Understanding Cross-site Linking in Online Social Networks

Yang Chen[1], Chenfan Zhuang[2], Qiang Cao[1], Pan Hui[3]
[1]Department of Computer Science, Duke University, Durham, NC 27707, USA
[2]School of Software, Tsinghua University, Beijing 100084, China
[3]CSE Department, Hong Kong University of Science and Technology, Hong Kong
{ychen,qiangcao}@cs.duke.edu, zhuangcf11@mails.tsinghua.edu.cn, panhui@cse.ust.hk

ABSTRACT

Online social networks (OSNs) have attracted billions of users, and play an important role in people's daily life. A user often has accounts on multiple OSN sites. In this paper, we study the emerging "cross-site linking" function, which is supported by many OSNs. Our study is based on Foursquare, a representative location-based social networking (LBSN) service. We conduct a data-driven analysis by using crawled public profiles of almost all (if not all) Foursquare users. Our analysis has shown that the cross-site linking function is widely adopted by Foursquare users, and the users who have enabled this function are more active than other users. We have also found that users who are more concerned with online privacy have a lower probability to enable the cross-site linking function. Besides analyzing crawled Foursquare user profiles, we further explore cross-site linking between Foursquare and other OSN sites, i.e., Facebook and Twitter. The study on "Foursquare-Facebook" linking indicates that users have a high probability to provide consistent information to different OSNs. Meanwhile, "Foursquare-Twitter" linking is used to demonstrate the usefulness of aggregating user-generated content across multiple OSN sites. We present a gender-based analysis of Twitter, which is made accurate by leveraging cross-site links between Foursquare and Twitter.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Online Social Networks, Cross-site Linking, Foursquare, Measurement, Facebook, Twitter

1. INTRODUCTION

Nowadays online social networks (OSNs) [6] have become extremely popular. There are a number of OSN sites with different functionality. A person might use Facebook to interact with friends [21], use Twitter to share breaking news [8], use LinkedIn for job search, and use Foursquare for location-centric social applications [14, 10, 24]. It is quite common for an individual user to have multiple accounts on different OSN sites.

Managing multiple accounts simultaneously could be troublesome for an OSN user. To improve the user experience, today's major OSN sites, such as Pinterest [1], Foursquare [2], SoundCloud [3], Quora [4], Google+ [5], and Taringa [6], have introduced a "cross-site linking" function. This function allows a user to link his account on one OSN site to his accounts on other OSN sites. For example, Foursquare, the representative LBSN service, allows a user to link his public profile to Facebook and Twitter. This function can provide many viable benefits for OSN users. We list three of them as below.

First, as a user might wish to publish the same content (e.g., a photo) on multiple OSNs, cross-site linking can make cross-site content posting easy. For example, if a user performs a "check-in" on Foursquare, the cross-site linking function can automatically publish this message on Facebook and Twitter.

Second, cross-site linking avoids repeated efforts in social connection establishment. If a user has accounts on multiple OSN sites, he might want to connect with the same people on each of these sites. Manually sending contact invitations to them on every site is tedious. However, cross-site linking can make it easier by importing added contacts from other sites. For example, a user only needs to add his friends on Facebook once, and his Foursquare account can simply import his contact list from Facebook. If he wants to create social connections with the same people on Foursquare, the contact invitations can be sent together by just one click, instead of manually inviting each of them.

Third, cross-site linking provides more information of a user, beyond that stored on a single OSN site. For example, if Alice receives a contact invitation from Bob via Foursquare, she can read through Bob's Foursquare page to see whether to accept the invitation. If Bob has linked his Foursquare profile to both Facebook and Twitter, Alice can

SNAKDD'14, August 24, 2014, New York City, NY, USA

http://dx.doi.org/10.1145/2659480.2659498.

[1]http://www.pinterest.com/
[2]http://foursquare.com/
[3]http://soundcloud.com/
[4]http://www.quora.com/
[5]http://plus.google.com/
[6]http://www.taringa.net/

not only check Bob's information on Foursquare, but also view Bob's activities on Facebook and Twitter. This can help Alice know more about Bob.

Given these advantages, cross-site linking has become an important function in many popular OSN sites. Due to the lack of systematic investigation of this widely-used function, we introduce a measurement-based study to shed light on it. We adopt Foursquare as our main focus, and have crawled profiles of almost all (if not all) its users. To the best of our knowledge, our work is the first study for the emerging cross-site linking function using the entire user population of a mainstream OSN site.

In § 3, we introduce the concept of "linking option", and we investigate the linking option distribution among all users. We have found that about 60% Foursquare users have enabled the cross-site linking function. We also study the linking option distribution among different users groups, and we demonstrate the behavioral difference among users with different linking options.

Intuitively, cross-site linking might cause concerns for users who care a lot about their privacy. In § 4, we examine the optional fields in Foursquare profiles, and figure out the relationship between cross-site linking and user privacy. Our hypothesis is that users who complete optional fields are less concerned with online privacy. Our results show that adding contents to an optional field in the user profile always indicates a higher chance of enabling the cross-site linking function.

Besides studying cross-site linking using collected Foursquare profiles, we also explore the cross-site linking between different OSNs in § 5, i.e., "Foursquare-Facebook" linking and "Foursquare-Twitter" linking. We investigate two viable problems. For instance, cross-site linking allows us to examine the information consistency among an user's accounts on different OSN sites. Our study on *cross-site information consistency* has shown that a user has a high probability to provide consistent information across linked accounts on Foursquare and Facebook. In addition, given the diverse focuses of different OSN sites, profiles on different sites contain different information fields. For an individual user, putting the information of his profiles on different OSN sites together will provide more comprehensive knowledge of this user. By using *cross-site information aggregation*, many previously impossible tasks will become possible. For example, it was difficult to perform gender-based analysis for Twitter users in an accurate way. Now we can utilize Foursquare's user gender information to bridge this gap.

