Privacy Leakage in Event-based Social Networks: A Meetup Case Study

TAEJOONG CHUNG*, Northeastern University
JINYOUNG HAN, Hanyang University
DEAJIN CHOI, Seoul National University
TED "TAEKYOUNG" KWON, Seoul National University
JONG-YOUN RHA, Seoul National University
HYUNCHUL KIM, Sangmyung University

Event-based social networks (EBSNs) are increasingly popular since they provide platforms on which online and offline activities are combined. Despite the increasing interest in EBSNs, little research has paid attention to the privacy issues coming from the unique features of EBSNs; the on-site information of users is highly relevant to real lives. In this paper, we try to investigate privacy leakages in Meetup, one of the most popular EBSN service. More specifically, we answer what private information can be inferred from the site's publicly available data. To this end, we conduct a measurement study by crawling webpages from `Meetup` containing 240K groups, 8.9M users, 27M group affiliations and 78M topical interests. By analyzing the dataset, we find that LGBT status of users, which is one of the most sensitive privacy information, can be predicted with 93% accuracy. Finally we discuss the cause of the privacy leakage on EBSNs and its possible ensuing damages.

ACM Reference format:
Taejoong Chung*, Jinyoung Han, Deajin Choi, Ted "Taekyoung" Kwon, Jong-Youn Rha, and Hyunchul Kim. 2017. Privacy Leakage in Event-based Social Networks: A Meetup Case Study. *Proc. ACM Hum.-Comput. Interact.* 1, 2, Article 35 (November 2017), 22 pages.
https://doi.org/10.1145/3134670

1 INTRODUCTION

Recent years have seen the increased popularity and rapid growth in usage of event-based social networks (EBSNs), such as `Meetup`[1], `Eventbrite`[2], and `Douban`[3]. Their support for connecting people by their interests around offline social events or activities is one of the major sources of the rapid growth; EBSNs not only support typical online social interactions

[1]http://www.meetup.com/
[2]http://www.eventbrite.com/
[3]http://www.douban.com/

*This work was done when the author was a postdoctoral researcher at Seoul National University.
Corresponding Author: Jinyoung Han (jinyounghan@hanyang.ac.kr).

2573-0142/2017/11-ART35
https://doi.org/10.1145/3134670

as in other online social networks (OSNs), but also provide convenient and easy-to-use online platforms for users to organize, find, share, participate, comment, and recommend offline social events as well, such as cocktail parties, musical concerts, business meetings or political manifestations. To date, many of these services have had a large number of people subscribed, rapidly growing in business. For example, as of Apr 2017, Meetup has 30.3 M users, creating 600 K events every month [35].

The rise of EBSNs opens up new avenues of research, as it allows researchers to access unprecedentedly large scale social data, which details both online social interactions and offline activities, for understanding how online interactions in EBSNs affect offline activities. For example, recent studies have revealed valuable insights into offline group behaviors [12, 20, 30, 45, 54], such as cohesive structural property of EBSNs [30, 54] (compared to conventional OSNs), effect of offline interactions to the online community and social ties [46], social composition of people in offline events [45], relations between OSNs and offline attendance behaviors [20], prediction of offline activity attendances [12], or offline event recommendation [42].

However, most of these studies paid little attention to privacy leakage issues in EBSNs, which may be attributed to their unique property. That is, user information that can be gathered from the EBSN sites, whether willingly disclosed or unintentionally revealed, may disclose the details of users' private interests or lives. For example, since the offline meetings are arranged on the basis of group affiliations, a user's personal interests can be inferred by analyzing the information about his/her group affiliations. Note that offline meetings or activities usually require 'accurate' and 'detailed' user information, e.g., via questionnaires when joining a group, a user's private information can be inferred with much accurate data. Therefore, we believe that understanding potential privacy threats in EBSNs can shed light on designing secure EBSN platforms for protecting an end-user's privacy.

This paper investigates such privacy issues in `Meetup`, one of the most popular EBSN that has 30.3 M members across 180 countries, and 270K groups (as of Apr, 2017). People who want to share their interests such as politics, books, religions, or even sexual identities (e.g., lesbian, gay, bisexual, transgender (LGBTs)) can make or join a group for offline activities. By analyzing the datasets obtained from `Meetup`, we seek to address the following question that have not been thoroughly investigated to our knowledge: *Can user's sensitive interests, such as LGBT status, which are not publicly disclosed, be revealed when participating in online* `Meetup` *groups? If so, how easily private information can be inferred?* To answer the above questions, we analyze privacy leakage problems in `Meetup` using the collected dataset from January 28, 2014 to August 18, 2014, by crawling *publicly accessible* web pages in `Meetup`. The dataset consists of 240 K groups (89% of total groups), which includes 8.9 M users, 27 M group affiliations, and 78 M topical interests. In particular, we focus on *how a user's LGBT status can be inferred based on simple machine-learning techniques*, which may be one of the most sensitive personal traits [18].

We highlight our key findings of this paper as follows:

- **Measurement:** To our knowledge, this is the first large-scale measurement study to comprehensively investigate how privacy information can be leaked easily in Meetup, one of the most popular event-based social network services.
- **Key Findings:** We find that a user and his/her affiliated group(s) share similar topical interests, which suggest that an user's trait can be inferred by observing his/her affilated groups. We also show that 75% of users' group affiliation information can be inferred *thoroughly* by investigating group membership pages. We reveal that a

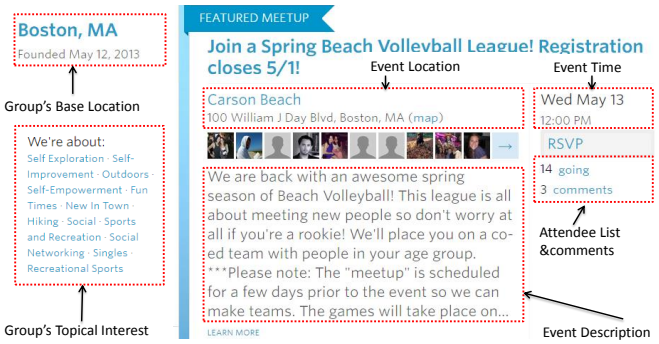


Fig. 1. An illustration of an event page of a group in Meetup.

user's LGBT status, which is one of the most sensitive personal information, can be accurately inferred (93% accuracy) by considering (i) the user's group affiliation, (ii) topical interests, (iii) vocabulary usage patterns, and (iv) networking pattern with friends in other OSNs (e.g., Facebook, Twitter).

2 BACKGROUND

2.1 Meetup Overview

Meetup [7, 41, 45] is an EBSN that entwines online interactions and offline activities. The main function of Meetup is to let people meet together offline depending on their interests, by allowing users to create groups so that people sharing similar interests can meet together offline. Users in Meetup can search and join groups based on their interests. Figure 1 illustrates an event page of a group in Meetup. We now describe the main components in Meetup as follows.

- User's Profile/Topical Interests: Each user has a profile page. A user is required to register his/her location information by specifying a postal code so that Meetup can recommend groups based on the location. For the similar reason, Meetup requires a user to choose a set of topical interests from a pool of topics; the total number of topics in Meetup exceeds 100 K recently. Meetup allows a user to make the information of his/her group affiliation and/or topical interests public or not.
- Group/Category: A user can make a group as a group organizer. A group belongs to one of the 33 categories provided by Meetup, varying from "Arts and Culture" to "Movement and Politics" to "LGBTs (Lesbian, Gay, Bisexual, and Transgender)". The list of categories is described in Table 1. Each group also specifies a set of topical interests from the same pool of topics that users have. Some groups require a user to answer a questionnaire such as the self-introduction or motivation of joining them, and the organizer can choose to make list of members of the group public, as well as the questionaires and answers of the members.
- Event: In a group, only the group organizer can hold an offline event by specifying the title, description, exact location (which is supported by Google map), and time. A member in a group can express his/her intention to aitend the event via an RSVP function.

