Understanding Cross-Site Linking in Online Social Networks

QINGYUAN GONG, YANG CHEN, and JIYAO HU, School of Computer Science, Fudan University, China, and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China
QIANG CAO, Department of Computer Science, Duke University, USA
PAN HUI, Department of Computer Science, University of Helsinki, Finland, and CSE Department, Hong Kong University of Science and Technology, Hong Kong
XIN WANG, School of Computer Science, Fudan University, China, and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China

As a result of the blooming of online social networks (OSNs), a user often holds accounts on multiple sites. In this article, we study the emerging "cross-site linking" function available on mainstream OSN services including Foursquare, Quora, and Pinterest. We first conduct a data-driven analysis on crawled profiles and social connections of all 61.39 million Foursquare users to obtain a thorough understanding of this function. Our analysis has shown that the cross-site linking function is adopted by 57.10% of all Foursquare users, and the users who have enabled this function are more active than others. We also find that the enablement of cross-site linking might lead to privacy risks. Based on cross-site links between Foursquare and external OSN sites, we formulate cross-site information aggregation as a problem that uses cross-site links to stitch together site-local information fields for OSN users. Using large datasets collected from Foursquare, Facebook, and Twitter, we demonstrate the usefulness and the challenges of cross-site information aggregation. In addition to the measurements, we carry out a survey collecting detailed user feedback on cross-site linking. This survey studies why people choose to or not to enable cross-site linking, as well as the motivation and concerns of enabling this function.

CCS Concepts: • **Human-centered computing** → **Social networking sites**; • **Networks** → *Online social networks*;

Additional Key Words and Phrases: Online social networks, cross-site linking, measurement, survey

25

This work is supported by the National Natural Science Foundation of China (No. 61602122, No. 71731004), Natural Science Foundation of Shanghai (No. 16ZR1402200), Shanghai Pujiang Program (No. 16PJ1400700), and Projects 26211515 and 16214817 from the Research Grants Council of Hong Kong. A preliminary version of this article has been published in Proceedings of the 8th Workshop on Social Network Mining and Analysis (SNAKDD'14).
Authors' addresses: Q. Gong, Y. Chen (corresponding author), J. Hu, and X. Wang, School of Computer Science, Fudan University, Shanghai 200433, China, and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai 200433, China; emails: {gongqingyuan, chenyang, hujy13, xinw}@fudan.edu.cn; Q. Cao, Department of Computer Science, Duke University, Durham, NC 27708, USA; email: qiangcao@cs.duke.edu; P. Hui, Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland, and Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; email: panhui@cs.helsinki.fi.

1559-1131/2018/09-ART25 $15.00
https://doi.org/10.1145/3213898

ACM Reference format:
Qingyuan Gong, Yang Chen, Jiyao Hu, Qiang Cao, Pan Hui, and Xin Wang. 2018. Understanding Cross-Site Linking in Online Social Networks. *ACM Trans. Web* 12, 4, Article 25 (September 2018), 29 pages.
https://doi.org/10.1145/3213898

1 INTRODUCTION

The blooming of social networks [26] has brought a significant portion of social activities online. A person can interact with friends on Facebook [62], share news on Twitter [29], set up a job-search profile on LinkedIn [49], and publish locations (e.g., "check in") on Foursquare [56]. Advanced online social networking (OSN) sites, such as Foursquare [38, 45, 56], Pinterest [22], Quora [55], Google+ [19], and Momo [7], have introduced a *cross-site linking* function to interconnect a user's accounts on multiple sites.

The linkage of user accounts with cross-site linking has brought tangible benefits to users and influenced the way that OSNs are being used. First, cross-site linking can reduce manual work for cross-site content posting, when a user wishes to publish content (e.g., a photo) to multiple OSN sites. For example, if a user does a "check-in" on Foursquare, he or she can publish this message to Facebook and Twitter [23, 38, 42, 45, 46] automatically via cross-site linking. Second, cross-site linking is useful for mirroring social connections across sites [70], when a user wishes to stay in touch with his or her friends on multiple OSNs. For example, after linking to his or her Facebook account, a Foursquare user can import contacts directly from his or her Facebook friend list through a single click. Third, cross-site linking is a simple yet effective way for self-presentation in the digital era. Cross-site links often connect more user information than what can be found on a single OSN site [67].

Existing work has studied data correlation among multiple sites, e.g., identifying accounts on different OSN sites belonging to the same user [25, 32], evaluating the correlations of a user's activities on different OSN sites [13, 40, 58, 70], and assembling user profiles using information available on multiple OSN sites [8, 14, 15, 67]. However, these studies are based on datasets with only a subset of users, which can lead to inaccurate analysis results. As a result, there lacks a systematic study to understand "cross-site linking" as an integrated whole.

This work carries out large-scale measurements, extensive linkage analysis, and a global user survey with the objective to attain a thorough understanding of cross-site linking, including its adoption by OSN users and its implications on OSN applications and user privacy. In particular, we crawled profiles and social connections of all Foursquare users as of August 2015 and refer to this complete Foursquare user base for estimating the adoption of its cross-site linking function. We conduct group-based analysis to uncover the distribution of user-chosen cross-site linking options and their correlation with users' demographic information, online behavior, registration time, and privacy concerns. We further extract features based on the correlated factors, evaluate their ability in predicting the enablement of cross-site linking, and identify a set of discriminative features.

We study information linkage across sites using the datasets we collected. We crawled cross-site hyperlinks on Foursquare and retrieved the linked Facebook and Twitter profiles and the social activity counters of these profiles. We formulate *cross-site information aggregation* that uses cross-site links to stitch together site-unique information fields for OSN users. As an example application of cross-site information aggregation, we develop a gender-based analysis of Twitter to investigate talkativeness of users by borrowing gender information from Foursquare. This Twitter analysis is otherwise impossible due to the absence of gender information on Twitter. Therefore, it sheds light on how an OSN application could benefit from cross-site linking.

We use an online survey to understand user motivations and activity characteristics on cross-site linking, e.g., asking an OSN user why and why not he or she chooses to enable this function. We design a 3-minute survey covering questions not adequately addressed by measurements and perform an in-depth analysis of the survey responses. This survey has been completed by a few hundred participants from around the world. Our analysis of the collected responses has identified detailed uses of cross-site linking, revealed potential privacy concerns and user agnosticism of the cross-site linking function, and clarified scenarios where information mutation and inconsistency across OSN sites can occur.

This work targets a comprehensive study of OSN cross-site linking. We have made the following key contributions.

- We conduct a measurement study on the entire user base of Foursquare (in Section 3). To the best of our knowledge, our work is the first measurement study on cross-site linking using the entire user population of a mainstream OSN site. We have found that about 57.10% of Foursquare users have enabled cross-site linking. In addition, we introduce a supervised machine-learning model to study the predictability of the enablement of cross-site linking using features from user profiles and activities.
- We go beyond the Foursquare dataset and further explore the cross-site links between Foursquare and two leading OSN sites, i.e., Facebook and Twitter (in Section 4). We formulate the cross-site information aggregation problem and outline the possibilities and challenges of consolidating user information from multiple OSN sites.
- We conduct an online survey with 369 participants from around the world. Participants have provided their detailed opinions about cross-site linking (in Section 5). This survey is complementary to our measurements. Findings from the survey are consistent with measurement results.

2 THE CROSS-SITE LINKING FUNCTION ON FOURSQUARE

Among the OSN sites that support the cross-site linking function, we choose Foursquare for the following reasons. First, Foursquare is a representative location-based social network (LBSN) service, and it is one of the most popular OSN sites. Second, Foursquare supports cross-site linking. It allows its users to link their accounts to both Facebook and Twitter. Third, every Foursquare user has a publicly accessible profile page. We can obtain a user's links to Facebook/Twitter on this page.

In this section, we first present a quick overview of the Foursquare social network and an example on how a Foursquare user profile is linked to the leading OSN sites, i.e., Facebook and Twitter. In addition, we describe how we crawl the profiles of all Foursquare users in a collaborative way.

