Table IX. Jaccard Coefficient between Different Methods

	BC-Tag	BC-Post	LJ-Interest	LJ-post
AGP / DGP	0.80	0.42	0.22	0.04
AGP / EDGP	0.73	0.32	0.07	0.01
DGP / EDGP	**0.85**	**0.71**	**0.31**	**0.14**

that is, if $x_1 < y_1$ and $x_2 < y_2$, or $x_1 > y_1$ and $x_2 > y_2$. x and y form a discordant pair if the relative rank of the two does not agree, that is, if $x_1 < y_1$ yet $x_2 > y_2$, or $x_1 > y_1$ yet $x_2 < y_2$. The Kendall τ coefficient is defined as

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\frac{1}{2}n(n-1)}.$$

Its value is between -1 (one ranking is the reverse of another) and +1 (two rankings are the same). Two ranks have no correlation if their Kendall Tau Coefficient is 0.

The τ coefficients on all the four datasets are listed in Table VIII, with entries in bold face to denote the highest similarity in each column. It is observed that all methods demonstrate a positive correlation. Among them, DGP and EDGP often output similar rankings. It is observed that the coefficient on LiveJournal data is much smaller than that on BlogCatalog. This difference might be due to more noise embedded in the LiveJournal data.

If the ordering effect is ignored, one might be only interested in the set of top-ranking attributes. Thus, we computed the Jaccard similarity [Jaccard 1901] between the top-ranking attributes output by different methods. Given two sets A and B, Jaccard similarity is defined as

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

Its range is between 0 and 1. The average Jaccard similarity between the top-10 attributes as selected by different profiling methods are reported in Table IX.

Again, DGP and EDGP are quite similar, especially on the BlogCatalog data. This fact explains why their ratings are similar as reported in Section 5.1. It also suggests that by comparing one group with its neighborhood, rather than the whole network, it is often sufficient to extract a discriminative group profile.

5.4. Further Analysis

5.4.1. Understanding Evaluation Results. We noticed that different groups receive distinctive ratings even for the same group-profiling method. What might be the reason leading to this differences? Is there any connection between group size and ratings? Figure 10 plots individual group ratings of EDGP on BC-Post. The groups are sorted, from left to right, by group sizes in a descending order. No evident correlation is found between the group size and the quality of group profiling. Large groups such as "personal" can receive low ratings, and small groups like "auto repair" can have high ratings. We observed similar patterns on other datasets with different profiling methods.

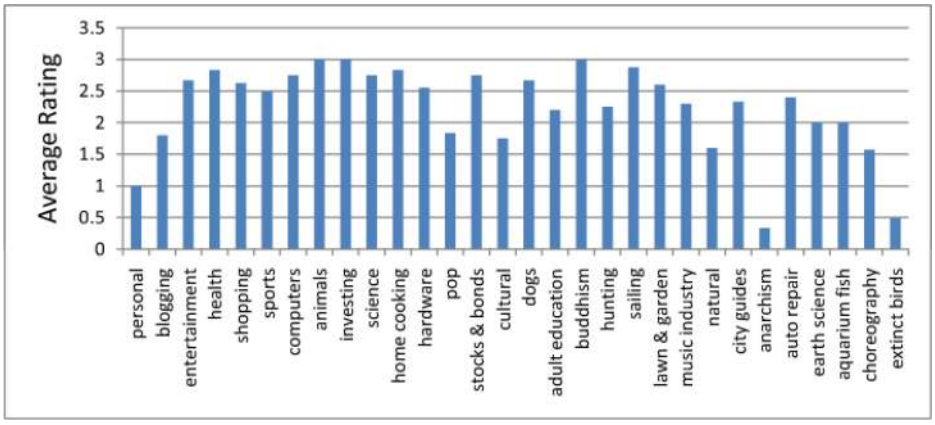

Fig. 10. Rating of individual groups based on EDGP on BC-Post. The groups, from left to right, are sorted by group size in descending order. No significant correlation between ratings and group sizes is found.