2.2 Related Work

Online Communities for Offline activities: EBSNs such as `Meetup` and `Douban` have been increasingly popular by providing an online playground that helps people sharing similar interests meet together offline. While traditional OSNs mostly focuses on social interaction in online spaces, EBSNs aim at supporting offline activities. Hence, the online space of EBSNs mostly focuses on supporting users sharing similar interests to have offline activities easily. This in turn has led many researchers to investigate online/offline behaviors of users in EBSNs [9, 12, 20, 30, 45, 54]. Sander showed that most of social interactions in offline events tend to occur among friends rather than strangers [45]. Liu *et al.* identified communities among members based on their co-attendances to the same meeting events, and showed that the structural pattern of the co-attendance network tends to be more cohesive than those of other OSNs such as location-based social networks [30]. Xu *et al.* investigated how offline events affect online social interactions in `Douban` [54]. On the other hand, Han *et al.* examined how online social relationships affect the attendance to offline events in `Douban` [20]. Du *et al.* proposed a method to predict the attendances of users by considering both online and offline factors such as user's preference or the spatial/temporal information of an event in `Douban` [12]. Cranshaw *et al.* suggested a method to infer online social relationships by analyzing location trails of users using Locaccino [44], which is a web-application sharing their location through Facebook [9]. Our work complements the previous work, as they mostly focused on user behaviors in EBSN. Instead, we focus on how users' private information like LGBT status can be leaked in ESBNs.

Inferring a user's LGBT status: There have been studies how to infer a user's sensitive data such as gender, age, or sexual orientation using available information in OSNs, like social relationships among users [8, 11, 24] or user generated contents [33, 52]. Dey *et al.* estimated the birth years of users by exploiting the underlying social network structures such as ages of friends or ages of friends of friends [11]. Jernigan *et al.* estimated the LGBT status

Table 1. Meetup categories with indexes are summarized.

1	Arts and Culture	18	Movies and Film
2	Career and Busineess	19	Music
3	Cars and Motorcycles	20	New Age Spirituality
4	Community and Environment	21	Outdoors and Adventures
5	Dancing	22	Paranormal
6	Education and Learning	23	Parents and Family
7	Fashion and Beauty	24	Pets and Animals
8	Fitness	25	Photography
9	Food and Drink	26	Religion and Beliefs
10	Games	27	Sci-Fi and Fantasy
11	LGBT	28	Singles
12	Movements and Politics	29	Socializing
13	Health and Wellbeing	30	Sports and Recreation
14	Hobbies and Crafts	31	Support
15	Language and Ethnic Identity	32	Tech
16	Lifestyle	33	Women
17	Literature and Writing		

of a user using a portion of the LGBT friends within her friendship network in Facebook [24]. Casas *et al.* identified the user's gender using both his/her generated content and social network information in Google+ [8]. Kosinski *et al.* suggested a method to identify users' attributes such as their marriage status using Facebook 'Like' button logs acquired from volunteers [26]. While these studies used various online records for inferring a user's sensitive privacy, our work focuses on privacy issues in `Meetup` where online interactions and offline activities are combined.

Privacy leakage in OSNs: As OSNs have become a part of our daily lives, there have been attempts to study privacy leakages in various OSNs, such as home location, profession, personality, etc. For instance, the information of geographical coordinates recorded on OSNs can be used in inferring a user's location information [39, 40]. Pontes *et al.* inferred the home city (or state, country) of a user from publicly available information in Foursquare [39], which is one of the most popular location-based social networks. They showed that the location where a user has visited most frequently is highly likely to be in his/her home city or state. Popescu *et al.* analyzed photos having geographical tag information in Flicker and estimated home locations of users [40]. Instead of using direct information such as the geo-tagged information, some studies have considered user generated content or social relationships in OSNs to estimate the locations of users [1, 5, 10, 22, 27, 32, 36]. Twitter messages containing gazetteer terms or popular place names can be used to estimate the home location of the sender [5, 22]. Mahmud *et al.* considered temporal characteristics of posting times in Twitter to estimate user locations [32]. Backstrom *et al.* used social relationships in OSNs in estimating user locations; spatial proximity characteristics between of Facebook friends are used [1]. Clodoveu *et al.* inferred a user's location using reciprocal relationships (i.e., following-followee) in Twitter [10], based on the observations of the previous work that a user who has reciprocal relationships with less than 2,000 friends is likely to be geographically close to her friends [27]. Some researchers developed models for predicting users' personality using diverse attributes of users observed in social media [15, 23, 43, 52]. For example, Wagner *et al.* showed that a user's profession or her personality can be identified based on his/her Twitting behaviors or attributes [52]. Golbeck *et al.* used personality data from 335 Twitter users and found that users' personality can be accurately predicted based on the publicly available information on their Twitter profiles. Elena *et al.* also used friendship information to infer sensitive attributes [13], but the group is defined implicitly based on the densification of friendship network. Lu *et al.* focused on the potential system's flaws in the social platform; for example, they showed that side channels such as newsfeeds, tagging, relationship status in Facebook can reveal privacy-sensitive information to users through indirect mechanisms [31].

Our work complements these, as they are mostly focused individual social network. Instead, we focus primarily on the group based social network, which doesn't require to have individual relationship between each users. Moreover, the dataset we use is considerable broader, including 8.9 M users and their 27 M affiliations.

Homophily in Social Networks: Homophily is a tendency of individuals to group together with similar others [34]. Homophily is often observed in many areas including online social networks [3, 27, 37, 48], online dating platforms [14, 47], and even in games such as MMORPGs (Massively Multiplayer Online Role-Playing Games) [6]. For example, Chat *et al.* investigated homophily phenomena in an online social curation site, Pinterest, which showed that homophily drives sharing content: people share content from other users who share their interests and follow users who have similar interests [4]. However, homophily sometimes can be used to infer personal traits or interests; for example, Guha *et al.* examined

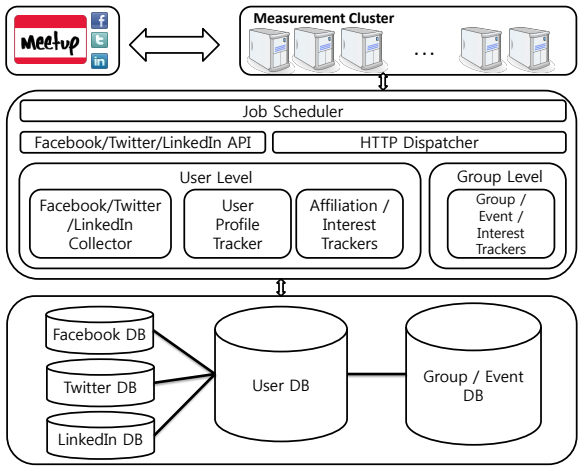


Fig. 2. The architecture of our Meetup crawling and analysis system is depicted.

homophily in friendship and surveillance networks for users of the Foursquare service and found a potential privacy leakage through social surveillance in these networks [19]. Similarly, Jenigar *et al.* tried to predict LGBT status based on the insight where gay users tend to have more gay friends [24].

In our study, we also apply homophily in a group to track personal traits; in other words, as a Meetup online group is formed based on the similar interests shared by members (*i.e.,* exhibiting *homophily*), we try to infer personal traits from other people who have similar interests.

3 METHODOLOGY

In this section, we explain our measurement methodology for data collection, and describe the dataset we used in this paper.