2.1 Overview of Foursquare and Its Cross-Site Linking Function

Foursquare has been established as an LBSN since 2009. A Foursquare user can leave a "tip" for any venue, which is visible to the public. Also, using the recently released "Swarm" application [9], a user can use a "check-in" function to claim that he or she is visiting a selected nearby place. Since Foursquare is a social networking service, its users can "follow" each other. For example, user A can follow user B by adding B to A's list of followings. Accordingly, A will become one of B's followers. The following relationship of Foursquare can be modeled as a directed graph $G = (V, E)$. V is the set of all Foursquare users, and E is the set of social connections between these users. A node in V denotes a Foursquare user, and an edge in E represents a social connection. When A follows B, a directed edge (v_A, v_B) will be added to E.

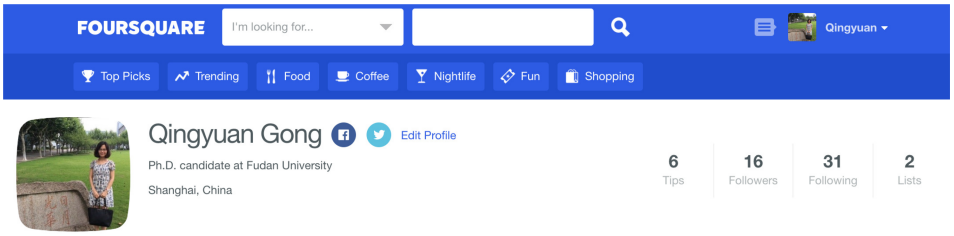


Fig. 1. A Foursquare user's public profile page.

Each Foursquare user has a public profile, which can be accessed by everybody. The information fields in the user's profile, such as his or her first name, number of check-ins, and number of tips, are available to the public. Also, this profile page provides hyperlinks to the lists of his or her followers and followings. Optionally, a user can choose to add the last name, a profile photo, a biography, the location, and the gender information. A user can link his or her public profile to his or her accounts on Facebook and Twitter through two steps. First, the user needs to submit his or her Facebook/Twitter user ID to Foursquare. This can be done when he or she registers an account, or anytime later. Second, by using the OAuth 2.0 framework, Foursquare creates the cross-site link after it verifies that the user owns the submitted Facebook/Twitter account. 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. Figure 1 shows a Foursquare user's public profile page.[1] This user has linked her Facebook and Twitter accounts to her profile. By clicking the Facebook icon, we can access her Facebook profile page. Similarly, by clicking the Twitter icon, we can visit her Twitter profile page.

2.2 Data Collection

Given the large user population, previous work is only able to use a small subset of users to conduct data-driven studies of Foursquare [38, 45, 56]. In our work, we aim to analyze the entire Foursquare user base, which overcomes the disadvantages of biased sampling. However, Foursquare is growing rapidly, and had attracted more than 60 million registered users by July 2015. Quickly crawling massive data is challenging, as the data fetching speed of a single IP address is strictly bounded by a predefined threshold. Using the crowd crawling approach proposed by Ding et al. [11], we split the overall crawling task into some independent small tasks and used 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 numeric user ID. If we know the ID of a user, we can access the URL http://foursquare.com/user/ID to view his or her profile page, which is open to the public. This ID is assigned in an ascending order. In other words, if user A registers earlier than user B, his or her numeric user ID would be smaller than user B's ID. Therefore, we could obtain the maximum user ID by registering a new account. Note that not all IDs between one and the maximum user ID are assigned to users. Some IDs are reserved for venues (points of interest) or brands (business accounts). Also, some IDs are unused. Scanning through all user IDs between one and the maximum user ID can ensure that our crawling covers the entire Foursquare user base, instead of focusing on a biased subset.

We implemented a Python-based distributed crawler. This crawler applied the standard Foursquare API to download the data of each user. As Foursquare does not allow too many concurrent requests from a same IP address within a short time period, we have to allocate many servers with

[1] https://foursquare.com/user/396301975.

Table 1. Cross-Site Linking Options

Twitter	Facebook	Linking Option	Percentage
Y	N	TW only	3.99%
N	Y	FB only	41.71%
Y	Y	FB+TW	11.40%
N	N	Neither	42.90%

different public IP addresses to form a collaborative crawling cluster. We used 80 virtual instances in the East US data center of the Microsoft Azure platform. We registered a Foursquare account on July 25, 2015, and its ID was 137146878. We adopted it as the maximum user ID before we started the data crawling. We divided the whole ID space into 80 chunks evenly, and each virtual instance was responsible for crawling one chunk of IDs. We carefully controlled the request rate of each crawler. Putting all crawlers together, our data collection introduced 10.89KB/s traffic to Foursquare, and 99.03KB/s traffic from Foursquare. Therefore, our measurement introduced a limited network overhead for the Foursquare service. The crawling lasted more than 1 month, from July 25 to August 27 in 2015. We successfully crawled the public profiles of 61.39 million Foursquare users. These users have conducted 7.46 billion check-ins and published 57.62 million tips. We believe that we have crawled the entire set of Foursquare users' profiles and social connections at that time.

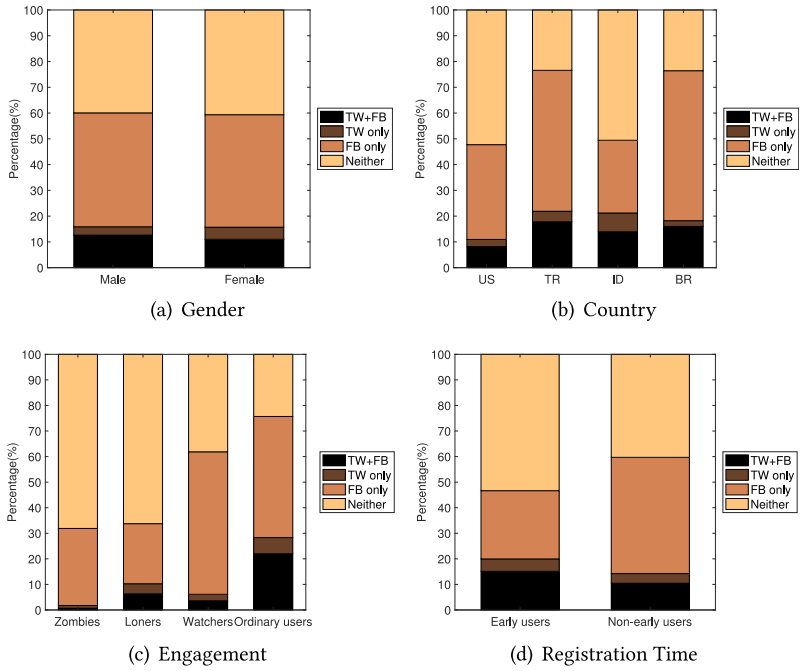
Note that we respect the privacy of the Foursquare users. We only collected publicly accessible information. By analyzing the crawled data, we were able to extract the values of information fields of each user, e.g., gender, number of check-ins, number of tips, linked Facebook account, and linked Twitter account. Also, for each user, we obtained full lists of followers and followings. Therefore, we were able to construct the complete social graph of Foursquare with 61.39 million nodes and 2.67 billion edges. To protect the users' privacy, we anonymized the ID of each user. Meanwhile, we stored and analyzed the data in an offline environment. Our study was reviewed and approved by the Research Department of Fudan University.

3 CROSS-SITE LINKING OPTIONS

In this section, we analyze the cross-site linking function on Foursquare and examine all the 61.39 million Foursquare users. In Section 3.1, we calculate the percentages of users that enable explicit linking to different external OSN sites, such as Facebook and Twitter. In Section 3.2, through a group-based analysis, we examine user preference on cross-site linking. We further investigate the behavioral difference among users with different linking options in Section 3.3 and with different registration time in Section 3.4, respectively. In Section 3.5, we study the impact of user privacy concerns on the use of cross-site linking. In Section 3.6, we explore to what extent the enablement of cross-site linking can be predicted.

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 OSN sites. 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" represent 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 the users are assigned to the "Neither" group, as they do not link to any external account.



(a) Gender (b) Country

(c) Engagement (d) Registration Time

Fig. 2. Group-based analysis (gender/country/engagement/registration time).

By examining all the crawled Foursquare user profiles, we calculate the account percentage of each linking option in Table 1. We can see that about 57.10% of Foursquare users have linked at least one account on Facebook or Twitter. Specifically, 53.11% of users have linked their Facebook accounts, and 15.39% of users have linked 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 helps us 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, engagement, and registration time, respectively.