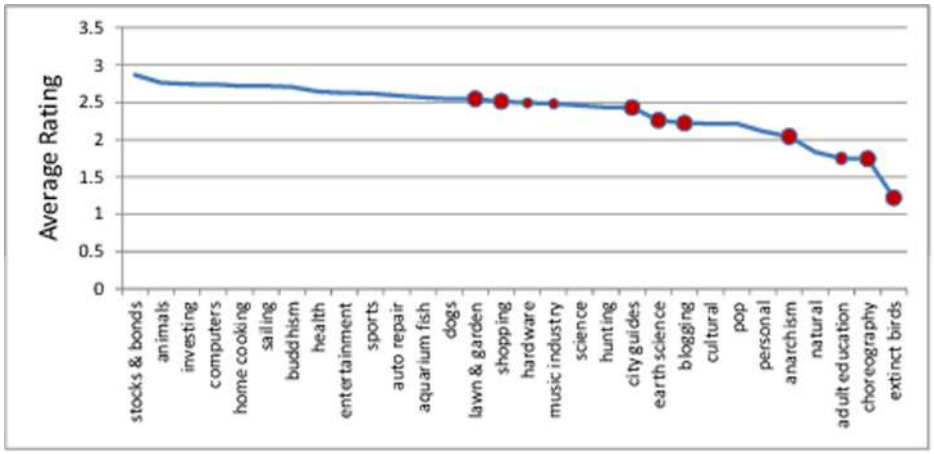

Fig. 11. Groups receiving "no idea" ratings. The curve Average ratings on BlogCatalog. The red circles varying in size represent the relative probability of the group receiving "No Idea" ratings. Groups are listed by ratings in descending order.

One interesting finding is that as the specificity of group increases, so does its rating. For instance, the largest group "personal" contains 11, 478 members but has an average rating of 1. Group "auto repair" with only 234 members receives a rating of 2.4. This result agrees with intuition that it is more difficult to describe general concepts, but easier to describe a specific one.

We further analyze the user evaluation behavior. We show the groups of BlogCatalog in Figure 11 sorted by their average ratings. The red circles in the curve highlight those groups receiving "no idea" during evaluation, with their sizes indicating the relative probability. The markers tend to reside at the tail of the curve, that is, when the rating is relatively low. When it is difficult for a human to judge what a particular group is about, it is not surprising that the performance of group profiling decreases as well.

5.4.2. Exploiting Group Internal Structures. For all our studied methods, we do not exploit the internal structure inside a group. Presumably, all groups have their influentials

Table X. Average Similarity of Profiles after Applying Degree-Centrality Weighting

	BC-Tag	BC-Post	LJ-Interest	LJ-post
AGP_{wo}/AGP_w	0.35	0.30	0.94	0.59
DGP_{wo}/DGP_w	0.20	0.05	0.06	0.01
$EDGP_{wo}/EDGP_w$	0.21	0.06	0.05	0.01

Table XI. Profiles for *City Guides* in BlogCatalog with or without Weighting

Without Weighting	With Weighting
olympic games	singapore sights
travel	singapore food
country	singapore recommendations
california	singapore places
tourism	singapore parks
islam	travel products
people	boutique hotels
lifestyle	travel deals
culture	travel style
reviews	luxury resorts

[Agarwal et al. 2008]. These are opinion leaders, and may play a more important role to reflect the peculiarity of a group. There are many ways to define the importance of a node. Commonly used measures include degree centrality, closeness centrality, betweenness centrality, or eigenvector centrality [Wasserman and Faust 1994]. Here, we take degree centrality as an indicator of a node's importance inside a group. The more connections one has inside a group, the more central role he plays in the group. The number of one node's connections inside a group is used as a weight when we compute the relevant statistics in Table I.

After applying this simple weighting for profiling, we observe that the top-ranking features are changed for many groups. Table X shows the average Jaccard similarity between methods with and without weighting. For DGP and EDGP, the weighting can change the profile a lot. Nevertheless, AGP is not affected as much by the weighting.