3.1 Data Collection

Since Meetup does not provide an official API for data collection, we developed our crawling and analysis system as shown in Figure 2. We fetched web pages in Meetup, from which the relevant information is extracted; the information of a group or a user can be extracted from a web page. Since a large number of web pages need to be collected from Meetup, we design the crawling framework in a distributed manner. Our measurement cluster consists of 25 PCs to which our job scheduler assigns tasks. Each of PCs sends HTTP requests continuously at moderate rate with a random sleep. The HTTP dispatcher processes the HTTP requests and responses according to the assigned tasks.

There are two main components in our measurement system: *group seeker* and *user seeker*. Meetup provides a portal to allow a user to search groups in terms of geographical distances and categories. By setting the geographical distance to 'Any Distance' and the category to 'All Meetups', we can find all groups. Our group seeker keeps searching the portal periodically (every 10 minutes), not to miss any new groups. For each of the groups we found, we obtained the relevant group information such as the category, group location, event histories containing all of the attendee lists, event location/time, member lists, and so on by fetching the web pages of groups. Some groups have questionnaires, which are to be

answered by users who wish to join the groups. As some of the answers of the questionnaires are open to *public*, we analyze the answers to extract users' traits like the motivation for joining a group.

We collected user lists with two different ways: (i) extracting members from individual group pages and (ii) crawling the user information from individual user pages by the user seeker. Our user seeker searched a new user using a breath first search (BFS) with feeds from new users who have sent messages in their guestbooks. We noticed that a user's profile URL has a fixed format as 'http://www.meetup.com/members/*user_id*', where *user_id* is an integer number. Our user seeker keeps discovering a new user by increasing *user_id* by 1 K.

For the discovered 8,943,065 users, we fetched their profiles containing their names, location information, the dates they joined `Meetup`, messages they have received in their guestbooks[4], their group affiliation information, and their interests, if they are open to public. `Meetup` also allows a user to connect her account to five major OSNs: Facebook, Twitter, Tumblr, Flickr, and LinkedIn. Thus, we further collected a user's public (i) Facebook profile containing his/her gender, country, and friends list (if available), (ii) Twitter profile containing her description, location, and following/followee list, and (iii) LinkedIn profile.

3.2 Dataset

Our dataset had been collected for 202 days from January 28, 2014 to August 18, 2014. For the groups measured in this period, we fetched all of their event histories from their births. The oldest event was held at April 17, 2002. We kept track of 209,412 public groups out of 241,197 groups for the 33 categories, which contain 8,877,653 events, 1,351,361 photos, 26,025,179 messages among members, 40,384,144 questions and their corresponding answers. Among the discovered 8,943,065 users, we also fetched their profiles from the other OSNs if they connected their accounts to Facebook, Twitter, and/or LinkedIn; we obtained 625,017 Facebook profiles and their 3,834,431 friend pairs, 298,533 Twitter profiles and their 1,993,950 follower pairs[5], 2,053,580 followee pairs, and 170,287 LinkedIn profiles. Top five countries in terms of the number of users and groups in our datasets are United States (70.5% in users and 69.6% in groups), United Kingdom (6.2% in users and 6.0% groups), Canada (5.7% in users and 5.5% in groups), Australia (3.4% in users and 3.9% in groups), and India (1.7% in both users and groups).

3.3 Ethics

Our methodology brings up a few ethical issues, and we would like to discuss them explicitly before showing our results. We first note that we only use "publicly accessible" information which can be obtained from `Meetup`, hence no authorizations are required to obtain them. We also note that the users recognize that their public information may be accessible or made public[6]; users can make their information (e.g., personal interests, location, or etc.) private if they do not want to reveal them. Furthermore, it is worth to note that we seriously take care of sensitive user information in our dataset. We use a hashing technique to encrypt personal identifiers (i.e., name), and make hard to track them with reverse engineering. In our data collection, we try to minimize our impact on `Meetup` service and its users, by scraping data with moderate rate (approx. 0.5 query/sec) of HTTP requests.

[4]As of Apr 2017, this function is changed to private conversations, and hence messages are not available any more.
[5]A follower pair means that both a user and her follower have `Meetup` accounts.
[6]http://www.meetup.com/terms/

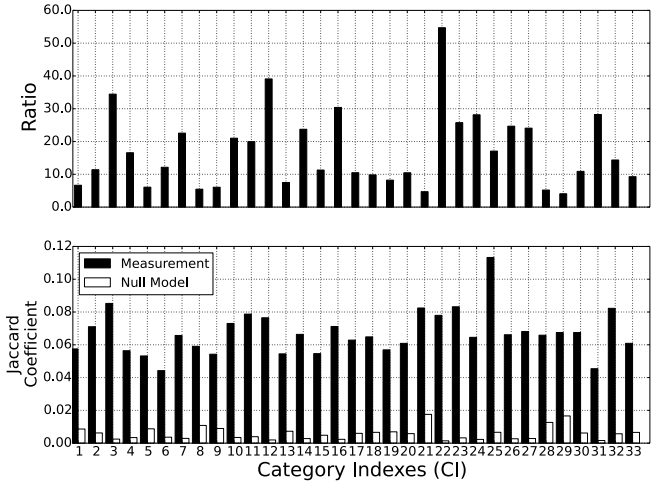


Fig. 3. Jaccard coefficient of the interests of a group and its affiliated users is much higher than that of a null model.

4 GROUP: CARRIERS OF INDIVIDUAL'S INFORMATION

In this section, we investigate how groups in `Meetup` are associated with their affiliated users' characteristics in terms of topical interests.

4.1 Interest Similarity: A Group and Its Affiliated Users

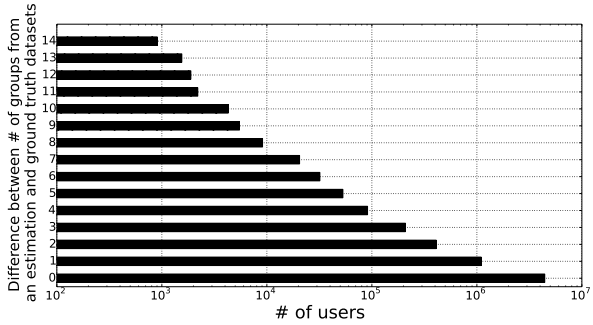
`Meetup` provides a set of topical interests (over 100 K), such as baseball and Indian food. From the interest pool, a group can specify its interests so that a user can search groups with her interests. A user in `Meetup` is also required to choose a set of topical interests from the *same* interest pool.

We first try to understand how a group and its members share similar interests. Our intuition behind this is that if the topical sets of users and the group which they affiliated with are similar, we might be able to infer users' hidden interests by looking the groups they are affiliated with.

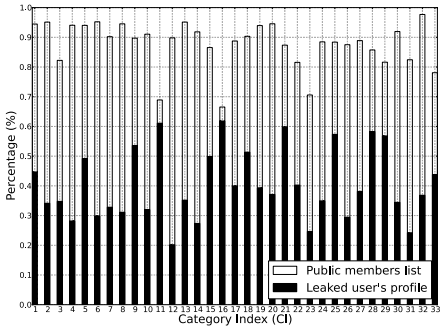
To this end, we calculate the Jaccard coefficient between the interest sets of a group and its member who make their profiles public. We denote the interest sets of a group k and of a user n by $I(G_k)$ and $I(U_n)$, respectively. Then the Jaccard coefficient between a group k and a user n is $J(G_k, U_n) = \frac{I(G_k) \cap I(U_n)}{I(G_k) \cup I(U_n)}$.

Figure 3 shows the average Jaccard coefficient (referred to as *Measurement*) of all the groups and their corresponding members across the 33 categories. For the comparison purposes, we also calculate the Jaccard coefficient between a group and randomly selected 100,000 users who are not in the group, which is referred to as *a Null Model*. We also plot the ratio of the two Jaccard coefficients in the right y-axis.