2. THE CROSS-SITE LINKING FUNCTION ON FOURSQUARE

Among the OSN sites that support the cross-site linking function, we choose Foursquare which provides us the following advantages. First, Foursquare is a representative LBSN service, and it is one of the most popular OSN sites. Second, Foursquare supports the cross-site linking function, and it allows its users to link their accounts to both Facebook and Twitter. Third, every Foursquare user has a publicly visible profile page, and we can access the information of a user's linked Facebook/Twitter account through this page.

In this section, we first present a quick overview of the Foursquare social network, and how a Foursquare user profile is linked to other OSN sites. In addition, we describe how we crawl the profiles of all Foursquare users in a collaborative way.

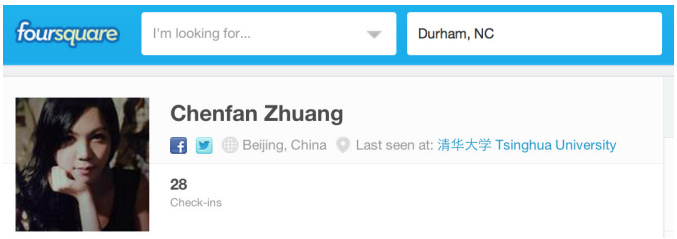


Figure 1: A Foursquare User's Public Profile Page

2.1 Overview of Foursquare and Its Cross-site Linking Function

Due to the rapid development of mobile devices, LBSN services have become one of the most popular mobile applications. In Foursquare, a representative LBSN, a user can use a "check-in" function to claim that he has been to a selected nearby place. The wide use of GPS-enabled mobile devices makes such "check-in" convenient. In such a social network service, a user's friend can get updates of his latest geo-location. The recently released "swarm" application allows Foursquare users to easily find the nearby friends. Moreover, a Foursquare user can leave a "tip" for any venue, which is a publicly viewable comment on the venue according to the user's experience. There has been existing measurement work on Foursquare [14, 10, 24].

On Foursqure, a user can link his public profile to his accounts on external OSN sites through two steps. First, he needs to submit his Facebook/Twitter user ID to Foursquare. Second, he needs to prove he owns the submitted Facebook/Twitter account. Therefore, he needs to authorize Foursquare to access his Facebook/Twitter account. During the authorization, his Facebook/Twitter username and password are needed. Once Foursquare has verified the ownership, the Facebook/Twitter IDs could be formally added to his public profile.

Note that each Foursquare user is allowed to add only one Facebook account and one Twitter account. Meanwhile, each Facebook/Twitter account can be linked by only one Foursquare user at the same time. For a certain Facebook or Twitter account which has been linked by a Foursquare user, if yet another Foursquare user is authorized to link to it, the previous cross-site link will be removed.

Fig. 1 shows a Foursquare user's public profile page. This user has linked her profile to her Facebook and Twitter accounts. By clicking the Facebook icon, we can access her Facebook profile page. Similarly, by clicking the Twitter icon, we can access her Twitter profile page.

2.2 Data Collection

Given the large user population, previous work only uses a small subset of users to conduct data-driven studies related to Foursquare [14, 10, 24]. In our work, we aim to analyze the entire Foursquare user base, and this can avoid the disadvantages of biased sampling. However, Foursquare is growing rapidly, and has reached 50 million registered users in May 2014. Quickly crawling massive data is not trivial, as a single IP address's data fetching speed is strictly

Table 1: Cross-site Linking Options

Twitter	Facebook	Linking Option	Percentage
Y	N	TW only	3.82%
N	Y	FB only	44.19%
Y	Y	FB+TW	11.96%
N	N	Neither	40.03%

bounded by a modest threshold. As in [3], we split the overall crawling task into some independent small tasks, and use a number of servers to crawl the profiles of all Foursquare users in a distributed and timely fashion.

Every Foursquare user is identified by a unique numerical user ID. This ID is assigned in an ascending order. In other words, if user A registers earlier than user B, his numerical user ID would be smaller than user B's user ID. If we know the ID of a user, we can access the URL `http://foursquare.com/user/ID` to obtain his public profile. We registered a new Foursquare account on July 22th, 2014, and its ID is 90990730. We adopt it as the maximum user ID before we start the data crawling. Note that not all IDs between [1, 90990730] are assigned to users. Some IDs are reserved for venues (point of interests), some are brands (business accounts), and some are unused.

We implement a Python-based distributed crawler, and every user's profile would be downloaded and saved as an HTML page. As Foursquare does not allow too many concurrent requests from the same IP address within a short time period, we have to allocate many servers with different public IP addresses to form a collaborative crawling cluster. We use 100 PlanetLab nodes around the United States. We evenly divide the whole ID space into 100 chunks, and each node is responsible for crawling one chunk of IDs. The crawling lasted one week, from July 22th to July 29th, and we have successfully crawled the public profiles of 51.15 million Foursquare users. These users have conducted 6.11 billion check-ins. Since Foursquare was launched in March 2009, and had reached 50 million users in May 2014 [7], we believe that we have crawled almost all (if not all) Foursquare users' profiles. To the best of our knowledge, this is the largest data set of the Foursquare user profiles used for research.