It is noted that the group profiles with a weighting scheme demonstrate some interesting patterns. Those more specific attributes might appear in a profile. For example, Table XI shows the DGP profiles for group *City Guides* with or without weighting. Both types of profiles are sensible. The profiles without weighting seem to be more general whereas some specific terms related to Singapore appear frequently on the right column as the central node is quite interested in visiting there. It is difficult to conclude which type is better. However, it is clear that the group internal structures can play a role in the construction of different informative profiles. We expect that the group internal structure as well as connections to members outside the group can affect the profiling output, and this influence requires further research.

6. POTENTIAL APPLICATIONS OF GROUP PROFILING

Group profiling can help describe groups. The group description can be further used in various types of applications. For instance, group profiles can be used to enrich user profiles. User profiling [Shmueli-Scheuer et al. 2010] is one fundamental task in targeting and advertising. However, some users might have very few features. In this case, borrowing features from their group profiles can help improve targeting [Shi et al.

2010]. Group profiles can also be used to understand the formation of implicit groups, assist community tracking, and group search. Next, we showcase two applications of group profiling: one for understanding implicit groups, and the other for group search and retrieval.

6.1. Understanding Implicit Groups

Social media provides a large volume of network interactions which can be used to study human interactions on an unprecedented scale. These large-scale networks present strong community structures [Chakrabarti and Faloutsos 2006]. Group profiling can help understand those implicit groups behind these diverse interactions. Here, we show some interesting findings of group profiling applied to a Flickr network.

Flickr[17] is a photo sharing Web site where photos are organized in a collaborative way such that both the owner and browsers can upload tags to them. We crawled user names, their contacts, and tags associated with their uploaded photos, ending up with 39,933 users and more than 3.59 million connections after 2 weeks. We applied the EdgeCluster algorithm [Tang and Liu 2009] to find overlapping communities inside the network. EdgeCluster defines a community as a set of edges, rather than a set of nodes like the majority of existing work. By partitioning edges into disjoint sets, it allows the resultant communities to overlap. We obtained 171 clusters with varying sizes. After applying group-profiling methods to those clusters, we have several interesting observations.

— People are usually gathered by their nationality. Flickr is an international social media site, and people from different countries might speak different languages. This is intuitive since people tend to tag places and events in their own languages. We found groups extensively focused on Italian, Arabic, Indian, Malaysian, Farsi, Spanish, and so on. A representative profile for an Italian group is shown next (only the top 15 keywords are included).

> *bimba, italians, Italians, ritratto, amicizia, ombrello, abbandono, autunno, viaggio, luce, amica, dolcezza, colori, nuvole, gambe*

  All keywords except *italians* and *Italians* are all in Italian. For instance, *bimba* means infant, ritratto means picture or portrait. The other words starting from *amicizia* can be translated to friendship, umbrella, neglect, autumn, travel, light, friend (female), sweetness, colors, clouds, legs, respectively. The topic for this large community appears to be unfocused. But, based on group profiling, we know that the communication at a high level is mainly about people speaking the same language. We can also apply group profiling to subcommunities to understand each community at a finer granularity.

— People connect to like-minded peers. Their shared interests are reflected in group profiles. For example, the top keywords for one of these groups is shown here.

> *TheUnforgettablePictures, TopShots, platinumphoto, SuperShot, GoldStarAward, RubyPhotographer, NaturesElegantShots, ourmasterpiece, SOE, Cubism, GoldDragon, AnAwesomeShot, ABigFave, WorldWideLandscapes*

  These keywords are highly similar in semantics, reflecting users' consensus in their preference. We found that most keywords are actually titles of some *explicit* interest groups in Flickr. Though people subscribe to different interest groups with different titles, they interact with each other frequently, thus forming an *implicit* group with

[17]http://www.flickr.com/

similar interests. This indicates the usefulness of group profiling in understanding community structures in social media.