We find that the Jaccard coefficient of *Measurement* is much higher than that of *the Null Model*, which indicates that a user's choice of groups is highly related to the overlapping interests between hers and group's ones. We also find that the ratio of the two Jaccard coefficients is substantially different across the different categories. For example, the coefficient ratio of "Movements and Politics" (Category Index (CI) 12) is 39.09, meaning that a portion of common interests between the group and its affiliated members is 39.09 times higher



(a) Difference of the number of groups between the backtracked dataset and ground truth one is plotted. Note that x axis is in log scale and the affiliation information of 6.7 M users (75.34%) can be *completely* identified by showing no differences between them.



(b) The portion of a group having public member list and its leakage is plotted.

Fig. 4. Private interests such as LGBTs, religion, and political attitude of a user can be leaked by collecting the information of her affiliated groups.

than the one between the group and non-affiliated users. The top 5 categories in terms of the coefficient ratio are "Cars and Motorcycles" (CI 3), "Movements and Politics" (CI 12), "Paranormal" (CI 22), "Hobbies and Crafts" (CI 14), "Support" (CI 31) which indicates that members in those groups have highly similar interests with their groups. The coefficient ratio of "LGBT" (CI 11) also shows a high similarity of interests (20.22) between users and groups. On the other hand, the bottom 5 categories in terms of the coefficient ratio are "Fitness" (CI 8), "Food and Drink" (CI 9), "Outdoors and Adventures" (CI 21), "Singles" (CI 28), and "Socializing" (CI 29), which are more general topics compared with the top 5 ones.

4.2 Membership Information Leakage in Meetup

Since groups in `Meetup` are used for sharing interests with offline activities, personal data of users can be found or inferred from their affiliated group pages as we saw in previous analysis. For example, an administrator of a group can make their members' profiles public in its group page so that any user can find the members of the group. In this case, even

if a member makes her profile that includes her group affiliation private, her affiliation to the group is obtained easily. By crawling such information from group pages, we obtain the group affiliation information of 247,517 users, which is not available in their profile pages. Note that a user can make her profile public or private. Likewise, a group can make its member list public or private.

We validate how a user's group affiliation information obtained from the group pages is similar to her actual affiliation information in her profile page as shown in Figure 4(a). To this end, we consider only users who make their group affiliation information public in their profile pages, and compare the information with the one acquired from the group pages. Surprisingly, as shown in Figure 4(a), the affiliation information of 6.7 M (75.34%) users can be *completely* identified. 8.46 M (95.00%) users show the differences of up to two different groups between the groups from the profile pages and the ones from the group pages.

Since a group's interests are closely related to their members' interests as shown in Figure 3, the exposure of a user's group affiliation may bring up a severe privacy issue such as her sensitive interests. To quantify such a potential privacy problem, we measure (i) how many groups open their members' profiles to the public and (ii) how many members hide the group affiliation information on their profiles in each group, for each category in Figure 4(b). Figure 4(b) shows portion of groups that have public member lists. Even for the sensitive categories "LGBTs" (CI 11), "Religion and Belief" (CI 26), and "Movements and Politics" (CI 12), we find that the portion of groups having public member lists are 68.9%, 87.51%, and 89.82% respectively. We also find that many groups have privacy leakage problems as they open their lists of members even if their members do not make their profiles public. We observe that 61.1% of "LGBT" (CI 11) groups have privacy leakage issues, which might raise a serious problem such as an unintentional outing.

Discussion: Disclosure at Individual Level vs. at Group Level. As shown in Figure 4(a), group affiliation information of a user, who has not made her affiliation information private, can be inferred by back-tracking of publicly disclosed member lists. This happens because group organizers in `Meetup` have the authority to disclose member lists to public at their discretion. To further explore this issue, we analyze the member information disclosure behavior of the *group organizers* in `Meetup`. As shown in Table 2, 90% of the organizers disclose their group profiles to public, and 96% among them have their member list open to public. The default setting for profile in `Meetup` is public, which indicates that 10% of the organizers intentionally hide their profiles. However, 93% among those who change their profiles so that their information were hidden to public make their member lists public, which shows the inconsistency in their behaviors. Such a discrepancy might be due to the conflicted position of the organizers. On the one hand, the organizer might consider that the member information of the group is sensitive and thus want to keep it private (or safe). On the other hand, the organizer might want to make his/her group popular, hence want to disclose member information (e.g., gender or self-introduction) for marketing purposes in hope that new members could be recruited. To remedy this problem, we believe that providing an abstract members information (such as gender or locality distribution) or testimonials can satisfy the organizers both in privacy protections of members and marketing purposes to recruit new users. As an administrator of a group in other event-based social networks (e.g., Douban) also has an authority to control the visibility of membership information, we believe that the same problem and remedy can be applied to other platforms as well.

Table 2. The large portion of organizers of groups who hide their profile, made member's list public

		Organizer's Profile	
		Public (90%)	Hidden (10%)
Members' List	Public	286370 (96%)	22984 (93%)
	Hidden	12272 (4%)	1776 (7%)

Table 3. 32 Categories of psychological words in LIWC and their examples

Category	Examples	Category	Examples	Category	Examples
social	mate, talk, they, child	anger	hate, kill, annoyed	inhibition	block, constrain, stop
body	cheek, hands, spit	family	daughter, husband, aunt	sad	crying, grief, sad
inclusion	and, with, include	health	clinic, flu, pill	friend	buddy, friend, neighbor
cognitive	cause, know, ought	exclusion	but, without, exclude	sexual	horny, love, incest
human	adult, baby, boy	insight	think, know, consider	percept	observing, heard, feeling
ingest	dish, eat, pizza	affect	happy, cried, abandon	cause	because, effect, hence
see	view, saw, seen	relative	area, bend, exit, stop	positive	love, nice, sweet
discrepancy	should, would, could	hear	listen, hearing	motion	arrive, car, go
negative	hurt, ugly, nasty	tentative	maybe, perhaps, guess	feel	feels, touch
space	down, in, thin	anxiety	worried, fearful, nervous	certain	always, never
bio	eat, blood, pain	time	end, until, season		

In the following section, we explore one of the privacy leakage problems in `Meetup`, predicting an user's LGBT status, which happens since the personal data of members in a group can be unintentionally leaked from the information of offline activities of the members.

5 LEAKAGE OF A USER'S LGBT STATUS

LGBT status is one of the most sensitive privacy information that should not be collected without an explicit consent [18]. In particular, LGBTs who have not disclosed their LGBT status often express profound fear and anxiety concerning what would happen to their relationship with their families, friends and colleagues, if their LGBT status gets accidentally revealed. A few recent studies have shown that people's LGBT status can be inferred from OSNs by inspecting their friend network in Facebook [24] or 'Like' history [26]. In this section, we examine how `Meetup` users' LGBT status can be leaked even if they make their LGBT status hidden using `Meetup`'s privacy settings. Unlike the prior work where only a single type of information was used such as friend networks or 'Like' histories, we use various attributes of users that we can obtain *publicly*, such as one's traits (i.e., topical interest), social networks, linguistic characteristics, affiliation information, in order to infer LGBT status.

5.1 Prediction on Sexual Interests

We observed the similarity between a group and its affiliated users by comparing their topical interests. Inspired from this, as a first step to test the feasibility of privacy leakage from group affiliations, we try to predict whether a user is interested in LGBT types of interests by considering only his affiliations. To this end, for all users, we classified the users into two groups; (i) the users who have LGBT types of interest on their interest sets (e.g., Gay, Lesbian, Bisexual, or Transgender, and etc.) and (ii) the users who do not have. As a result we obtained 120,844 users interested in LGBT types of interests.