By parsing all retrieved HTML pages of user profiles, we can obtain every Foursquare user's information (e.g., gender, number of friends, number of checkins, number of tips, linked Facebook account, and linked Twitter account).

3. CROSS-SITE LINKING OPTIONS

In this section, we analyze the cross-site linking function on Foursquare. We examine the entire Foursquare user base using 51.15 million crawled user profiles. In § 3.1, we study the percentage of users that enables explicit linking to different external OSNs, such as Facebook and Twitter. In § 3.2, through a group-based analysis we investigate the preferences of user groups on cross-site linking. In § 3.3, we further examine the behavioral difference among users that enable cross-site linking to different OSNs.

3.1 Linking Option Distribution of the Entire Foursquare Population

On Foursquare, users can choose to expose links that connect to their accounts on other OSNs. According to the exposed cross-site links we assign a unique "linking option" to every Foursquare user, as shown in Table 1. There are four possible linking options, i.e., "TW only", "FB only", "FB+TW", and "Neither". "TW only" and "FB only" represents the users with links only pointing to their Twitter accounts and Facebook accounts, respectively. "FB+TW" includes the users who have linked both Facebook and Twitter accounts. The rest of users are assigned to the "Neither" group, as they do not link any external account.

By examining all the crawled Foursquare user profiles, we compute the account percentage of each linking option in Table 1. We can see that about 60% of Foursquare users have linked at least one account on Facebook or Twitter. Specifically, 56.15% users have added their Facebook accounts, and 15.78% users have added their Twitter accounts. These numbers indicate that the cross-site linking function is widely used among Foursquare users.

3.2 Group-based Analysis of the Linking Option Distribution

Besides studying all Foursquare users as a whole, we consider different features to classify users into disjoint groups. This enables us to examine user preference on cross-site linking. In particular, the distribution of cross-site linking options varies among different groups of users. In this subsection, we study user groups that are derived by dividing the entire user base according to a user's gender, country, and certain activity, respectively.

3.2.1 Gender

It has been reported that user gender has significant influence on user behavior in online social networks. Previous work studied the gender influence on Facebook [20, 13], Flickr [11], Twitter [19], and Myspace [15]. From this perspective, we aim to study the existence of the difference of the cross-site linking behavior between male and female users.

Among the collected Foursquare user profiles, 51.52% are male and 42.92% are female. The rest 5.56% of the users do not publish their gender information. As can be seen in Fig. 2(a), there is very little difference between male users and female users in terms of the distribution of linking options.

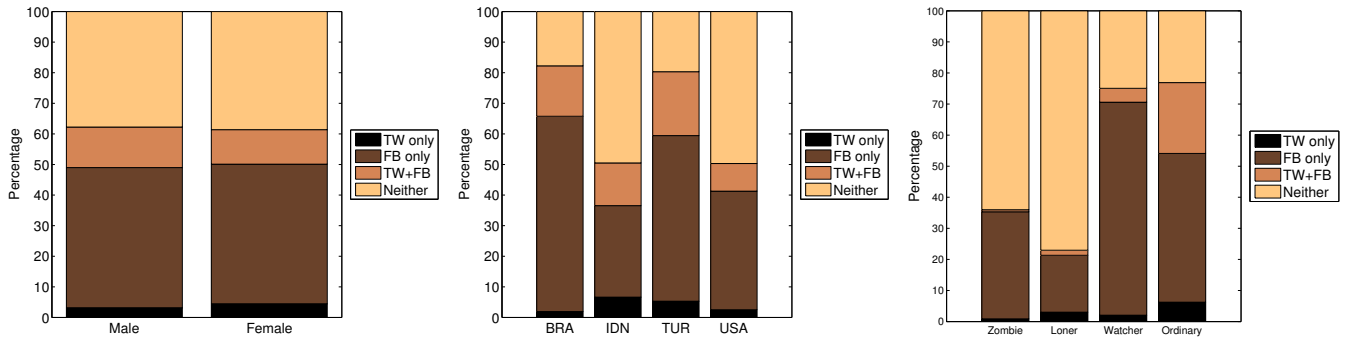
3.2.2 Country

Country-based analysis is also widely used to understand social networks. Existing literatures have reported country-based analysis on Twitter [7] and Facebook [16]. We infer a Foursquare user's country according to the "residential location" field in his crawled profile. By using the Google Geocoding API[8], we are able to obtain the country information of 89.44% of Foursquare users. Since 8.32% of Foursquare users choose to hide their residential location, only the rest 2.24% of Foursquare users' country information cannot be determined.

According to our analysis, the top four countries that have the largest Foursquare user populations are the United States (USA), Turkey (TUR), Indonesia (IDN), and Brazil (BRA). They cover 27.61%, 9.49%, 8.17% and 7.04% of

[7] `http://www.foxbusiness.com/technology/2014/05/15/foursquare-unveils-new-swarm-app/`

[8] `https://developers.google.com/maps/documentation/geocoding/`



(a) Gender (b) Country (c) Activity

Figure 2: Group-based Analysis (Gender/Country/Activity)

Foursquare users, respectively. As shown in Fig. 2(b), users from different countries have quite different distributions of the cross-site linking options. Among these countries, Brazil has the highest percentage (82.25%) of users that link their Foursquare profiles to other OSNs. In contrast, the United States has the lowest percentage (50.33%) of users to use this function.

3.2.3 Activity

We also group users by their activities. As a LBSN service, social interactions and location-centric activities (e.g., check-ins and leaving tips) play key roles in user behavior on Foursquare. Based on these two factors, we divide all users into the following four groups.