6.2. Group Search and Retrieval

On social networking sites, users may want to subscribe to different groups. Some groups might match their interests, but with a misleading group name. In this case, it is difficult for a user to locate groups of interest. On the other hand, advertisers would like to launch campaigns target those groups with desired properties, such as age, gender, education level, interest, etc. Group profiling, by providing an expanded and discriminative description of groups, can be used to build a better group recommendation system. As a proof of concept, we present one example to show how to retrieve and rank related groups to a query based on the result of group profiling. More advanced techniques may be borrowed from the tasks in BlogTrec [Macdonald et al. 2010].

A query can have multiple words $q = \{w_1, w_2, \cdots, w_\ell\}$. Given a group profile, that is, the ranked list of top-k features, we deem a group relevant if at least one word in q appears in the list. We determine each word's ranking score $r(w_i)$ by its position in the group profile. That is, $r(w_i) = m$ if a word w_i appears in the m-th position of the profile. 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 as recommended. For instance, in the LiveJournal dataset, the search of "street fashion" results in 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

We can tell that most groups are reasonable by looking at the group names. Some results like *thriftybusiness*[18] seem irrelevant at first glimpse. But once we look at the pictures uploaded by its members, we notice that the majority of the uploaded pictures are indeed about clothes and accessories, confirming the relevance of the group to the query. This example showcases the power of group profiling. The LiveJournal Web site also provides a group search engine. It sorts returned groups by recency of one group being active. The group profiling strategy can find groups based on relevance. In practice, ranking can be accomplished following a hybrid criterion of group activeness and group-query relevance can be explored.

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. 2008] present the group profiling problem in terms of topics shared by the group. They propose to classify online documents associated with groups, and then aggregate the class labels to represent the shared group interests. To capture the latent semantic relationship between different groups, 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

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

new content arrives. Note that the work Tang et al. [2008] concentrates on topic taxonomy adaptation. Group profiles are constructed by aggregation. In this work, we systematically study different approaches for effective group profiling.

Group profiling is also applied by sociologists to understand politics and culture in the Persian blogosphere [Kelly and Etling 2008]. In the study, bloggers are first clustered based on their link structure. Then, human beings are hired to assign topics and write a short summary for each blog site. Based on the description, the authors analyze profiles associated with each group. They also count frequencies of Iranian related terms occurring in each group and report interesting patterns associated with different groups, such as which terms occur frequently in one particular group, or what common terms are shared by two different groups. All the preceding analysis requires a lot of human effort. That is where our automatic group-profiling techniques can help extend the analysis to a much larger scale.

In this work, we map group-profiling problems to feature selection. Feature selection chooses a subset of features to represent the original high-dimensional data, in order to improve prediction performance or reduce time and space complexity [Guyon 2003]. It has been widely used in various domains. Different metrics are used to measure the importance of features. Take text as an example, term frequency, document frequency, tf-idf weight [Jones 1972], χ^2 statistic, information gain, and mutual information are commonly used ones to select terms in documents. Term frequency selects most frequent terms. Document Frequency (DF) simply measures the number of documents in which a term appears. tf-idf weight is a combination of term frequency and document frequency to balance between term specialty and popularity. It is commonly used in information retrieval and text mining. χ^2 statistic (CHI) measures the divergence between a term and a category from the χ^2 distribution if one assumes the independence of the term and category. This measure is not reliable for extremely infrequent terms [Dunning 1993]. Information Gain (IG) chooses features with maximal information increment for classification. Mutual Information (MI) is the extra bits required to differentiate two random variables X and Y if their joint distribution is given. [Yang and Pedersen 1997] show that χ^2 statistic and IG perform better than the others. Bi-Normal Separation (BNS) compares probabilities of a feature appearing in positive and negative classes. It outperforms other measures when class distribution is highly imbalanced [Forman 2003; Tang and Liu 2005]. Since one group is often relatively small compared with a network, BNS is adopted in this work. Of course, alternative measures mentioned before may be used or developed for group profiling.