For each user, we construct an affiliation vector of binary values, which we call *Group Affiliation* class, each of which indicates whether a user is affiliated with a group or not. It is worth noting that we *exclude* the groups whose categories is "LGBT" (CI: 11), since our focus is to predict whether a user is interested in LGBT types of interests or not *without* any

explicit indicators of the LGBT status. We also exclude the groups whose activities are related to *LGBTs* such as 'Gay Artist Groups', even if its category is 'Arts and Culture' (CI: 1) not "LGBT" (CI: 11). Finally, we obtain 237,361 groups which are not directly related to LGBTs amongst 241,197 groups; thus, the size of an affiliation vector is 237,361. We also construct an affiliation vector for each user who does not make her affiliation information public on her profile page by applying the backtracking technique to infer each user's affiliation as shown in Figure 4(a). Using the obtained vectors, we build a classification model using the Logistic Regression as a classifier. To resolve the class imbalance problem [21], where the performance of a learning model becomes severely low in the presence of underrepresented data (i.e., the class distribution is severely skewed), we compose the users who have LGBT types of interests and the others as 1:1 ratio to provide a balanced distribution [2]. It is worth noting that because of its very high dimensionality of vector sets (241,688 users and its 237,361 affiliations, which results in a 241,688 $\times$ 237,361 matrix), the dimensionality of the vector sets was reduced using singular-value decomposition [17].

Figure 5 shows the prediction accuracy of dichotomous variables expressed in terms of the area under the receiver-operating characteristic curve (AUC), which is equivalent to the probability of correctly classifying two randomly selected users, one from each class (i.e., a user who has LGBT interests and a user who has not), by changing the top k SVD components. We first observe that by considering only 100 top SVD components, we can correctly classify users with LGBT interests in 82% of cases. We also find out that even if we consider more than k SVD components (i.e. more than 200 components), the performance improvement of classifiers is marginal (84.1 $\sim$ 86.0%), which indicates that few core groups are directly related to LGBT types of interests, even though we remove the groups whose categories is "LGBT" (CI: 11). In the following subsection, instead of predicting whether a user has the LGBT types of interests or not, we try to directly predict a user's LGBT status (i.e., whether the user is a gay or not) to study the limit of the privacy leakage problem in EBSNs.

5.2 Predictions on LGBT Status

In this subsection, we try to predict user's LGBT status, which is one of the most sensitive privacy data [18, 51], using the only publicly available information. We first introduce the

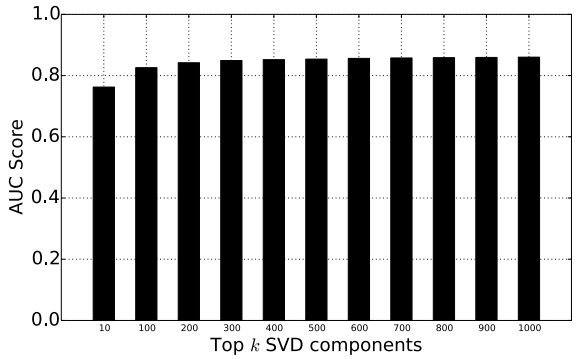


Fig. 5. AUC (Area Under Curve) scores are represented depending on top k SVD components using singular value decomposition (SVD)

ground truth dataset that contains user's LGBT status information, and then evaluate a machine learning-based model to infer a user's LGBT status.

5.2.1 Ground Truth Dataset. When users request to join a group, they typically are asked to fill out the group's questionnaire for the group organizer to decide whether to approve their request. We find that some groups keep their questionnaire and answers of the members publicly available, some of which are about their LGBT status. Among all the collected 243,149 questions and their 40,384,144 answers in our dataset, we find 342 questions are related to LGBT status, containing words such as 'sex', 'gay', 'transgender', 'lesbian', 'queer', 'bisexual', or 'orientation'. Then we also collect corresponding 35,549 answers. From those questions like "What is your LGBT status?", we can get users' answers, like "Male Gay". In this way, we obtain the ground truth dataset on LGBT status of 1,065, 654, 361, and 73 users who are answered their LGBT status as a gay, lesbian, bisexual, and transgender respectively; thus, there are total 2,153 LGBT people. In this study, we focus on predicting whether a user is gay or not, as they constitute a majority (around 50%) in our ground truth dataset.

We first investigate what categories of groups are more favored by gay users or non-gay users from our ground truth dataset. We collect the group affiliation information of gay and non-gay users, and calculate the probabilities of a gay and a non-gay user to be affiliated with a specific category in Figure 6. Interestingly, we observe the distinguishing patterns of group affiliation between gay and non-gay users; gay users are more affiliated with groups in the categories of "Lifestyle" (CI 16), "Literature and Writing" (CI 17), "Sci-Fi and Fantasy" (CI 27), and "Support" (CI 31). On the other hand, they seldom choose groups in the categories of "Career and Business" (CI 2), "Movements and Politics" (CI 12), "Parents and Family" (CI 23), "Singles" (CI 28), and "Tech" (CI 32). We also confirmed the different patterns in choosing topical interests between gay and non-gay users, even when we remove "LGBT" related topics (Due to the page limits, we omit their distributions). This suggests that different patterns of group affiliations could provide an important implication for predicting a user's LGBT status, to be detailed in the next section.

5.2.2 The Features Linked to LGBT Status. We first present the features linked to LGBT status, all of which are open to public. Table 4 summarizes the description of the features.

Group Affiliation: For each user, we construct an affiliation vector of binary values, each of which indicates whether a user is affiliated with a group or not based on 237,361 groups. It is worth noting that we construct an affiliation vector for each user who makes her affiliation information public on her profile page. For those who do not reveal their affiliation information on their profile pages, we apply the backtracking technique to infer their affiliations by investigating the member lists in each groups, and use them as a validation set which will be detailed in the following subsection.

Topical Interest: Like an affiliation vector, we build an interest vector that consists of binary values, each of which indicates whether a user has the particular interest or not. Similarly, we exclude the interests related to LGBT such as 'Gay', 'LGBT', etc. Finally, we obtain 98,990 topical interests.

Linguistic Characteristics: Some studies have shed lights on the differences of linguistic characteristics between LGBTs and non-LGBTs in speech patterns [29, 38] or in vocabulary usages [28]. To investigate whether online linguistic characteristics can also be a discriminator, we collect a corpus of comments generated by `Meetup` users. We obtain 30,138,548 and 26,004,737 comments written on event pages and users' guestbook pages from 2,743,163 and 1,768,385 users, respectively. To quantitatively measure the linguistic characteristics from the

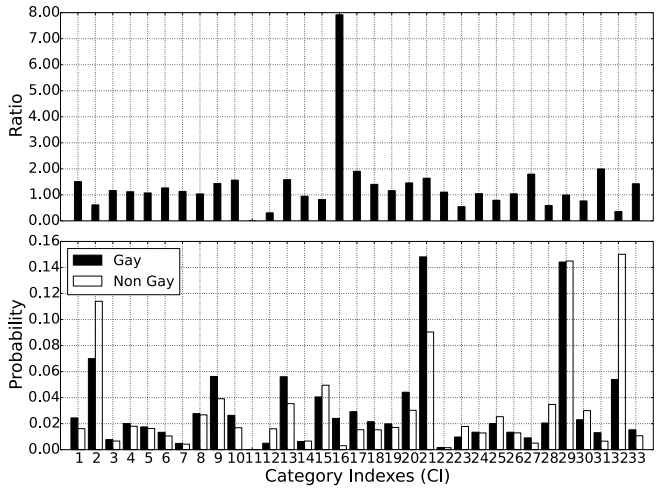


Fig. 6. The probability of a gay user that joins a group in a particular category is compared with that of a non-gay user. Note that the probability for "LGBT" category (CI 11) equals unity as we obtain gay users from LGBT groups by investigating their questionnaire; We omit the plot for "LGBT" category.