- *Zombies*: Users who have none friend, and have never performed any location-centric activity. This group of users have no interaction with other Foursquare users. Many of them are likely to be crawlers, or newly registered users.
- *Loners*: Users who have none friend, but have performed location-centric activity. Such users do not connect with other Foursquare users, but still perform check-ins and leave tips.
- *Watchers*: Users who have at least one friend and zero location-centric activity. These users are silent, and they use only the OSN features of Foursquare.
- *Ordinary users*. For the rest of users, we put them into the fourth group. They have at least one friend, and have either perfomed check-ins or left tips.

The percentage of these four groups of users are 28.23%, 9.50%, 14.02%, 48.25%, respectively. According to Fig. 2(c), 64.02% of zombies and 77.05% of loners have not enabled cross-site linking. We believe it is because those users are socially isolated.

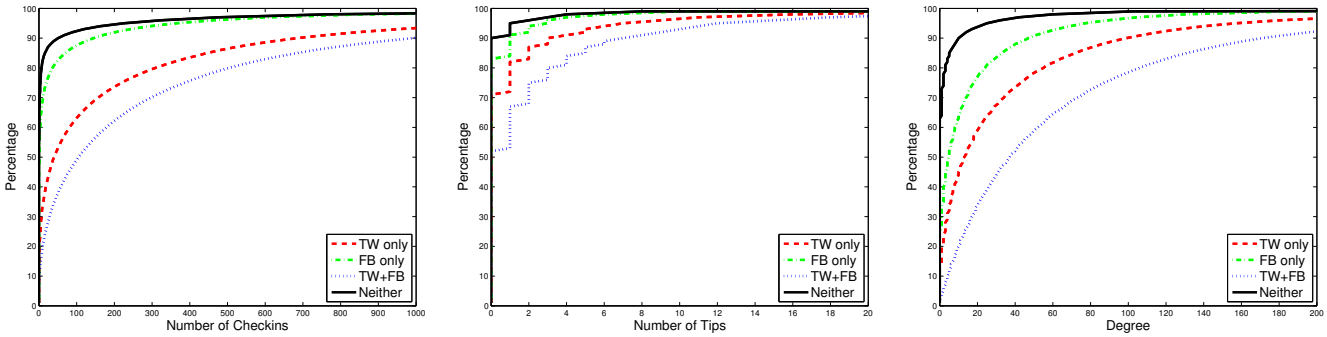
We compared watchers with ordinary users. On one hand, 6.51% of watchers and 29.02% of ordinary users have linked their accounts to Twitter. Watchers are silent and do not perform any location-centric activities. It is likely that they are less motivated to join Twitter, which is a news spreading platform [8]. On the other hand, 73.01% of watchers and 70.69% of ordinary users have linked their accounts to Facebook. As users from both of these two groups are connected with other Foursquare users, they have similar percentages of accounts that are linked to Facebook.

3.3 Behavioral Difference among Users with Different Linking Options

In this subsection, we discuss the behavioral difference among users with different linking options. We examine a user's behavior from two important aspects, i.e., content generation behavior and social connectivity. For content generation behavior, we examine two key metrics on Foursquare, i.e., number of check-ins and number of tips. For social connectivity, we study a user's number of friends. Fig. 3 shows the cumulative distribution function (CDF) of the number of check-ins/tips/friends of users with each linking option.

According to Fig. 3(a) and Fig. 3(b), we witness that the users who have enabled "cross-site linking" function are more "active", i.e., they submit check-ins more frequently and leave more tips. Particularly, the users whose linking option are "FB+TW" are most active in terms of content generation. Therefore, although all Foursquare users have identical functionality in performing check-ins and leaving tips, the cross-site linking function will deliver the newly published contents to more prospective audience. As a result, these users have more motivation to publish. We also see that the "TW only" users are in general more active than the "FB only" users. We believe that it is because "TW only" users have more incentive to publish. As a news spreading platform [8], Twitter can quickly spread Foursquare users' check-ins or tips as publicly viewable tweets through the microblogging network. In contrast, by default a Facebook user's status is only visible to friends, which limits the number of possible audiences.

From Fig. 3(c), we can see that the cross-site linking function can help users connect to more friends on Foursquare, as it can help users import existing friends from other OSNs. Particularly, if a user who has linked both the Facebook and Twitter accounts to his profile, he will have a higher probability to acquire more friends. For the users who have linked to only one external account, the average number of friends of "TW only" users is larger than that of "FB only" users. The difference between "TW only" users and "FB only" users is similar to the results shown in Fig. 3(a) and Fig. 3(b). Since "TW only" users are more active in content generation, they would also have higher chance to get more friends.



(a) Number of Check-ins (b) Number of Tips (c) Number of Friends

Figure 3: Linking Options v.s. Number of Check-ins/Tips/Friends

4. IMPACT OF USER PRIVACY CONCERNS

Although there are about 60% users on Foursquare have enabled the cross-site linking function, the rest of the users decided not to link their public profiles to Facebook/Twitter. As aforementioned, if a Foursquare user has enabled this function, other people can know more about him by accessing his Facebook and/or Twitter accounts. Intuitively, if a user cares about their privacy and do not want to expose additional personal information to the public, we believe that it could be a concern for him to use this function. In this section, we investigate the impact of user privacy concerns on the use of the cross-site linking function.

Foursquare allows users to customize their profiles according to their privacy concerns. It provides five optional fields for a user to set up his public profile, i.e., profile picture, gender, residential location, last name, and bio. A user can choose whether or not to fill and disclose each field to the public. We consider users who do not disclose profile fields as privacy concerning. We summarize our findings as below.