Another line of research relevant to group profiling is to extract annotations from relational data with text. For instance, [Roy et al. 2006] construct a hierarchical structure as well as corresponding annotations based on a complicated generative process. The model complexity and scalability hinder its application to large-scale networks. [Chang et al. 2009] propose NUBBI (Networks Uncovered By Bayesian Inference) to infer descriptions of entities in a text corpora as well as relationships between these entities. 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.

Other research extends topic models to extract groups based on network and text information together. Conventionally, a collection of documents are modeled as a set of latent topics, and each topic represents a distribution of words. Link-LDA [Erosheva et al. 2004] treats citations of papers the same way as normal words, that is, the citation is generated based on a multinomial distribution over documents. Pairwise Link-LDA [Nallapati et al. 2008] essentially combines the topic model [Blei et al. 2003] and the mixed membership stochastic block model [Airodi et al. 2008] by sharing the same latent mixture of communities for both word topics and relation

topics. Link-PLSA-LDA [Nallapati et al. 2008] 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. 2008] treat connections between documents in a different fashion. They enforce the connected documents to share similar topics and use the network information as regularization to extract topics. Topic-Link LDA [Liu et al. 2009] models the probability of connections between two nodes as depending on their similarities in terms of both latent topics and latent community memberships.

These works differ from group profiling as they aim to extract latent topics of a collection of documents, while group profiling aims to extract representative attributes that are descriptive of a given group. After extracting topics, the question of which topic or which words from the topics should be chosen to represent the given group remains unanswered. However, we agree that the two approaches are relevant to some extent. For instance, the group-profiling techniques discussed here can be applied to select topics for each group as well.

8. CONCLUSIONS AND FUTURE WORK

In social media, users form implicit communities by interaction. It is intriguing to understand the formation of these social structures. In some real-time social Web sites such as Twitter, a transient crowd may form in a short time [Kamath and Caverlee 2011]. A clear understanding of the burgeoning communities may help in cultural modeling and trend detection. The group profiling discussed in this work is one technique to find out the likely reason that causes all people of a community to connect to or interact with each other.

In this work, we adopt a group-profiling approach to extract descriptive features for a given group. Different group-profiling strategies are investigated. A natural approach would be aggregating individual attributes and considering which attribute is shared most frequently inside a group. This method has been commonly used in plotting tag clouds in social media. We found that aggregating individual attributes is applicable only in a relatively noise-free environment. But if profiles are constructed from noisy attributes, such as user blog posts or self-reported interests, differentiation-based methods, which differentiate a group from either the global network or only its neighbors, consistently outperforms the aggregation-based approach. More interestingly, an egocentric view for group profiling works as well as a global view. That is, by selecting those attributes that differentiate a group from their 1-hop away neighbors, we are able to construct reasonably good profiles. This fact suggests that we can simply examine those actors that are 1-hop away from the group to understand a particular group, which can be very efficient in navigating a large-scale network with numerous communities.

This work is a solid yet initial study of group profiling. Many extensions of group profiling can be explored. We list some directions that merit further research: (1) In a dynamic network environment, communities evolve. They can grow, merge, split, or even dissolve. We expect that group profiling can be used to understand group evolution and capture group interactions and relationships. It is also imperative to find out groups and their profiles simultaneously facing a stream of interaction information. Some preliminary work following this line has been published [Wang et al. 2010a]. (2) Another direction is joint modeling of group influence and profiles. In Section 5.4.2, we investigated the internal structure of a group and study how that affects the group-profiling performance. More research following this line will be appreciated. A related problem is to identify actors that play a critical role in group formation or swaying group opinions [Agarwal et al. 2008]. On the other hand, [Watts and Dodds 2007] challenge the conventional influential hypothesis. They suggest that many so-called "influentials" in social media are actually accidental. It remains unclear which group

profiles are determined by few actors, and which group profiles are determined by the majority. (3) In current work, we propose to understand emerging social structures based on group profiles. Is it possible to zoom into finer levels, say, each connection? Given a network and related node attributes or pairwise interaction information, can we label the connections to explain why two nodes are connected? Of course, this is a much more difficult task than group profiling. Hopefully, with the explosion of user-generated content and mountains of interaction information, we may be able to make sense of social networks on a scale that has never been achievable before.