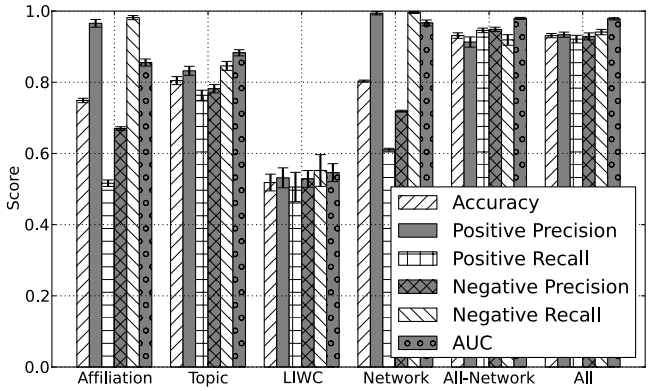


Fig. 7. The performance of binary classification is plotted for each of 5 classes in Table 4.

corpus, we use Linguistic Inquiry and Word Count (LIWC) [49], which is a transparent text analysis software that counts words into psychologically meaningful categories [16, 50, 53]. Table 3 describes the 32 categories in LIWC (which are related to psychological emotions) and their corresponding examples. Using the LIWC score that is calculated as the fraction of the words in each category, we construct a linguistic vector, which consists of LIWC scores, for each user.

Network Topology: It has been reported that a Facebook user's LGBT status can be inferred using her Facebook friend information [24], which is inspired from the saying, 'Birds of a feather flock together'. That is, a ratio of gay users among a user's neighbors in his Facebook network can be used to infer his LGBT status. Our work goes one step further: we consider (i) multiple OSNs and (ii) one or two hop neighbors in such OSNs. It is also worth noting that we do not use any personal information obtained from other social networks,

Table 4. Prediction Features for identifying a gay user

Features	Feature Description
Group Affiliation Class	
237,361 groups not related with LGBT	a group vector that a user has joined
Topical Interest Class	
98,990 interests not related with LGBT	an interest vector that a user has
LIWC Class	
32 psychological features	an LIWC score vector of comments
Network Class	
Comment1hopNumGay Facebook1hopNumGay Twitter1hopNumGay Comment1hopPortionGay Facebook1hopPortionGay Twitter1hopPortionGay	# and % of gay users in his first hop neighbors in Comment, Facebook, and Twitter Networks
Comment1+2hopNumGay Facebook1+2hopNumGay Twitter1+2hopNumGay Comment1+2hopPortionGay Facebook1+2hopPortionGay Twitter1+2hopPortionGay	# and % of gay users in his first and second hop neighbors in Comment, Facebook, and Twitter Networks
Comment2hopNumGay Facebook2hopNumGay Twitter2hopNumGay Comment2hopPortionGay Facebook2hopPortionGay Twitter2hopPortionGay	# and % of gay users in his second hop neighbors in Comment, Facebook, and Twitter Networks

but only extract network information such as friendship in Facebook or follower/following relationship. In this way, we assume that we know whether a user's friends are gay or not in advance. We model a social network S as a graph $S = (V, E)$, where V is the set of users, and E is the set of undirected edges between two users who have a social relationship. We construct three types of social networks in `Meetup`: (i) a commenting network where an edge is defined between two users if at least one of them sends message(s) to the other in their guestbooks, (ii) a Facebook friend network where an edge is defined between two users who share a friend relationship in Facebook, and (iii) a Twitter following/follower network where an edge is defined between two users who have a follower/following relationship.

All but Network Topology: To investigate whether a user's LGBT status can be inferred *without* any pre-knowledge such as her friends' LGBT status, we consider all the features described above without the Network Topology features. To normalize the LIWC scores in the Linguistic Characteristics for balancing with other features, we use $L2$ normalization, also known as the Euclidean norm, which divides each element by its size of the vector.

All: Finally, we collectively consider all the above features with the same normalization method as 'All but Network Topology'.

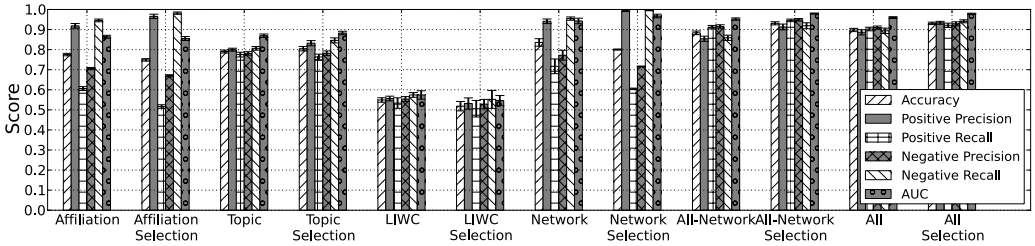


Fig. 8. The performance of binary classification that uses only 10% of selected features is comparable or even better than the one with all the features across the 5 classes.

5.2.3 Classification Results. We build a simple classification model using the Logistic Regression as a classifier, based on the above features. To resolve the class imbalance problem [21], where the performance of a learning model becomes severely low in the presence of underrepresented data (i.e., the class distribution is severe skewed.), we compose gay and non-gay users as 1:1 ratio to provide a balanced distribution [2]. To avoid over-fitting of our classification, we perform a 10-fold cross-validation, which splits the dataset into 10 different sets and use (i) a single set as a validation set and (ii) the other 9 sets as a training set, for 10 times with randomly selected non-gays (1,065 out of non LGBT users). With calculated tp (true positive), fp (false positive), tn (true negative), and fn (false negative), we report the following performance metrics:

- Accuracy: $\frac{tp+tn}{tp+tn+fp+fn}$.
- Positive Precision: $\frac{tp}{tp+fp}$.
- Positive Recall: $\frac{tp}{tp+fn}$.
- Negative Precision: $\frac{tn}{tn+fn}$.
- Negative Recall: $\frac{tn}{tn+fp}$.
- AUC: Area Under the receiver operating characteristic (ROC) Curve, which is a common evaluation score for binary classification problems. If a classifier is not better than random guessing, the score would be around 0.5. If it perfectly determines who is a gay user or not, the score becomes 1.

Figure 7 shows the average of accuracy, positive precision (predictability of a gay identification), positive recall, negative precision, negative recall, and AUC with error bars. As shown in Figure 7, we find out that the model based on 'Network Topology' performs better than the other models based on the other features. The accuracy and AUC of the model are 0.835 and 0.897, respectively, which indicates that the LGBT status of a user can be accurately inferred with the 'Network Topology' features. The model based on 'Affiliation' features also shows a good performance (i.e., the accuracy and AUC are 0.775 and 0.864, respectively), even though all the LGBT related groups are excluded, implying that the LGBT status of a user can be identified by analyzing her non-LGBT group affiliations. On the other hand, the model based on the 'Linguistic Characteristics' features shows the lowest performance, which means the linguistic patterns between gay and non-gay people may not be substantially distinguishable. The model based on 'All but Network Topology' features shows the higher performance than the one based on 'Network Topology' (0.883 vs. 0.835),

which indicates that a user's LGBT status can be inferred *without* any pre-knowledge of LGBT status of her friends.

5.2.4 Feature Selection. Since the irrelevant features may degrade the performance of classification in terms of both speed (due to high dimensionality) and predictive accuracy (due to irrelevancy) [25], we explore the most relevant and effective features to identify a user's LGBT status. We first extract the top 10 features that most contribute in classifying a user's LGBT status in each model based on the 'Topic' and the 'Network Topology' features in Table 5. Because every group title in the 'Group Affiliation' features is searchable in `Meetup`, we do not include the top 10 features in the model based on the 'Group Affiliation' features in Table 5; a substantial portion of the top 10 group titles include sexually biased terms such as 'Male Massages' or 'Men Only'.