Profile Picture: By default, the Foursquare platform assigns each new user a blank profile photo. A user can choose to upload his/her own profile picture. Among all users, 66.99% have uploaded their own profile photos, while 33.01% choose not to upload their own photos. In Fig. 4(a), among the users who have uploaded profile photos, 77.23% of them have enabled the cross-site linking function. In contrast, the percentage is only 24.94% for the users who have not uploaded. Therefore, whether or not uploading personalized profile photo is an indicator for the adoption of cross-site linking.

Gender: Gender is another optional field in a Foursquare user's profile. Among all users, 5.56% of them have chosen to hide their gender information. According to Fig. 4(b), for users who have disclosed their gender information, about 61.85% of them have linked their accounts to Facebook or Twitter. In contrast, for users who have hided their gender information, only 27.99% of them have linked their accounts to Facebook or Twitter. Similarly, the availability of the gender information is another indicator for the use of cross-site linking.

Residential Location: A user's residential location is also optional in his/her profile. Among all users, 8.32% of them have chosen not to disclose their residential location. According to Fig. 4(c), for users choosing to make their location information public, 61.64% of them have linked their accounts. Differently, only 41.60% of the users that choose to hide their residential locations have linked their accounts.

Last Name: A user is required to provide his first name to Foursquare. However, the last name information is optional. About 94.61% of Foursquare users choose to add their last name to Foursquare profiles, and 5.39% of Foursquare users choose to hide this information. According to Fig. 4(d), for users who have provided their last names, 60.63% of them have enabled the cross-site linking function. For users who choose to hide their last names, 48.34% have enabled the cross-site linking function.

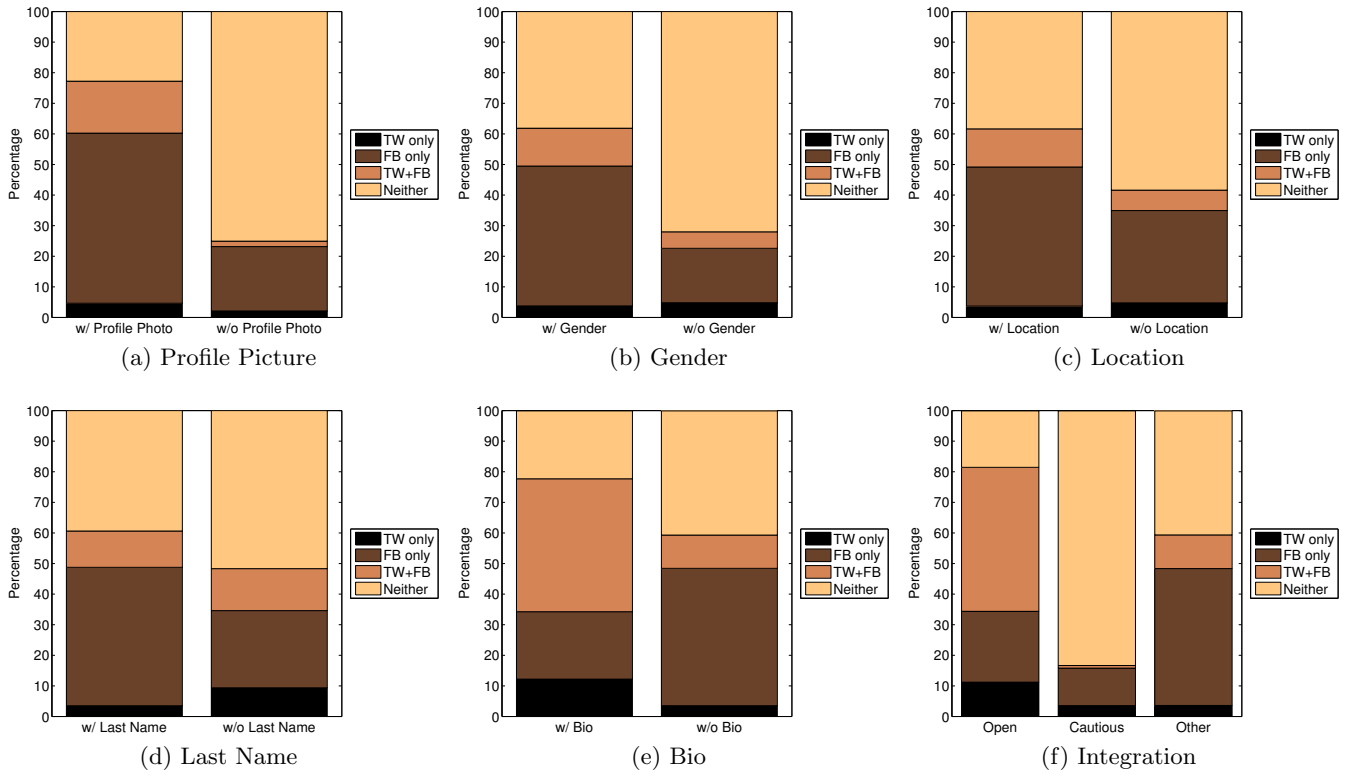
Bio: A Foursquare user is allowed to add a description about himself in the optional "Bio" field. About only 3.29% users have entered content in this field, and the rest 96.71% users choose to leave this field empty. As we have shown in Fig. 4(e), for users who have provided their bio, as many as 77.79% of them have enabled the cross-site linking. For users who have not entered their bio, 59.36% of them have enabled the cross-site linking.

The above analysis shows that enabling any of these five optional profile fields indicates a higher probability of using the cross-site linking function. In other words, for users who care more about their privacy, they have a smaller probability to enable the cross-site linking function.