3.2.1 Gender. It has been reported that gender has influence on user behavior in OSNs. Previous work studied the gender influence on Facebook [43, 60], Flickr [39], Twitter [59], and MySpace [50]. Among the collected Foursquare user profiles, 51.17% are male and 41.42% are female. The rest, 7.41%, of the users do not disclose their gender information. As can be seen in Figure 2(a), there is very little difference between male 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 [28] and Facebook [51]. We obtain a Foursquare user's country according to the "Location" field in his or her crawled profile. By using the Google Geocoding API,[2] we are able to obtain the country information of 83.32% of

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

Foursquare users, while 6.49% of Foursquare users' country information cannot be determined. The rest, 10.19%, of Foursquare users choose to hide their residential location. According to our analysis, the top four countries that have the largest Foursquare user populations are the United States (US), Turkey (TR), Indonesia (ID), and Brazil (BR). They cover 24.16%, 11.60%, 7.21%, and 5.95% of Foursquare users, respectively. As shown in Figure 2(b), users from different countries have quite different distributions of cross-site linking options. Among these countries, Turkey and Brazil have the highest percentages (76.57% and 76.44%, respectively) of users that link their Foursquare accounts to other OSNs. In contrast, the United States has the lowest percentage (47.74%) of users to use this function.

3.2.3 Engagement. We also group users by their engagement. Since Foursquare is an LBSN service, content generation (e.g., performing check-ins and leaving tips) and social connectivity play key roles in the engagement of users. Based on these two factors, we divide all users into the following four groups:

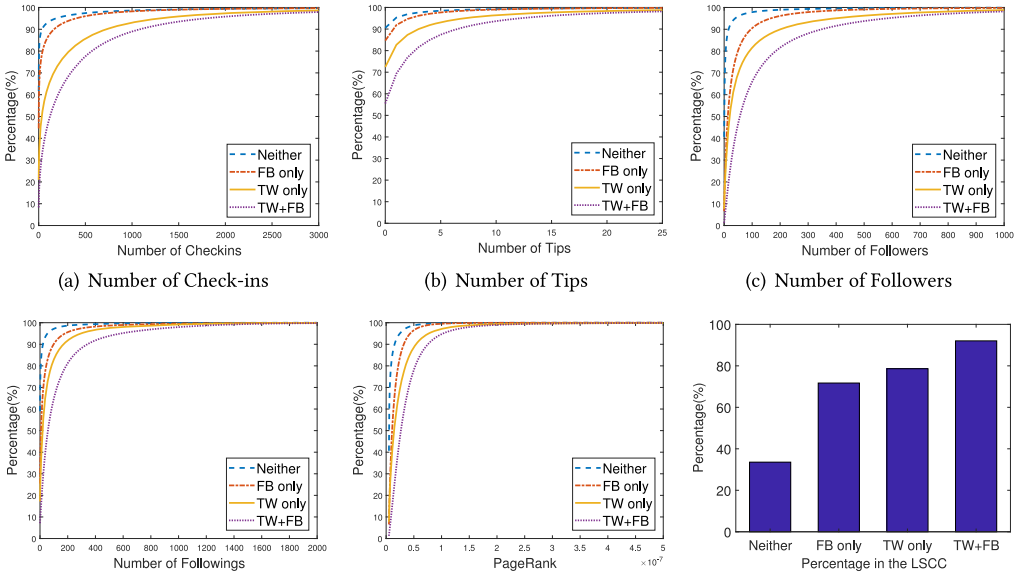
- *Zombies*: Users who have not followed any other user and have never posted any check-in or tip. This group of users have no interaction with other Foursquare users. Many of them are likely to be crawlers or newly registered accounts.
- *Loners*: Users who have not followed any other user but have published at least one check-in or tip. Such users do not connect with other Foursquare users but still perform check-ins and leave tips.
- *Watchers*: Users who have followed at least one user and have published nothing. These users are silent, and they use only the social networking features of Foursquare.
- *Ordinary users*: For the rest of the users, we put them into the fourth group. They have followed at least one user and have either performed check-ins or left tips.

The percentages of these four groups of users are 27.52%, 9.91%, 17.26%, and 45.31%, respectively. According to Figure 2(c), 68.11% of zombies and 66.23% of loners have not enabled cross-site linking, since they are socially isolated. Differently, 61.88% of watchers and 75.70% of ordinary users have enabled cross-site linking. Also, we compare watchers with ordinary users. On one hand, 59.34% of watchers and 69.36% of ordinary users have linked their accounts on Facebook. As users from these two groups are willing to connect with other users, they both have significant percentages of accounts that are linked to Facebook. On the other hand, 6.14% of watchers and 28.33% of ordinary users have linked their accounts on Twitter. Watchers are silent and have not published anything. It is likely that they are less motivated to join Twitter, which is known as a news spreading platform [29].

3.2.4 Registration Time. We also group users according to their registration time. As mentioned in Section 2.2, a user who registers earlier has a smaller user ID. Therefore, by referring to the user IDs, we can group all users into two groups according to the registration time. The earliest 20% of users are denoted as "early users," and the other 80% are denoted as "nonearly users." As shown in Figure 2(d), nonearly users have a higher probability of enabling cross-site linking. This indicates a growth in popularity of the cross-site linking function.

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 and social connectivity. For content generation, we examine two key metrics on Foursquare, i.e., number of check-ins and number of tips. Regarding social connectivity, we study a user's number of followers, number of followings, and PageRank value [41]. PageRank is a metric that quantifies the



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

(d) Number of Followings (e) PageRank (f) % in LSCC

Fig. 3. Linking options versus user behavior (content generation/social connectivity).

Table 2. Spearman's Rank Correlation Coefficients of Metric Pairs

Metric	# of Check-ins	# of Tips	# of Followings	# of Followers	PageRank
# of Check-ins	1.0	0.464	0.632	0.633	0.673
# of Tips	0.464	1.0	0.330	0.314	0.362
# of Followings	0.632	0.330	1.0	0.775	0.753
# of Followers	0.633	0.314	0.775	1.0	0.938
PageRank	0.673	0.362	0.753	0.938	1.0

importance of different nodes in a network. It has been used for ranking websites for the Google search engine and for quantifying the user influence [29, 47, 48, 61] in OSNs. For any node of a selected network, its PageRank value is between 0 and 1. A larger PageRank value indicates that the corresponding node is more important. Based on the Foursquare social graph, we are able to compute the PageRank values of all users. Similar to other OSNs [17], the largest strong connected component (LSCC) of Foursquare has covered 57.94% of the entire user base. For each linking option group, we also examine the percentage of nodes within the LSCC. The results are shown in Figure 3. For the five metrics we studied in Figures 3(a) to 3(e), we use Spearman's rank correlation coefficient to examine the monotonic correlation between each pair of them. The value of a Spearman's rank correlation coefficient is between $[-1, 1]$. A value close to 1 means a similar rank between two metrics, and a value close to -1 indicates that the two metrics have a negative correlation between the ranks. Also, a value close to 0 represents that the two metrics have almost no correlation. Results are shown in Table 2. We find that different metric pairs reveal different levels of correlation. The number of followers and PageRank are highly correlated, with a Spearman's rank correlation coefficient of 0.938.

For the four linking option-based user groups, i.e., "TW+FB," "TW only," "FB only," and "Neither," we use Welch's t-test to study the difference between each pair of them concerning different