Table 5 shows that the sexual related topics such as 'Mens social' and 'Male Massage' are the most popular ones which gay users are interested in. Among the 'Network Topology' features, we find that the ratio of gay friends within two hop neighbors in a commenting network shows the best predictive power to identify a gay user, which implies that using not only the friends of a user but also the friends of his friends can increase the prediction performance, which was not explored in [24] that considers only the one hop neighbors. We further observe that the commenting network features show the most predictive power followed by the Facebook and the Twitter networks. This may be due to the fact that the commenting behavior usually occurs among people who actually have met at events in `Meetup`, which means they are likely to belong to the same group or share the same interests.

We next choose only the top 10% features that most contribute in classification, which are applied to our models. Figure 8 shows the performance results of the 10%-only models; for the comparison purposes, we also plot the performance results based on all the features. We find out that the performance of the 10%-only model based on the 'Network Topology' increases, which is due to the positive effect of eliminating some noises from the features related to the Twitter network. Interestingly, the 10%-only model based on the 'All but Network' features outperforms the 10%-only model based on 'All' features (AUC: 0.979 vs. 0.978), which implies that a user's LGBT status can be effectively identified *without* any pre-knowledge and using a simple machine learning technique.

Table 5. Top 10 positive features of topic and network topology classes are listed. In Network column, feature lists are the same ones in Table 4.

Rank	Topics	Network
1	Mens Social	Comment1+2hopPortionGay
2	Male Massage	Facebook1+2hopNumGay
3	New York	Comment1hopNumGay
4	Performing Arts	Comment1hopPortionGay
5	Baltimore	Comment1+2hopNumGay
6	Friends	Facebook1hopNumGay
7	Professional	Facebook1hopPortionGay
8	New York City	Facebook1+2hopPortionGay
9	Social	Twitter1+2hopNumGay
10	Public Speaking	Twitter1+2hopPortionGay

5.2.5 Discussion: Cross-Service (Personal) Information Compilation and Privacy Leakage Spread. As many online social services including `Meetup` allow a user to login and connect with other major service accounts (such as Facebook, LinkedIn, and Twitter), information from multiple services could be gathered and compiled together, and during such process unintended personal information disclosure might happen. At the same time, the consequence of privacy breach in one service would easily be transferred to other services. Our result showed that sensitive information such as LGBT status can be easily disclosed by inspecting a user's group affiliations and other participation behaviors. We believe that the case might be similar with other sensitive information such as religion, medical status, and political attitudes. If such sensitive interests can be inferred from one service and are transferred to other services or even to offline life (e.g., to friends or at workplace), the consequences would be very severe. For example, by mining our dataset, we are able to identify 625,017 users who have Facebook accounts, to whom we can easily send messages to their friends by easily obtaining friend lists from Facebook. Also, 170,280 users who have their Linkedin accounts are identified, which specified their career or current work information. Further research is required to analyze threats of such cross-service privacy breach and also to devise plans for privacy protection in such context.

6 CONCLUDING DISCUSSION

Before concluding our paper, our results indicate there are a number of topics to be discussed.

6.1 EBSNs vs. OSNs

While traditional OSNs mostly focuses on social interaction solely in online spaces, EBSNs support offline activities as well as online interactions. In this way, more diverse information comparing to traditional OSNs are obtainable in EBSNs. For example, a user's interests, location/time of the events a user has attended, attendance history of the events, group affiliation information of a user are publicly accessible.

One of the main problems that can cause the privacy issue in EBSNs is that a member in a group does not have a control to reveal or hide such user information. In particular, for example, an organizer of a group can disclose its member list public even though its members do not want to reveal their association information to others. This also happens in Facebook groups; a member does not have a control to hide his association information in the group since only its organizer can decide to open/hide its member list. We believe this is due to a conflict of interest between organizers and group members as discussed more in detail in the following subsection; an organizer may want to advertise the group, hence disclose member information (e.g., gender); group members do not want to reveal their association information. Therefore, understanding such a conflict is essential in designing secure EBSNs.

6.2 Other Potential Privacy Leakages in Meetup

There are many privacy leakage issues including sexual orientation, political opinion, or location information in EBSNs. Location privacy leakage is a good example; there have been attempts to estimate a user's home location based on a variety of information (e.g., geotagged information or social interactions) available in OSNs. Pontes *et al.* tried to infer the home city (or state, country) of a user from publicly available information in Foursquare [39], which is one of the most popular location-based social networks. However, the granularity of location leakage was not precise; they showed that the location a user has visited most frequently is highly likely to be in his/her home city or state. However, in Meetup, a public

event page usually contains 1) the *exact* location of event and time information and 2) list of members who join the event, so it can be easily predicted who will attend to which event, which may raise a severe location privacy issue. Moreover, we found that some users hold events in their houses, e.g., a home party, In our dataset, we observed that the name of place for such an event posted on the group page contains expressions like "one's house"; our dataset includes 86,535 events with exact address information.

6.3 Limitation, Generalization, and Extensibility

We obtained ground truth dataset (i.e., Gay users) by looking at their questionnaires which are publicly available. However, we were not able to differentiate them whether they are willing to publicize their sexual orientation or they do not want to disclose their sexual orientation. Hence, we just identify the sexual orientation of the given user no matter what he/she wants to reveal/hide his/her sexual orientation. If we can differentiate the both cases, e.g., via surveys, we can evaluate our classifiers more accurately. We leave it for future work. We believe our approach can be easily extended to other online social networks where users' topical interest can be inferred, e.g., through group affiliation information, network topology, keywords. For example, Douban also allows users to create an offline event based on shared interests such as film, books, and music. Also, they can specify their interests and track other's attendance history of the events. As our privacy implication model based on the group affiliation (which represents homophily) and similarity of users' interest we believe our model can be applied easily. Interestingly, Douban allows users to follow others *freely*, without any permissions based on the interests. Considering the recent research work [4], which showed that Pinterest users tend to follow others who have similar interests, a follower/following (homophily) network could provide more powerful classifier to predict user's traits based on its social network. Also, Kosinski et al. [26] used machine learning technique to identify a user's marriage status or other personal traits using their 'Like history' on Facebook obtained from volunteers who provided all Like history.

6.4 Conclusions

We have conducted a comprehensive measurement study to explore privacy leakage problems that exist on `Meetup`, a representative EBSN. Using the collected dataset, we showed a user's LGBT status, which is highly private, can be leaked in `Meetup`, as not only online interactions but also offline real world trails can be captured in `Meetup`. Our classification models can accurately identify a user's LGBT status with over 90% accuracy regardless of the user's intention. We believe this work can give an important implication on designing and building more secure EBSN platforms that entwine online interactions and offline activities.

7 ACKNOWLEDGEMENTS

This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2016R1D1A1A09919378), and National Research Foundation of Korea through PF Class Heterogeneous High Performance Computer Development (NRF-2016M3C4A7952587).

REFERENCES

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the World Wide Web (WWW'10)*.