Besides studying the five optional fields individually, we further consider them as an integrated whole, and divide users into groups according to the combination of their privacy settings in each field. For users who have filled all five aforementioned optional fields, we call them "open users", as they keep their public profiles as complete as possible. In contrast, for users who have not provided any information to all these five optional fields, we call them "cautious users", as they do not reveal any non-mandatory information. The rest of users are simply denoted as "other users". The percentages of open users, cautious users, and other users are 2.66%, 0.08%, and 97.26%, respectively. According to Fig. 4(f), we can see that 81.52% of the open users have enabled the cross-site linking function, while only 16.69% of the cautious users did so.

5. EXPLORING CROSS-SITE LINKING BETWEEN DIFFERENT OSNS

Our previous analysis is solely based on the crawled Foursquare user profiles. In this section, we explore cross-site linking



(a) Profile Picture (b) Gender (c) Location

(d) Last Name (e) Bio (f) Integration

Figure 4: Cross-site linking v.s. User Privacy

Table 2: Cross-site Information Aggregation: an Example

User	Foursquare					Twitter				
Info	ID	Gender	Checkins	Tips	...	ID	Tweets	Favourites	Language	...
User A	1	m	36	10	...	13213	1545	21	en	...
User B	2	f	520	51	...	9682	100	13	es	...
User C	5	f	1	0	...	1293213	5	0	pt	...
User D	9	m	1209	1	...	8876	23	17	en	...
...	...	...	...	...	...	...	...	...	...	...

between different OSNs. We study Foursquare users who have connected their Foursquare profiles to their Facebook and Twitter accounts. In particular, we crawled the account profiles and user activities on Facebook and Twitter through the cross-site links. This provides us additional user information from external OSNs. We first examine whether a Foursquare user has entered the same content for a specified profile field on different OSN sites, and accordingly we evaluate the cross-site information consistency in § 5.1. Second, we aggregate information from multiple OSN sites for a Foursquare user, and we study the cross-site information aggregation in § 5.2.

5.1 Cross-site Information Consistency

Different OSN sites share several common personal information fields for user profiles, such as gender, first name, last name, etc. If a user owns accounts on different OSN sites, he might choose to expose the same or different personal information on different sites. The cross-site linking function enables us to accurately evaluate the cross-site information consistency [2].

We first examine Foursquare and Facebook. User profiles on these two sites share three information fields in common, i.e., gender, first name, and last name. By examining these three fields, we evaluate the cross-site information consistency between Foursquare and Facebook.

Among the crawled Foursquare user profiles, we randomly sample 100,000 Foursquare users who have linked their profiles to their Facebook accounts. Based on a Facebook ID, we can crawl the corresponding user's basic information using the Facebook Graph API, i.e., accessing the URL: `http://graph.facebook.com/User_ID`. For every user, we can extract his/her gender, first name, and last name from the crawled page.

Determining whether the user has entered the same gender on both Foursquare and Facebook is simple, as the user can only choose from "male" and "female" if he/she has chosen to specify the gender. Differently, checking whether two names are the same is more complicated for non-English speaking users, as some languages have different spellings for the same name, e.g., in German the character O-umlaut ("ö") can also be written as "oe". For simplicity, we focus on

Table 3: Gender-Analysis for Twitter: Usage of Selected Optional Fields (%)

	URL	Description	Location
Male	32.07	62.82	56.87
Female	25.02	64.73	57.35

the users whose default language on Facebook is "EN-US" or "EN-UK".

We define the *consistent percentage* as the percentage of users who have entered exactly identical information in a selected field on both Foursquare and Facebook. According to our crawled data, the consistent percentage of the first name field and the last name field are 89.84% and 87.02%, respectively, while that of the gender field is 99.30%. We believe such difference is due to the limited number of choices for the user gender, i.e., one can only pick from "male" and "female". Nevertheless, even for the two name-related information fields, we can still find that the users have a high probability to manifest cross-site information consistency.

5.2 Cross-site Information Aggregation

In the last subsection, we evaluate the cross-site information consistency between Foursquare and Facebook. For Twitter, these investigated information fields are not available in users' profiles [9]. However, we can look at cross-site linking from another angle, i.e., cross-site linking makes it possible to aggregate information from multiple OSN sites.

By aggregating the information of the same user from different OSN sites, one could infer more about a user. Cross-site information aggregation is useful for different parties, including the OSN service providers and OSN application providers. They can use the cross-site linking function to understand their users better, and improve the user experience from various aspects, such as friend suggestion, point-of-interest recommendation, personalized advertising, and malicious account detection.

Table 2 shows an example of cross-site information aggregation, using the information fields of Foursquare and Twitter. If a user exposed a link that connects his Foursquare profile and his Twitter account, we aggregate the information fields of his profiles from both sites, and forming an aggregate table. Intuitively, rather than focusing on a single OSN site, cross-site information aggregation allows us to obtain the contents of more information fields of a user.

We conduct a gender-based analysis of Twitter as a viable example of the cross-site information aggregation. As we mentioned in § 3.2.1, gender-based study is very important for different social networks. For Sina Weibo, the second largest microblogging service in the world, there are some gender-based analysis work such as [17, 5], as the gender information is available in every Weibo user's profile. However, users are not able to specify the gender information on Twitter. As a result, gender-based user behavior study on Twitter is very difficult. [19, 22] used a user's first name to infer the user's gender. However, such gender estimation is inaccurate due to the large number of unisex names in each language.

[9]Instead of specifying "first name" and "last name" separately, Twitter only has a "full name" field. Besides, the gender information is not included in Twitter profiles.