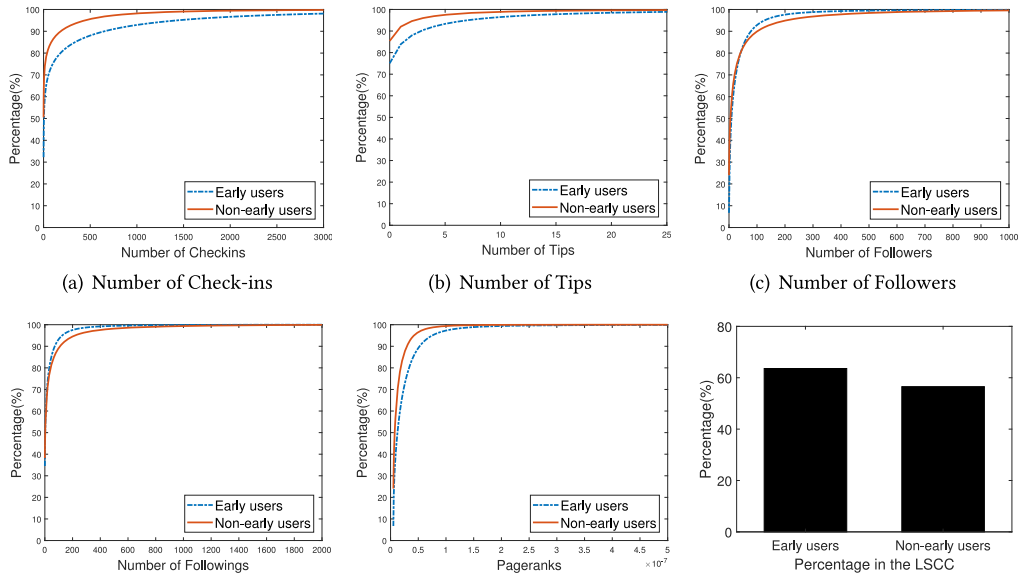
metrics, i.e., number of check-ins, number of tips, number of followers, number of followings, and PageRank. Given a selected metric, we examine each two user groups and see whether any group pairs are statistically different according to Welch's t-test. For each pair, we calculate the corresponding p-value. If the p-value is smaller than 0.05, we can conclude that these two groups are significantly different. By going through all relevant metrics, we find that for nearly every two user groups in terms of each metric, the corresponding p-value of the Welch's t-test is smaller than 0.001. These results indicate that every two user groups are significantly different. The only exception happens when we consider the "number of tips" metric. When we evaluate the difference between the "TW + FB" group and the "TW only" group using Welch's t-test, the p-value is as large as 0.447. In other words, the difference between these two user groups is not that remarkable in terms of the "number of tips" metric. We believe that it is because the "tips" function is less frequently used. For users in the "TW+FB" group, more than 55.49% of them have not published any tip. For users in the "TW only" group, 72.39% of them have not published any tip. Therefore, considering the "number of tips" metric, the difference between the "TW + FB" and "TW only" groups is not significant.

In Figure 3(a) and Figure 3(b), we show the cumulative distribution function (CDF) of the number of check-ins and tips. Although all Foursquare users have identical functionality in performing check-ins and leaving tips, the cross-site linking function will automatically post the newly published contents in Foursquare to Facebook and/or Twitter. Therefore, the contents will be read by a more prospective audience, and accordingly these users have more motivation to publish. We observe that the users who have enabled cross-site linking are more "active" in content generation; i.e., they conduct check-ins more frequently and leave more tips. Particularly, the "FB+TW" users are the most active in terms of content generation. We also see that in general, the "TW only" users publish more than the "FB only" users. We speculate that it is because "TW only" users have more incentive to publish. As a news spreading platform [29], Twitter can quickly spread Foursquare users' check-ins or tips as publicly accessible tweets through the microblogging network. In contrast, a Facebook user's posts are only visible to friends by default.[3] According to the study by Minkus et al. [36], a dominant portion of Facebook users have chosen to only allow friends to see their posts, which limits the number of possible audience.

From Figure 3(c), we find that users who have enabled cross-site linking have a larger average number of followers than those who have not. Among the four groups, "FB+TW" users have the highest average number of followers. Meanwhile, the average number of followers 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 Figure 3(a) and Figure 3(b). According to Figure 3(d), we can reach the same conclusion according to the distribution of the number of followings. As we have mentioned in Section 1, a key advantage of cross-site linking is known as social bootstrapping [70]. In other words, the enablement of cross-site linking can help Foursquare users import contacts from Facebook/Twitter, resulting in a larger average number of followers and followings. In Figure 3(e) and Figure 3(f), we also observe a similar trend.

Since the check-ins in Foursquare are not publicly accessible, many of the existing works use Twitter to obtain the check-in history of Foursquare users [23, 38, 42, 45, 46]. According to our findings in Figure 3, the users who have linked their Twitter accounts to Foursquare are more active in generating content and creating social connections. Therefore, when using such user datasets for user behavior study, we need to consider such bias carefully. In other words, we should not use this set of users directly to represent the behavior of the entire Foursquare population.

[3] https://www.theverge.com/2014/5/22/5739744/facebook-changes-default-privacy-of-posts-from-public-to-friends.



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

(d) Number of Followings (e) PageRank (f) % in LSCC

Fig. 4. Registration time versus user behavior (content generation/social connectivity).

3.4 Behavioral Difference among Users with Different Registration Time

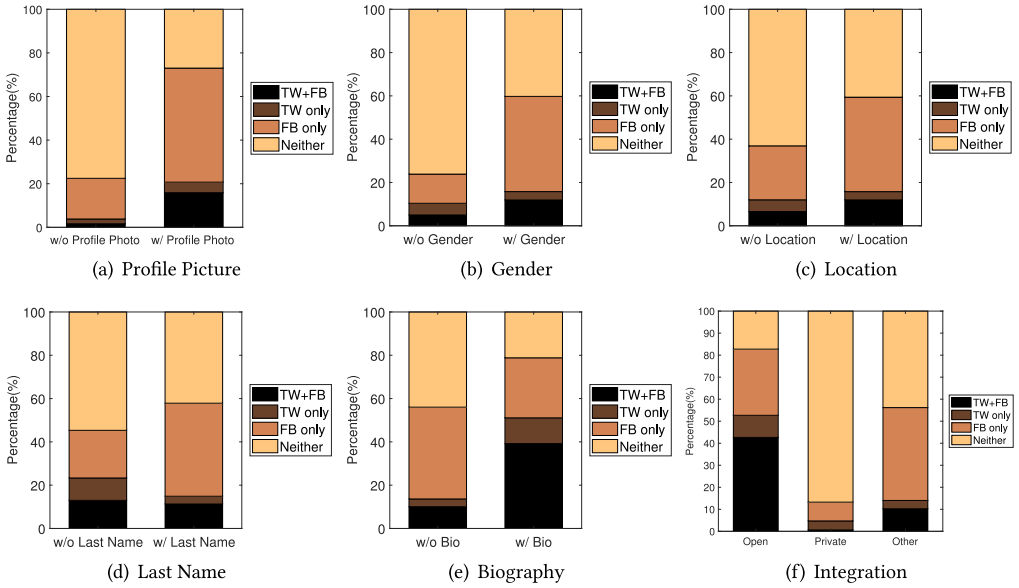
In this subsection, we group the users according to their registration time and see the behavioral difference between early users and nonearly users. As in Section 3.3, we study the six key metrics, i.e., number of check-ins, number of tips, number of followers, number of followings, PageRank, and percentage of nodes within the LSCC. For each metric, these two user groups are significantly different (p-value < 0.001, Welch's t-test). The results are shown in Figure 4.

According to Figure 4(a) and Figure 4(b), early users have published more check-ins and tips. Differently, according to Figure 4(c) and Figure 4(d), nonearly users have higher average numbers of followers and followings. As discussed in Section 3.2.4, nonearly users have a higher probability of enabling cross-site linking. The cross-site links are helpful for them to get bootstrapped by importing the contact lists from well-established OSNs like Facebook and Twitter. Figure 4(e) shows that early users have a higher average number of PageRank. Similarly, as shown in Figure 4(f), early users have a higher probability of being involved in the LSCC.

3.5 Impact of User Privacy Concerns

There are about 57.10% users on Foursquare who have enabled the cross-site linking function. The 42.90% of Foursquare users who have not enabled cross-site linking can be divided into two categories. Some of them might choose not to link their Facebook/Twitter accounts to their Foursquare accounts, while the rest of them might not even have accounts on Facebook/Twitter. To estimate how many Foursquare users have accounts on Facebook or Twitter, we refer to social hub services, a.k.a., social directory services [68], which allow a user to list his or her accounts on different OSN sites on his or her profile page. In particular, we focus on about.me,[4] a representative social hub service that has attracted millions of users around the world. To obtain a massive set of about.me user profiles, we used a list of 1.11 million about.me user IDs collected by Cao et al. [5]. We have

[4]http://about.me/.



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

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

Fig. 5. Cross-site linking versus user privacy.

crawled all these about.me user profiles using a Python-based crawler. Among the users who have accounts on Foursquare, 76.79% of them have accounts on both Facebook and Twitter, while 98.73% of them have accounts on either Facebook or Twitter. These numbers show the dominance of Facebook and Twitter in OSNs. We believe that a major portion of Foursquare users without enabling cross-site linking still belong to the first category. Intuitively, if a user cares about his or her privacy and does not want to expose additional personal information to the public, he or she might have concerns to enable cross-site linking. As discussed by Irani et al. [24], enabling an optional information field will lead to a higher chance of privacy leakage. In this subsection, we investigate the impact of user privacy concerns on the use of the cross-site linking function. As discussed by Vasconcelos et al. [52], there is a nonnegligible portion of malicious accounts on Foursquare, posting spam and advertisements. Note that we focus on the entire Foursquare population. Our dataset not only includes legitimate users but also covers fake profiles, social bots, and spammers.