[2] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor. Newsl.* 6, 1 (2004).
[3] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. 2010. Investigating Homophily in Online Social Networks. In *Proceedings of Web Intelligence and Intelligent Agent Technology (WIIAT)*.
[4] Shuo Chang, Vikas Kumar, Eric Gilbert2, and Loren Terveen. 2014. Specialization, homophily, and gender in a social curation site: findings from pinterest. In *Proceedings of the Computer Supported Cooperative Work and Social Computing*. ACM, 674–686.
[5] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'10)*.
[6] Taejoong Chung, Jinyoung Han, Daejin Choi, Taekyoung "Ted" Kown, Huy Kang Kim, and Yanghee Choi. 2014. Unveiling Group Characteristics in Online Social Games: A Socio-Economic Analysis. In *Proceedings of World Wide Web (WWW'14)*.
[7] CNN. 2011. From Howard Dean to the tea party: The power of Meetup.com. http://edition.cnn.com/2011/11/07/tech/web/meetup-2012-campaign-sifry. (2011).
[8] Diego Couto, Gabriel Magno, Evandro Cunha, Marcos André Gonçalves, César Cambraia, and Virgilio Almeida. 2014. Noticing the Other Gender on Google+. In *Proceedings of the ACM Web Science Conference (WebSci'14)*.
[9] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. 2010. Bridging the Gap Between Physical Location and Online Social Networks. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'10)*.
[10] Clodoveu Davis, Pappa L. Gisele, Diogo Renno Rocha de Oliveira, and Filipe de Lima Arcanjo. 2011. Inferring the Location of Twitter Messages Based on User Relationships. *T. GIS* 15, 6 (2011), 735–751.
[11] Ratan Dey, Cong Tang, Keith W. Ross, and Nitesh Saxena. 2012. Estimating age privacy leakage in online social networks. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'12)*.
[12] Rong Du, Zhiwen Yu, Tao Mei, Zhitao Wang, Zhu Wang, and Bin Guo. 2014. Predicting Activity Attendance in Event-based Social Networks: Content, Context and Social Influence. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'14)*.
[13] Lise Getoor Elena Zheleva. 2009. To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In *Proceedings of the World Wide Web (WWW'09)*.
[14] Andrew T. Fiore and Judith S. Donath. [n. d.]. In *Proceedings of Computer and Human Interaction (CHI)*.
[15] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting Personality from Twitter.. In *SocialCom/PASSAT*. IEEE, 149–156.
[16] Scott A. Golder and Michael W. Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333, 6051 (Sept. 2011), 1878–1881.
[17] G. Golub and W. Kahan. 1965. Calculating the singular values and pseudo-inverse of a matrix. *SIAM J. Numer. Anal.* 2, 2 (1965), 205–224.
[18] Australian Government. 2008. For Your Information: Australian Privacy Law and Practice (ALRC Report 108). http://www.alrc.gov.au/publications/report-108A. (2008).
[19] Shion Guha and Stephen B Wicker. 2015. Do Birds of a Feather Watch Each Other?: Homophily and Social Surveillance in Location Based Social Networks. In *Proceedings of the Computer Supported Cooperative Work and Social Computing*. ACM, 1010–1020.
[20] Junwei Han, Jianwei Niu, Alvin Chin, Wei Wang, Chao Tong, and Xia Wang. 2012. How Online Social Network Affects Offline Events: A Case Study on Douban. In *UIC/ATC*. 752–757.
[21] Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. *IEEE Trans. on Knowl. and Data Eng.* 21, 9 (Sept. 2009).
[22] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*.
[23] Hughes, David John, Rowe, Moss, Batey, Mark, Lee, and Andrew. 2012. A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior* 28, 2 (2012), 561–569.
[24] Carter Jernigan and Behram F. T. Mistree. 2009. Gaydar: Facebook Friendships Expose Sexual Orientation. *First Monday* 14, 10 (2009).

[25] Kenji Kira and Larry A. Rendell. 1992. A Practical Approach to Feature Selection. In *Proceedings of the Ninth International Workshop on Machine Learning (ML92)*. 249–256.
[26] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (April 2013), 5802–5805.
[27] Haewoon Kwak and *et al.* 2010. What is Twitter, a social network or a news media?. In *Proceedings of the World Wide Web (WWW'10)*.
[28] Ellen Lewin and William L. Leap. 2002. *Studying Lesbian and Gay Languages: Vocabulary, Text-making, and Beyond*.
[29] Sue Ellen Linville. 1998. Acoustic Correlates of Perceived versus Actual Sexual Orientation in Men's Speech. *Folia Phoniatr Logop* 50 (1998), 35–48.
[30] Xingjie Liu, Qi He, Yuanyuan Tian, Wang-Chien Lee, John McPherson, and Jiawei Han. 2012. Event-based Social Networks: Linking the Online and Offline Social Worlds. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD'12)*.
[31] Sai Lu, Janne Lindqvist, and Rebecca N. Wright. 2014. Uncovering Facebook Side Channels and User Attitudes. In *Proceedings of Web 2.0 Security and Privacy Workshop (W2SP)*. IEEE.
[32] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2014. Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology* 5, 3 (2014).
[33] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *Proceedings of the ACM workshop on Privacy in the electronic society (WEPS'11)*.
[34] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27 (2001), 415–444.
[35] Meetup.com. 2015. About Meetup. http://www.meetup.com/about/. (2015).
[36] Jimmy Lin Michael D. Lieberman. 2012. You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'10)*.
[37] Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. 2010. You are who you know: inferring user profiles in online social networks. In *Proceedings of Web search and data mining (WSDM)*.
[38] Benjamin Munson. 2007. The Acoustic Correlates of Perceived Masculinity, Perceived Femininity, and Perceived Sexual Orientation. *Language and Speech* 50 (2007), 125–142.
[39] Tatiana Pontes, Marisa Vasconcelos, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. 2012. We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'12)*.
[40] Adrian Popescu and Gregory Grefenstette. 2010. Mining user home location and gender from flickr tags. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'10)*.
[41] Huffington Post. 2014. Meetup.com: A Secret Weapon for Your Career and Personal Brand. http://www.huffingtonpost.com/stephan-spencer/using-meetupcom-as-a-bran_b_4767898.html. (2014).
[42] Zhi Qiao, Peng Zhang, Chuan Zhou, Yanan Cao, Li Guo, and Yanchuan Zhang. 2014. Event Recommendation in Event-Based Social Networks. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14)*.
[43] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *Proceedings of the Third International Conference on Social Computing (SocialCom) and the Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*. IEEE, 180–185.
[44] Norman M. Sadeh, Jason I. Hong, Lorrie Faith Cranor, Ian Fette, Patrick Gage Kelley, Madhu K. Prabaker, and Jinghai Rao. 2009. Understanding and capturing people's privacy policies in a mobile social networking application. *Personal and Ubiquitous Computing* 13, 6 (2009), 401–412.
[45] Thomas H. Sander. 2005. E-associations: using technology to connect citizens: the case of meetup.com. *American Political Science Association* (2005), 47.
[46] Lauren F. Sessions. 2010. How offline gatherings affect online communities – When virtual community members meetup. *Information, Communication & Society* 13, 3 (2010).
[47] Jan Skopek, Florian Schulz, and Hans-Peter Blossfeld. 2010. Who Contacts Whom? Educational Homophily in Online Mate Selection. *European Sociological Review* 27, 2 (2010), 180–195.
[48] B Tarbush and A Teytelboym. 2012. Homophily in Online Social Networks. *Proceedings of International Workshop on Internet and Network Economics* 7695 (2012), 512–518.

[49] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html. (2010).
[50] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM'10)*.
[51] Human Rights Library University of Minnesota. 2013. Study guide, Sexual Orientation and Human Rights. http://hrlibrary.umn.edu/edumat/studyguides/sexualorientation.html. (2013).
[52] Claudia Wagner, Sitaram Asur, and Joshua M. Hailpern. 2013. Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter.. In *International Conference on Social Computing (SocialCom'13)*. 303–310.
[53] Shaomei Wu, Chenhao Tan, Jon Kleinberg, and Michael Macy. 2011. Does Bad News Go Away Faster?. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM'11)*.
[54] Bin Xu, Alvin Chin, and Dan Cosley. 2013. On How Event Size and Interactivity Affect Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*.

Received April 2007; revised September 2017; accepted November 2017