ACKNOWLEDGMENTS

We thank Reza Zafarani for crawling the LiveJournal Data. We are grateful to Dr. John Salerno for insightful discussions and suggestions.

REFERENCES

AGARWAL, N., LIU, H., TANG, L., AND YU, P. S. 2008. Identifying the influential bloggers in a community. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08)*. ACM, New York, 207–218.

AIRODI, E., BLEI, D., FIENBERG, S., AND XING, E. P. 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res. 9*, 1981–2014.

ALMACK, J. 1922. The influence of intelligence on the selection of associates. *School Soc. 16*, 529–530.

BACKSTROM, L., HUTTENLOCHER, D., KLEINBERG, J., AND LAN, X. 2006. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM, New York, 44–54.

BAUMES, J., GOLDBERG, M., MAGDON-ISMAIL, M., AND WALLACE, W. 2004. Discovering hidden groups in communication networks. In *Proceedings of the 2nd NSF/NIJ Symposium on Intelligence and Security Informatics*.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

BOTT, H. 1928. Observation of play activities in a nursery school. *Genetic Psychol. Monographs 4*, 44–88.

CHAKRABARTI, D. AND FALOUTSOS, C. 2006. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv. 38,* 1, 2.

CHANG, J., BOYD-GRABER, J., AND BLEI, D. M. 2009. Connections between the lines: Augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*. ACM, New York, 169–178.

CHEN, W.-Y., CHU, J.-C., LUAN, J., BAI, H., WANG, Y., AND CHANG, E. Y. 2009. Collaborative filtering for orkut communities: Discovery of user latent behavior. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, New York, 681–690.

DUNNING, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist. 19,* 1, 61–74.

EROSHEVA, E., FIENBERG, S., AND LAFFERTY, J. 2004. Mixed-Membership models of scientific publications. *Proc. Nat. Acad. Sci. 101,* 90001, 5220–5227.

FIORE, A. T. AND DONATH, J. S. 2005. Homophily in online dating: When do you like someone like yourself? In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, 1371–1374.

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

GUYON, I. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

HECHTER, M. 1988. *Principles of Group Solidarity*. University of California Press.

HOPCROFT, J., KHAN, O., KULIS, B., AND SELMAN, B. 2004. Tracking evolving communities in large linked networks. *Proc. Nat. Acad. Sci. U.S.A. 101,* 1, 5249–5253.

JACCARD, P. 1901. tude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Socit Vaudoise des Sciences Naturelles 37*, 547–579.

JONES, K. S. 1972. A statistical interpretation of term specificity and its application in retrieval. *J. Document. 28,* 1, 1121.

KAMATH, K. AND CAVERLEE, J. 2011. Transient crowd discovery on the real-time social web. In *Proceedings of the ACM Intenational Conference on Web Search and Data Mining (WSDM)*.

KELLY, J. AND ETLING, B. 2008. *Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere*. Berkman Center for Internet and Society at Harvard University.

KENDALL, M. 1938. A new measure of rank correlation. *Biometrika 30*, 81–89.

KUMAR, R., NOVAK, J., AND TOMKINS, A. 2006. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM, New York, 611–617.

LAUW, H. W., SHAFER, J. C., AGRAWAL, R., AND NTOULAS, A. 2010. Homophily in the digital world: A livejournal case study. *IEEE Internet Comput. 14*, 15–23.

LESKOVEC, J., KLEINBERG, J., AND FALOUTSOS, C. 2007. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data 1,* 1, 2.

LESKOVEC, J., BACKSTROM, L., KUMAR, R., AND TOMKINS, A. 2008a. Microscopic evolution of social networks. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, 462–470.

LESKOVEC, J., LANG, K. J., DASGUPTA, A., AND MAHONEY, M. W. 2008b. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, 695–704.