Foursquare allows users to customize their profiles according to their privacy concerns. It provides five optional information fields for a user to set up the public profile, i.e., profile picture, gender, location, last name, and biography. A user can choose whether to make each information field available to the public according to his or her privacy concerns. We summarize our findings as below.

Profile Picture: A user can choose to upload his or her own profile picture. Among all users, 68.49% have uploaded their own profile photos, while 31.51% choose to use the default blank one instead of uploading their own photos. In Figure 5(a), among the users who have uploaded profile photos, 73.03% of them have enabled the cross-site linking function. In contrast, the percentage is only 22.47% for the users who have not uploaded. Therefore, whether or not a personalized profile photo has been uploaded is an indicator for the adoption of cross-site linking.

Gender: Among all users, 7.41% of them have chosen to hide their gender information. According to Figure 5(b), for users who have disclosed their gender information, about 59.76% of them have linked their Facebook or Twitter accounts. In contrast, for users who have hid their gender

information, only 23.86% of them have linked their Facebook or Twitter accounts. Similarly, the availability of the gender information is another indicator for the use of cross-site linking.

Location: Among all users, 10.19% of them have chosen not to disclose their location. According to Figure 5(c), for users choosing to make their location information public, 59.39% of them have linked their accounts. Differently, only 36.91% of the users that choose to hide their locations have linked their accounts.

Last Name: A user is required to provide his or her first name to Foursquare. However, the last-name information is optional. About 93.94% of Foursquare users choose to add their last name to Foursquare profiles, and 6.06% of Foursquare users choose to hide this information. According to Figure 5(d), for users who have provided their last names, 57.86% of them have enabled the cross-site linking function. For users who choose to hide their last names, 45.33% have enabled the cross-site linking function.

Biography: A Foursquare user is allowed to add a description about him- or herself in the optional "Bio" field. Only about 4.68% users have entered content in this field, and the rest, 95.32%, of the users choose to leave this field empty. As we have shown in Figure 5(e), for users who have provided their biographies, as many as 78.84% of them have enabled cross-site linking. For users who have not entered their biographies, only 56.03% of them have enabled cross-site linking.

The above analysis shows that enabling any of these five optional information fields indicates a higher probability of using cross-site linking. Besides studying the five optional information 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 information field. For users who have filled in all five 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 five optional fields, we call them "private users," as they do not reveal any nonmandatory information. The rest of the users are simply denoted as "other users." The percentages of open users, private users, and other users are 3.58%, 0.09%, and 96.33%, respectively. According to Figure 5(f), we can see that 82.72% of the open users have enabled the cross-site linking function, while only 13.25% of the private users have enabled this function.

3.6 Prediction of the Enablement of Cross-Site Linking

In this subsection, we aim to explore the correlation between the selected user activity features and the enablement of cross-site linking. We study the entire Foursquare population as a whole, from a statistical perspective. Among all Foursquare users, there are different reasons for enabling cross-site linking. Although many users apply cross-site linking intentionally, there are also some other users who might activate this function casually. Therefore, the entire Foursquare population is a natural mixture of these two types of users, and we use this complete set of users for our analysis. We investigate to what extent the enablement of cross-site linking can be predicted by referring to the metrics we considered in previous measurements. Based on the observations from our measurement analysis, we identify four sets of features that could help distinguish between users enabling cross-site linking or not.

Social connection features (five features): This set of features includes the social-graph-related metrics used in Section 3.3, i.e., number of followers, number of followings, PageRank, whether the user is within the LSCC, and whether the user is within the largest weakly connected component (LWCC).

Home country features (four features): As discussed in Section 3.2, among the top four countries, users from Turkey and Brazil are more inclined to link their accounts across OSN sites than those from the United States and Indonesia. We introduce four features related to the home

Table 3. Prediction of Linking Options

Algorithm	Parameter	Precision	Recall (TPR)	F1-Score	FPR	AUC
XGBoost	Refer to Table 4	0.816	0.880	0.847	0.264	0.886
RF	95 trees, 6 features/tree	0.798	0.836	0.817	0.281	0.847
J48	Confidence factor = 0.125, Instance/leaf M = 8	0.814	0.880	0.845	0.268	0.874
LIBLINEAR	L2-regularized L2-loss support vector classification, Cost = 7	0.778	0.864	0.819	0.328	0.843
Naive Bayes	Default	0.795	0.797	0.796	0.274	0.833

country, i.e., representing whether the user is from the United States, Turkey, Brazil, or Indonesia, respectively.

Privacy preference features (five features): This set of features reflects a user's preference over the optional information fields. There are five features, indicating whether a user has enabled the biography, profile photo, last name, location, and gender fields, respectively.

User-generated content (UGC) features (two features): This set of features represents a user's activeness in content generation. We use two features, i.e., the number of check-ins and the number of tips.

Supervised machine-learning algorithms are widely used to quantify the statistical correlations between selected features and the categories they belong to. In this article, we classify users into two categories. One, "with linking," category is for users who have enabled cross-site linking. The other, "without linking," category covers users who have not enabled this function. We apply a machine-learning-based binary classifier to see how accurately we can predict the enablement of cross-site linking. We use all 61.39 million Foursquare users as instances to conduct the training, validation, and test. We randomly select 40.93 (66.67%) million users (instances) to form a training and validation dataset, and the rest, 20.46 (33.33%) million, of the users (instances) to form a test dataset. The percentages of the users' linking options in the entire Foursquare set are kept for both the training and validation dataset, and the test dataset. The percentages of users who have and have not enabled cross-site linking are 57.10% and 42.90%, respectively.

A number of classic supervised machine-learning-based algorithms are employed to do the prediction, including C4.5 decision tree (J48) [44], Random Forest [3], and linear support vector machine (LIBLINEAR) [12]. In particular, we experiment with the emerging XGBoost [6], a scalable end-to-end tree boosting system. XGBoost has been widely used in various machine-learning challenges such as Kaggle [6].

The prediction performance of each algorithm is evaluated by the following five representative metrics: precision, recall, F1-score, False-Positive Rate (FPR), and AUC. Precision is valued by the fraction of users predicted as "with linking" accounts who have really enabled the cross-site linking function. Recall, also known as the True-Positive Rate (TPR), means the fraction of "with linking" users who are identified accurately. F1-score is the harmonic mean of the above two metrics. FPR is the ratio between the number of "without linking" users who are wrongly predicted as "with linking" and the total number of users who have not enabled cross-site linking. AUC [16] is the area under the ROC (receiver operating characteristic) curve, which is equivalent to the probability that a selected classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. For each algorithm, once a set of parameters is chosen, we can calculate the above five performance metrics using 10-fold cross-validation. By adjusting the parameters carefully, we record the "best parameters set" that achieves the highest F1-score. We use the test dataset to evaluate the prediction performance of each algorithm based on the set of "best parameters." The results and corresponding parameters are shown in Table 3 and Table 4. Using McNemar's test [34] to calculate the statistical significance, we find that the prediction

Table 4. Parameters Set for XGBoost

Parameter	Value
learning_rate	0.1
min_child_weight	3
max_depth	14
gamma	94.93
subsample	0.8
colsample_bytree	0.8
booster	gbtree
objective	binary:logistic
seed	27
tree_method	gpu_hist
nthread	16

Table 5. χ^2 Statistic

Rank	χ^2	Feature	Feature Category
1	13957560.4832	Number of followers	Social
2	12916174.8448	PageRank	Social
3	9217807.0178	Has_photo?	Privacy
4	8270100.9999	Number of followings	Social
5	7496141.0560	Within_the_LSCC?	Social
6	6682871.8022	Within_the_LWCC?	Social
7	4214867.3672	Number of check-ins	UGC
8	1475568.5373	Has_gender?	Privacy
9	1186996.7389	Number of tips	UGC
10	831553.1457	From_TR?	Country

performance of every two classifiers is significantly different (p-value < 0.001, McNemar's test). We can see that XGBoost performs the best, with an F1-score reaching 0.847 and an AUC of 0.886. The F1-scores and AUC values of the other four algorithms are not as good as XGBoost, but even the lowest F1-score is as large as 0.796, and the lowest AUC is 0.833. According to the prediction results, we can conclude that the features we consider in our measurement work are closely relevant to users' linking options.