Cross-site information aggregation can solve this problem in an automatic yet accurate way, as the gender information is publicly available in Foursquare. We start from the Foursquare users who have added their Twitter accounts to their profiles. Using the aforementioned distributed crawling framework, we can obtain these users' Twitter profiles. In total, we have collected 8.07 million Twitter profiles. We extract the gender information of each user from his Foursquare profile. Among these Twitter users, 54.28% of them are male, 42.08% of them are female, and the rest 3.64% of them choose to hide their gender information. We pick three key information fields in Twitter profiles for our gender-based analysis, i.e., the number of tweets a user has published ("statuses_count"), the number of tweets a user has favorited ("favorites_count"), and the number of public lists a user involved in ("listed_count"). These fields indicate a user's activity on Twitter. Fig. 5 shows the CDF of these three metrics, by separating male and female users.

According to Fig. 5(a), we can see that female users publish more tweets than male users. The median number of published tweets of male and female users are 999 and 1948, respectively. In other words, we find that female users are more talkative in Twitter.

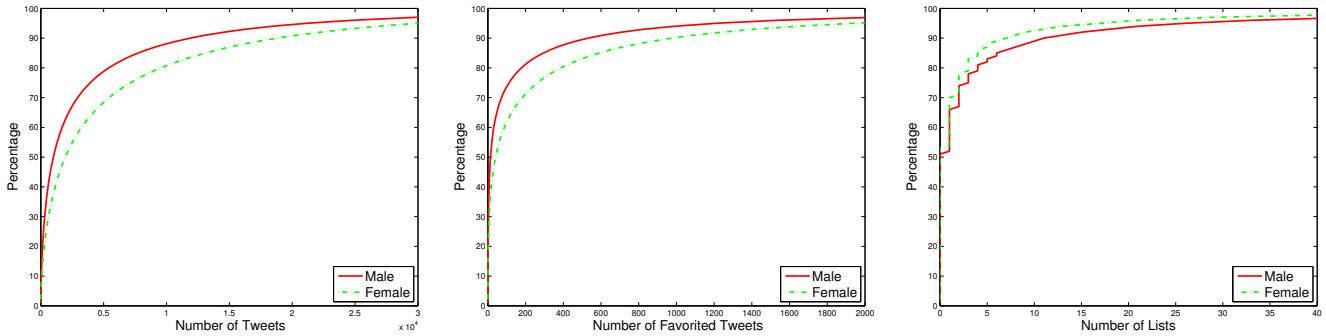
People can favorite a tweet by clicking a small star icon next to the tweet. From Fig. 5(b), we can see that female users have also added more tweets into their favorite lists than male users. The median number of favorited tweets of male and female users are 14, and 41, respectively. Therefore, female users are also more active in using the "favorite" function.

People can create a "list" by adding some Twitter users into it. The list timeline will be composed of a stream of tweets published by the added Twitter users. According to Fig. 5(c), we have found that male users are involved in more lists than female users. The 90th percentiles of the number of the lists male and female users involved in are 11 and 7, respectively. In addition, we are aware that more than half male users and female users have not been added by any public list.

There are some optional fields in a Twitter user's profile. We pick three representative optional fields, i.e., URL, description, and location. The URL field records a web page address of the user. The description field is a self-introduction of the user. The location field records the user's current geo-location. We study how much percent of male and female users have enabled each field. The results are shown in Table 3. We can see a clear difference between male users and female users in terms of adding a URL to their profiles. Male users have a higher probability to add such a URL. Differently, we also find that for the rest two fields, there is little gender difference. Both male and female users have a nearly 63% probability of adding a description, and a nearly 57% probability to add the location information.

6. RELATED WORK

Some previous works have studied cross-OSN linking. Ottoni et al. [12] studied the user behavior across Pinterest and Twitter. Chen et al. [2] have studied cross-OSN links between Google and 9 other OSNs. However, their studies were based on a very small sampled data set, with 30K, and 35K users, respectively. The sampling methods they used are biased. Differently, we present a holistic study of the



(a) Number of Published Tweets (b) Number of "Favorited" Tweets (c) Number of Lists

Figure 5: Gender-based Analysis of Twitter

cross-site linking function based on the entire Foursquare user base, and accordingly we can have accurate results for the entire user population.

Wang et al. [18] compared a series of user activity across Foursquare, Facebook, and Twitter using 200K randomly sampled users. In our work, we have explored more on the cross-site linking function, and addressed a number of unexplored viable issues, such as the distribution of different linking options, behavioral difference among users with different linking options, and user privacy. Also, we proposed cross-site information aggregation.

Goga et al. [4] and Liu et al. [9] investigated how to identify accounts on different OSN sites that all are owned by the same user. Both of these two solutions can achieve a high accuracy. For the OSN sites which do not support the cross-site linking function, their solutions are very useful for correlating users across OSN sites.

7. CONCLUSION AND FUTURE WORK

In this paper, we conduct a comprehensive study on the emerging cross-site linking function on OSN sites. We use Foursquare as an example site to conduct a measurement-based study on the entire Foursquare user base. Our results have revealed a number of unknown findings about this function. Moreover, we examine the cross-site links between Foursquare and two popular OSN sites, i.e., Facebook, and Twitter.

The major findings of this paper are as follows:

- About 60% of Foursquare users have enabled the cross-site linking function. These users are more active than other users, in terms of both content generation, and making social connections.
- Adding contents to an optional field in a Foursquare user's public profile indicates a higher probability of activating the cross-site linking function.
- If a Foursquare user has linked his account to Facebook, he will have a high chance to provide consistent information to both Foursquare and Facebook.
- The use of cross-site information aggregation helps us investigate the gender difference in using Twitter. According to our analysis, female users publish and favourite more tweets than male users, while male users are involved in more public Twitter lists.