LIU, H. AND MOTODA, H. 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

LIU, Y., NICULESCU-MIZIL, A., AND GRYC, W. 2009. Topic-Link lda: Joint modeling of topic and community for blog analysis. In *Proceedins of the 26th International Conference on Machine Learning*.

MACDONALD, C., SANTOS, R. L., OUNIS, I., AND SOBOROFF, I. 2010. Blog track research at trec. *SIGIR Forum 44*, 1, 57–74.

MCPHERSON, M., SMITH-LOVIN, L., AND COOK, J. M. 2001. Birds of a feather: Homophily in social networks. *Ann. Rev. Sociol. 27*, 415–444.

MEI, Q., CAI, D., ZHANG, D., AND ZHAI, C. 2008. Topic modeling with network regularization. In *Proceeding of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, 101–110.

NALLAPATI, R. M., AHMED, A., XING, E. P., AND COHEN, W. W. 2008. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, 542–550.

NEWMAN, M. 2003. The structure and function of complex networks. *SIAM Review 45*, 167–256.

PALLA, G., BARABASI, A.-L., AND VICSEK, T. 2007. Quantifying social group evolution. *Nature 446,* 7136, 664–667.

ROY, D. M., KEMP, C., MANSINGHKA, V. K., AND TENENBAUM, J. B. 2006. Learning annotated hierarchies from relational data. In *Proceedings of the Conference on Naval Information Processing System (NIPS)*. 1185–1192.

SHI, X., CHANG, K., NARAYANAN, V. K., JOSIFOVSKI, V., AND SMOLA, A. J. 2010. A compression framework for generating user profiles. In *Proceedings of the ACM SIGIR Workshop on Feature Generation and Selection for Information Retrieval*.

SHMUELI-SCHEUER, M., ROITMAN, H., CARMEL, D., MASS, Y., AND KONOPNICKI, D. 2010. Extracting user profiles from large scale data. In *Proceedings of the Workshop on Massive Data Analytics on the Cloud (MDAC'10)*. ACM, New York, 4:1–4:6.

SUN, J., FALOUTSOS, C., PAPADIMITRIOU, S., AND YU, P. S. 2007. Graphscope: Parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*. ACM, New York, 687–696.

TANG, L. AND LIU, H. 2005. Bias analysis in text classification for highly skewed data. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05)*. IEEE Computer Society, 781–784.

TANG, L. AND LIU, H. 2009. Scalable learning of collective behavior based on sparse social dimensions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. 1107–1116.

TANG, L. AND LIU, H. 2010. Toward predicting collective behavior via social dimension extraction. *IEEE Intell. Syst. 25*, 19–25.

TANG, L., LIU, H., ZHANG, J., AGARWAL, N., AND SALERNO, J. J. 2008. Topic taxonomy adaptation for group profiling. *ACM Trans. Knowl. Discov. Data 1,* 4, 1–28.

TANTIPATHANANANDH, C., BERGER-WOLF, T., AND KEMPE, D. 2007. A framework for community identification in dynamic social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, New York, NY, 717–726.

THELWALL, M. 2009. Homophily in myspace. *J. Amer. Soc. Inf. Sci. Technol. 60,* 2, 219–231.

WANG, X., TANG, L., GAO, H., AND LIU, H. 2010a. Discovering overlapping groups in social media. In *Proceedings of the 10th IEEE International Conference on Data Mining*.

WANG, Y., CONG, G., SONG, G., AND XIE, K. 2010b. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. ACM, New York, 1039–1048.

WASSERMAN, S. AND FAUST, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

WATTS, D. AND DODDS, P. 2007. Influentials, networks, and public opinion formation. *J. Consum. Res. 34,* 4, 441–458.

WELLMAN, B. 1926. The school child's choice of companions. *J. Educ. Res. 14*, 126–132.

YANG, Y. AND PEDERSEN, J. O. 1997. *A Comparative Study on Feature Selection in Text Categorization.* Morgan Kaufmann Publishers, 412–420.

Received September 2010; revised January 2011; accepted April 2011