To see which features are better correlated with the enablement of cross-site linking, we evaluate the discriminative power of all selected features by the χ^2 (Chi square) statistic [65]. The 10 most discriminative features are shown in Table 5. From the results, we can see that nearly all social graph features are ranked high concerning the influence on the prediction performance. In addition, the number of check-ins, the number of tips, and whether a user has added a profile photo can be taken as important features to predict if a user tends to enable the cross-site linking function.

4 CROSS-SITE INFORMATION AGGREGATION

Our previous analysis is solely based on the massive data obtained from Foursquare. In this section, we explore cross-site linking between different OSNs. We study Foursquare users who have

Table 6. Cross-Site Information Aggregation: Information Fields

OSN site	Facebook	Twitter	Foursquare
First Name	Alice	N/A	Alice
Last Name	Smith	N/A	-
Full Name	N/A	Alice Smith	N/A
Gender	f	N/A	m
Birthday	Oct. 2, 1994	Oct. 2, 1994	N/A
Location	Shanghai	-	Shanghai
Biography	-	Programmer	Programmer
School	Fudan Univ.	N/A	N/A
Work Unit	-	N/A	N/A
Hometown	Shanghai	N/A	N/A
Marital Status	Single	N/A	N/A

linked their Facebook and Twitter accounts to their Foursquare profiles. We first formulate the problem of cross-site information aggregation in Section 4.1. Afterward, we examine whether a Foursquare user has entered the same content for a specified information field on different OSN sites by referring to Foursquare and Facebook in Section 4.2. Finally, we demonstrate the usefulness of cross-site information aggregation using an example of gender-based analysis of Twitter in Section 4.3.

4.1 Problem Overview

Different OSN sites are designed with different purposes. Therefore, each site might choose its own set of information fields to record the demographic information of users. When looking at two OSN sites, although there might be some information fields in common, each of them might still have its own unique set of information fields. Also, even for the information fields available in both sites, a user might fill them at a different degree of completeness on each site. With cross-site linking, we can aggregate information from multiple OSN sites for the same user. Intuitively, rather than focusing on a single OSN site, cross-site information aggregation allows knowing more about a user. Therefore, cross-site information aggregation is useful for different parties, including the OSN service providers and OSN application providers. They can leverage the cross-site linking function to understand their users better and improve user experience from various aspects.

The enabled information fields of a user u on an OSN site s can be expressed as a set $\mathcal{F}_s^u$. This set covers a number of information fields. Note that some information fields are optional, and we only put the enabled information fields into $\mathcal{F}_s^u$. For example, in Table 6, Alice's profile on Foursquare can be represented as

$$\mathcal{F}_{Foursquare}^{Alice} = \{Firstname, Gender, Location, Bio\}. \quad (1)$$

The other seven information fields in Table 6 are not involved in $\mathcal{F}_{Foursquare}^{Alice}$ since the "Full Name," "Birthday," "School," "Work Unit," "Hometown," and "Marital Status" fields are not supported by Foursquare. Meanwhile, the user Alice chooses not to enable the optional "Last Name" field.

Since different information fields represent different aspects of a user, we can aggregate these fields across Foursquare, Facebook, and Twitter. Considering the enabled information fields of user u, the aggregate information fields for u can be denoted as $\mathcal{F}_{total}^u$. For example, Alice's aggregate

Table 7. Average Number of Enabled Information Fields of Foursquare

All	Gender		Country			
-	Male	Female	US	TR	ID	BR
3.81	3.92	3.90	3.85	4.04	4.00	3.97

Table 8. Average Number of Enabled Information Fields across Foursquare, Facebook, and Twitter

Foursquare Only	Facebook Only	Twitter Only	Foursquare+Facebook	Foursquare+Twitter	Foursquare+Facebook+Twitter
4.01	5.92	2.42	6.58	5.63	8.05

information fields are shown as follows:

$$\begin{aligned}\mathcal{F}_{total}^{Alice} &= \mathcal{F}_{Facebook}^{Alice} \cup \mathcal{F}_{Twitter}^{Alice} \cup \mathcal{F}_{Foursquare}^{Alice} \\ &= \{\textit{Firstname}, \textit{Last Name}, \textit{Nickname}, \textit{Gender}, \\ &\quad \textit{Birthday}, \textit{Location}, \textit{Bio}, \textit{School}, \textit{Work unit}, \\ &\quad \textit{Hometown}, \textit{Marital status}\}.\end{aligned}$$

We define an indictor variable to denote whether the user has entered any value for a certain information field in $\mathcal{F}_{total}$. For each information field α, the indicator $f(\alpha, u)$ tells whether this information field α is enabled by user u:

$$f(\alpha, u) = \begin{cases} 0 & \text{if } \alpha \notin \mathcal{F}_{total}^{u}, \\ 1 & \text{if } \alpha \in \mathcal{F}_{total}^{u}. \end{cases} \tag{2}$$

Therefore, for user u, the number of aggregate information fields A_u can be defined as

$$\mathcal{A}_u = \sum_{\alpha} f(\alpha, u). \tag{3}$$

Before showing the potential of the aggregation of information fields across Foursquare, Facebook, and Twitter, we examine the enabled information fields of Foursquare. As in Table 6, we focus on the following five fields of Foursquare: first name, last name, gender, location, and biography. Table 7 shows that the average number of enabled fields of all Foursquare users is 3.81. We further group the users according to their gender and country. Results show that there exists a slight difference between male and female users. Regarding the home country, users from the United States enable a fewer number of information fields than that from other countries.

Among the crawled Foursquare user profiles, we randomly select 100,000 Foursquare users who have linked their Facebook and Twitter accounts to their profiles. We implement a crawler to retrieve the profiles of these users on Foursquare, Facebook, and Twitter. Table 8 shows the average number of enabled information fields on Foursquare, Facebook, and Twitter and different kinds of combinations of them. We can see that aggregating the information fields across different OSN sites will provide a comprehensive set of demographic information of each user. Also, the coverage of cross-site information aggregation can also be expanded to the UGCs.

4.2 Cross-Site Information Consistency

Different OSN sites share several common personal information fields for user profiles, such as gender, first name, and last name. If a user has accounts on multiple OSN sites, he or she might

Table 9. Cross-Site Information Consistency

Gender	First Name	Last Name	Location (Zip Code)	Location (First-Level Admin. Division)	Location (Country)
99.48%	90.91%	86.94%	68.44%	76.36%	91.16%

choose to expose the same or different information on different sites. The cross-site linking function allows us to accurately evaluate the cross-site information consistency [8]. Using the same set of users as in Section 4.1, we focus on information fields including gender, first name, last name, and location. We first examine the consistency on gender between Foursquare and Facebook. Whether the user has entered the same gender on both Foursquare and Facebook can be determined instantly, as a user is only allowed to choose from "male" and "female" if this user has decided to specify the gender. In addition, we are also interested in whether a Foursquare user has entered the same first name and/or last name on Facebook. Note that 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 the users whose default language on Facebook is "EN-US" or "EN-UK." Last but not least, we look at the location consistency. In both Facebook and Foursquare, a user is allowed to specify his or her "location" information. Given the flexibility in presenting a location, we do not directly compare the corresponding location strings. Instead, we use the Google Geocoding API to translate the original location information in both sites into formatted addresses. By referring to the translated location information, we will be able to obtain three subfields, i.e., the zip code, the first-level administrative division, and the country of the corresponding location. We define the *consistent percentage* as the percentage of users who have entered identical information in a selected field or subfield on both Foursquare and Facebook. The results are shown in Table 9.

Based on the statistical result, 99.48% of the 100,000 users set the same gender information across Foursquare and Facebook. It can be concluded that Foursquare users linking both their Facebook and Twitter accounts are not likely to provide different gender information on two websites. Regarding the first name and the last name, 90.91% and 86.94% of the users provide consistent information. For the location information, 68.44% of the users have entered addresses with the same zip code on Facebook and Foursquare. In addition, 76.36% of the users set the same first-level administrative division and 91.16% of the users specify the same country.