We believe that there would be a number of interesting directions to better understand the cross-site linking function. We list them as our future work:

- Besides the three sites studied in this paper, we plan to investigate cross-site links among more mainstream OSN sites. By analyzing real data collected from multiple OSN sites, we aim to discover general patterns to characterize cross-OSN links.
- Without a user's permission, we are only able to access the publicly viewable part of his social network data. For example, we can view a Foursquare user's public profile, but his check-in history is only accessible by his friends. We plan to conduct a volunteer-based study to analyze the complete social network data of a set of users. This could allow a deep investigation into cross-site information aggregation. For example, we can reconstruct a user's mobility pattern [24] by aggregating the detailed geographic data from different OSN sites.
- We aim to explore the possibility of developing practical services/applications based on cross-site links. For example, malicious account detection is a critical challenge to most of the OSN providers. Previous solutions [1, 23] are based on users' social connections and activities within a single OSN site. With the user-created cross-site links, an OSN provider can further use additional information from other linked OSN sites to improve malicious account detection.

8. ACKNOWLEDGMENTS

We are grateful to Xiaohan Zhao and Tianyi Wang for their comments.

9. REFERENCES

[1] Q. Cao, M. Sirivianos, X. Yang, and et al. Aiding the Detection of Fake Accounts in Large Scale Social Online Services. In *Proc. of NSDI*, 2012.

[2] T. Chen, M. A. Kaafar, and et al. Is More Always Merrier? A Deep Dive Into Online Social Footprints. In *Proc. of ACM WOSN*, 2012.

[3] C. Ding, Y. Chen, and X. Fu. Crowd Crawling: Towards Collaborative Data Collection for Large-scale Online Social Networks. In *Proc. of ACM COSN*, 2013.

[4] O. Goga, G. Friedland, and et al. Exploiting Innocuous Activity for Correlating Users Across Sites. In *Proc. of WWW*, 2013.

[5] W. Guan, H. Gao, and et al. Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events. *Physica A: Statistical Mechanics and its Applications*, 395(0):340–351, 2014.

[6] L. Jin, Y. Chen, and et al. Understanding User Behavior in Online Social Networks: A Survey. *Communications Magazine, IEEE*, 51(9):144–150, 2013.

[7] J. Kulshrestha, F. Kooti, and et al. Geographic Dissection of the Twitter Network. In *Proc. of AAAI ICWSM*, 2012.

[8] H. Kwak, C. Lee, and et al. What is Twitter, a Social Network or a News Media? In *Proc of WWW*, 2010.

[9] S. Liu, S. Wang, and et al. HYDRA: Large-scale Social Identity Linkage via Heterogeneous Behavior Modeling. In *Proc. of ACM SIGMOD*, 2014.

[10] A. Noulas, S. Scellato, and et al. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of AAAI ICWSM*, 2011.

[11] N. O'Hare and V. Murdock. Gender-based Models of Location from Flickr. In *Proc. of the ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, 2012.

[12] R. Ottoni, D. de Las Casas, and et al. Of Pins and Tweets: Investigating how users behave across image- and text-based social networks. In *Proc. of AAAI ICWSM*, 2014.

[13] D. Quercia, M. Bodaghi, and J. Crowcroft. Loosing "Friends" on Facebook. In *Proc. of ACM WebSci*, 2012.

[14] S. Scellato, A. Noulas, and et al. Socio-spatial Properties of Online Location-based Social Networks. In *Proc. of AAAI ICWSM*, 2011.

[15] M. Thelwall. Social networks, gender and friending: An analysis of MySpace member profiles. *Journal of the American Society for Information Science and Technology*, 59(8):1321–1330, 2008.

[16] A. Vasalou, A. N. Joinson, and D. Courvoisier. Cultural differences, experience with social networks and the nature of "true commitment" in Facebook. *International Journal of Human-Computer Studies*, 68(10):719–728, 2010.

[17] K. wa Fu and M. Chau. Reality Check for the Chinese Microblog Space: A Random Sampling Approach. *PLoS ONE*, 8(3):e58356, 2013.

[18] P. Wang, W. He, and J. Zhao. A Tale of Three Social Networks: User Activity Comparisons across Facebook, Twitter, and Foursquare. *Internet Computing, IEEE*, 18(2):10–15, 2014.

[19] W. Wang, L. Chen, and et al. Cursing in English on Twitter. In *Proc. of ACM CSCW*, 2014.

[20] Y.-C. Wang, M. Burke, and R. E. Kraut. Gender, Topic, and Audience Response: An Analysis of User-generated Content on Facebook. In *Proc. of ACM CHI*, 2013.

[21] C. Wilson, B. Boe, and et al. User Interactions in Social Networks and Their Implications. In *Proc. of ACM EuroSys*, 2009.

[22] C. Xiao, L. Su, J. Bi, Y. Xue, and A. Kuzmanovic. Selective Behavior in Online Social Networks. In *Proc. of WI-IAT*, 2012.

[23] Z. Yang, C. Wilson, and et al. Uncovering Social Network Sybils in the Wild. *ACM Trans. Knowl. Discov. Data*, 8(1):2:1–2:29, 2014.

[24] Z. Zhang, L. Zhou, and et al. On the Validity of Geosocial Mobility Traces. In *Proc. of ACM HotNets*, 2013.