Comparing among these information fields, we can see that different fields have different levels of consistency. The highest level of consistency is achieved in the gender field. It is mainly because of the fewer choices of this field. Differently, the location field has the lowest level of consistency. This is partially due to the flexibility of presenting an address. Also, a user might relocate to somewhere else but fail to keep all his or her profiles on different OSN sites up to date.

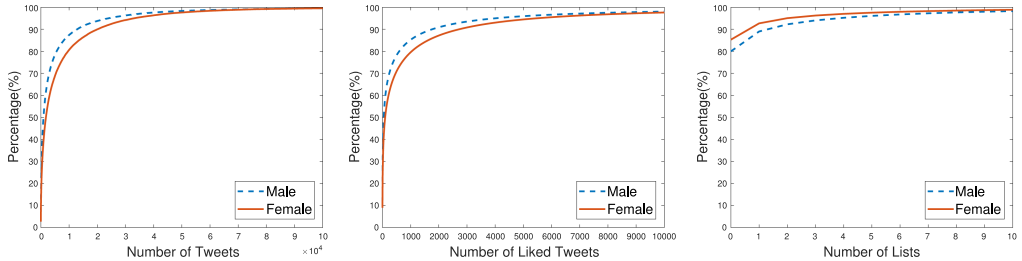
Note that when aggregating the information fields across multiple OSN sites, some information fields might end up with a conflict. For example, a user might enter different contents for a certain information field on different OSNs. According to our study, users have a high probability to enter the same information in the fields of gender and name. For information fields such as location, the level of consistency is not that high unless we focus on the country level. These findings can be helpful for handling possible conflicts, which would be a future issue to explore.

4.3 Cross-Site Information Aggregation Application: Gender-Based Analysis of Twitter

Table 10 shows an example of cross-site information aggregation, using the information fields of Foursquare and Twitter. If a user exposes a link that connects his or her Foursquare and Twitter

Table 10. Cross-Site Information Aggregation: An Example

User	Foursquare			Twitter				
Info	ID	Gender	...	ID	Tweets	Likes	Language	...
User A	1	m	...	13213	1545	21	en	...
User B	2	f	...	9682	100	13	es	...
User C	5	f	...	1293213	5	0	pt	...
User D	9	m	...	8876	23	17	en	...
...	...	...	...	...	...	...	...	...



(a) Number of Published Tweets (b) Number of Liked Tweets (c) Number of Lists

Fig. 6. Gender-based analysis of Twitter.

accounts, we aggregate the information fields of his or her profiles from both sites and build an aggregate table. We conduct a gender-based analysis of Twitter as an example of the cross-site information aggregation. As we mentioned in Section 3.2.1, gender-based study is very important for different social networks. For Sina Weibo, the second-largest microblogging service in the world, Fu et al. [54] and Guan et al. [21] have undertaken gender-based analysis studies, 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. Wang et al. [59] and Xiao et al. [63] proposed to use 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.

Cross-site information aggregation can solve this problem in an automatic yet accurate way, as the gender information is publicly available on Foursquare. We start from the Foursquare users who have added their Twitter accounts to their profiles. Using the aforementioned distributed crawling framework, we have collected 7.89 million Twitter users' profiles. We extract the gender information from each user's Foursquare profile. Among these Twitter users, 54.03% of them are male, 40.88% of them are female, and the rest, 5.09%, of them choose to hide their gender information on Foursquare. 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 liked ("likes_count"),[5] and the number of public lists a user has subscribed to ("listed_count"). These fields indicate a user's activity on Twitter. Figure 6 shows the CDF of these three metrics, by separating male and female users.

According to Figure 6(a), female users publish more tweets than male users. The median numbers of published tweets of male and female users are 974 and 1,819, respectively. In other words, we find that female users are more talkative on Twitter. From Figure 6(b), we can see that female users are also more active in using the "like" function. The median numbers of liked tweets of male

[5]People can "like" a tweet by clicking a small heart icon next to the tweet.

Table 11. Gender-Based Analysis of Twitter: Enablement of Selected Optional Fields (%)

	URL	Description	Location
Male	31.13	64.19	68.53
Female	24.01	64.41	61.78

and female users are 45 and 96, respectively. People can create a "list" by adding some Twitter users into it. The list timeline is composed of a stream of tweets published by the added Twitter users. According to Figure 6(c), male users subscribe to more lists than female users. The 95th percentiles of the number of the lists that male and female users have subscribed to are 4 and 2, respectively.

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 many percent of male and female users have enabled each field. The results are shown in Table 11. We can see a clear difference between male users and female users in terms of adding a URL or location to their profiles. Male users have a higher probability to add such a URL or location. Differently, we find that for the description field, there is little gender difference. Both male and female users have a nearly 64% probability of adding a description.

5 ANALYZING USERS' OPINIONS FOR CROSS-SITE LINKING: A SURVEY

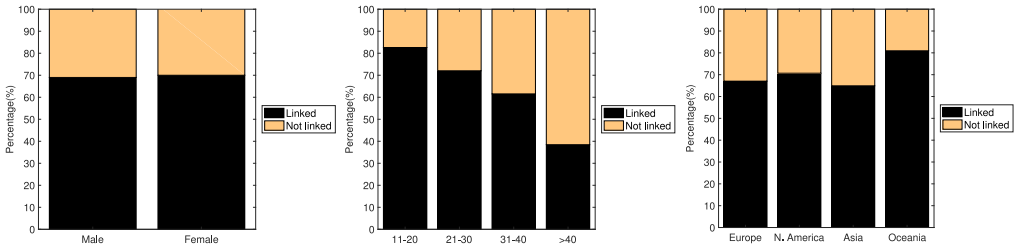
In Section 3 and Section 4, we have conducted a series of measurement studies on cross-site linking. There are still some questions that cannot be answered by referring to measurement results. For example, the results cannot tell why a user chooses to or not to enable this function, which is an important aspect to understand the cross-site linking function. We introduce an online survey using the SurveyMonkey platform[6] to fill this gap. Through the survey, we can directly ask the participants for their detailed opinions about cross-site linking. Therefore, this survey can serve as an important supplement to our measurement study. There are three primary focuses in our survey:

- Why do people want to link their accounts across different OSNs?
- Why do some people choose not to enable the cross-site linking function?
- How do people use the cross-site linking function to enhance the user experience?

The survey is light enough for a participant to finish within 3 minutes, and it covers the above key aspects. Our study went from March 10, 2017, until April 10, 2017, using the SurveyMonkey platform. We advertised the survey through social media such as Facebook, Twitter, and WeChat. In total, we recruited 369 participants from 22 different countries for our survey.

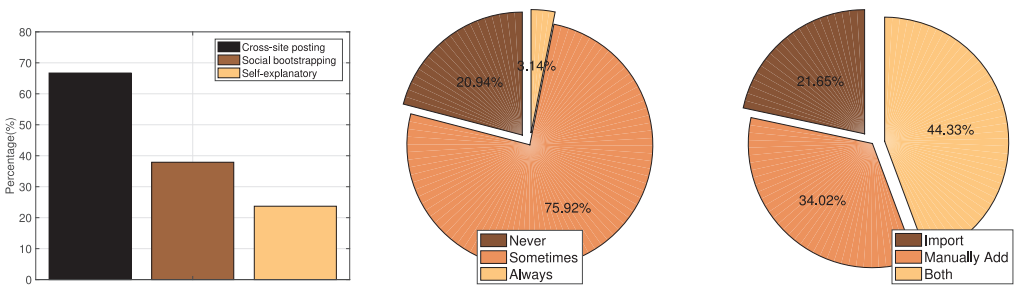
In the first few questions of this survey, we ask the participants to provide their demographic information, i.e., gender, age, and current continent. For the gender information, 56.10% of them are male and 43.90% of them are female. Regarding the age distribution, the percentages of 11 to 20, 21 to 30, 31 to 40, and 41+ are 6.23%, 64.80%, 25.75%, and 3.22%, respectively. In other words, more than 90% of the participants are between 21 and 40. Regarding the continent information, 36.31% of them are from North America, 25.21% are from Europe, 32.79% are from Asia, and the rest, 5.69%, are from Oceania. Therefore, our survey has global coverage. After answering the demographic questions, a participant is asked whether he or she has accounts on different OSNs; 361 (97.83%) of the participants provide a positive answer to this question. Therefore, we can conclude that the

[6] http://www.surveymonkey.com.



(a) Gender (b) Age (c) Continent

Fig. 7. Group-based analysis on survey results (gender/age/continent).



(a) Reasons for Linking (b) Frequency of Cross-site Posting (c) Social Bootstrapping

Fig. 8. Key reasons for enabling cross-site linking.

vast majority of the participants have accounts on multiple OSN sites. We exclude the other eight participants in our further analysis.

Why do people want to link their accounts across different OSNs?

Among the 361 participants who have accounts on multiple OSN sites, we ask each of them whether he or she has used the cross-site linking function; 68.42% of them answer "yes" to this question. For all these participants, we consider different demographic features to classify them into disjoint groups. Figure 7 shows the preferences of the different participant groups on cross-site linking. According to Figure 7(a), we can see that the difference between male and female users is very small, which is similar to our conclusions based on the measurement results in Section 3.2.1. As shown in Figure 7(b), we can see that the cross-site linking function is more popular among younger people. Figure 7(c) shows that participants from Oceania are more likely to link their accounts, covering 80.95% of the participants from this continent. The percentages of users enabling the cross-site linking function from the other three continents are 67.03%, 70.68%, and 64.91%, respectively.

We further investigate the participants who have chosen to enable the cross-site linking function. We show the percentage of participants choosing each of the three primary reasons for cross-site linking in Figure 8(a). Note that a participant is allowed to choose more than one of them when necessary. Among these three reasons, the most popular one is "cross-site content posting," covering 66.66% of these participants, since this feature can save users time from posting the same thing to multiple OSN sites manually. Meanwhile, "importing the friend list" and "self-explanatory" also cover 37.88% and 23.74% of these participants, respectively. Therefore, all the three reasons are viable and useful. Since the top reason for cross-site linking is the convenience of cross-site posting, we also ask the participants how often they use this function. Figure 8(b) shows the distribution of the participants' choices. It is not surprising that almost 79.06% of these participants have used

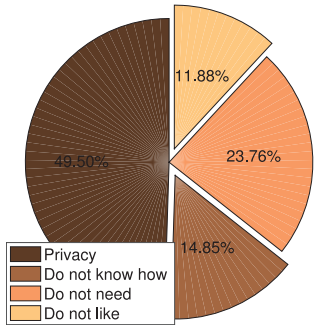


Fig. 9. Key reasons for not choosing cross-site linking.

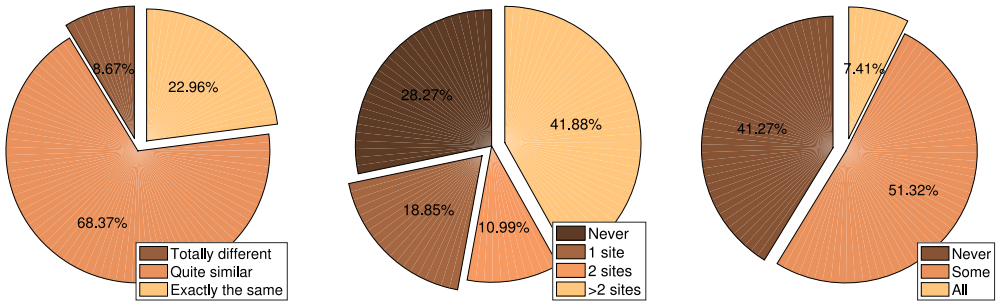
this function. Meanwhile, users are aware that this function will publish the same post to multiple OSN sites with a single click, and accordingly they are cautious when using this function. As a result, only 3.14% of these participants use this function frequently. We are also interested in how a user adds friends when he or she just starts using a new OSN site. After linking his or her accounts on well-established OSN sites, e.g., Facebook and Twitter, to his or her account on this new site, there are two main options for such social bootstrapping. One is to import the friend list from the linked well-established OSN sites directly. The other one is to manually add friends within the new site. According to Figure 8(c), we can see that only 34.01% of these participants have chosen to manually add friends only. In other words, importing friend lists is a widely used function of cross-site linking. This is consistent with the previous findings by Zhong et al. [70].

Why do some people choose not to enable cross-site linking?

Among the participants who have accounts on multiple OSN sites, 31.58% of them have chosen not to enable the cross-site linking function. For these participants, they are asked to provide some reason. As a subjective question, we allow participants to give their own ideas instead of picking from provided options. We then manually classify the answers into four groups. Figure 9 shows the distribution. We can see that 49.50% of these participants feel that enabling cross-site linking will lead to privacy problems. Therefore, this concern prevents them from enabling the cross-site linking function. Interestingly, 14.85% of them disclose that they just do not know how to use this function. We believe it is because the OSN sites do not provide a clear demonstration for the usefulness of cross-site linking. Also, 23.76% of them feel that they do not need this function, and the rest, 11.88%, of them claim that they do not like this function. These findings cannot be obtained by doing the measurements only, and provide us more information on why some users choose not to enable cross-site linking.

How do people leverage the cross-site linking function to enhance the user experience?

We also investigate three aspects of the user experience on cross-site linking based on the participants who have enabled cross-site linking. First, we study if people set the same attribute values across different OSN sites. Figure 10(a) suggests that about 68.37% of these participants prefer to set quite similar values for the same information field on different websites, 22.96% intend to use exactly the same values, and 8.67% would set totally different values. Therefore, the majority of these participants are willing to publish similar contents on different OSN sites. Second, since users can change their personal information on different sites at any time, we explore whether they want to update the information on multiple sites in a consistent way. From our survey results shown in Figure 10(b), 28.27% of these participants do not change any information at all, 18.85% of them just want to change their information on one site, and about 52.87% of these participants prefer to change their information on multiple sites. This indicates that the majority of these



(a) Cross-site Information Similarity (b) Changing Personal Information (c) Deleting Published Posts

Fig. 10. User experience on cross-site linking.

participants are willing to maintain consistent information across multiple OSNs, which is similar to our measurement results. Last but not least, since a user might want to delete a published post, we examine whether a user has considered deleting published contents on relevant sites. According to Figure 10(c), 41.27% of these participants will never delete any post, 51.32% of these participants delete published contents on some (but not all) sites, while only 7.41% of these participants delete published contents on all sites. Therefore, by aggregating the information from multiple sites, some deleted posts can still be recovered, while the users might be not aware of this.

In short, the results of our survey are consistent with our findings from measurements. In addition, this survey provides more insights that cannot be extracted from publicly accessible data on measured OSN sites. Our results support and extend our measurement results to understand the cross-site linking function better.

6 RELATED WORK

There is a series of related work in cross-site linking of online social networks. We classify them into three categories, i.e., cross-site account linkage (Section 6.1), correlation of cross-site user behavior (Section 6.2), and information aggregation from multiple sites (Section 6.3). We go through prospective applications of cross-site linking in Section 6.4.

6.1 Cross-Site Account Linkage

When a user has accounts on different OSN sites, he or she might choose not to disclose the linking between these accounts to the public. Identifying social linkage among different OSN sites has become an important problem. Goga et al. [18] investigated the spatiotemporal information and language style of the user-generated content, to identify accounts that belong to the same user. According to their study, the spatial information is the most powerful feature to find accounts belonging to the same user on different OSN sites. Liu et al. [32] studied the similarity of users' long-term behavior on multiple sites. Along with their analysis on social structure consistency, they proposed a solution framework HYDRA, which was able to correctly identify real user linkage across different OSN sites with high probability. Cao et al. [5] proposed an unsupervised approach to implement the account alignment between two OSN sites. Jain et al. [25] observed the phenomenon that a section of users could evolve their attributes over time. They compared the evolution of the attributes of real-world users and quantified the effect to users' linking options. All these solutions could achieve high accuracy. For OSN sites that do not support the cross-site linking function, or for the scenarios in which users conceal the cross-site links intentionally, these approaches are useful for correlating accounts across OSN sites.

6.2 Correlation of Cross-Site User Behavior

Although cross-site linking provides the convenience of cross-site posting, a user can still behave differently on different social websites. Evaluating the correlation between the same user's activities on different sites has become a viable problem. Kong et al. [27] studied how to find the correspondence between accounts of the same user on Twitter and Foursquare. Lee et al. [30] studied the topical preference of users and proposed a topic model to characterize both the common topics across OSNs and users' specific preferences for each topic on different social media platforms. Ottoni et al. [40] compared user behavior across Pinterest and Twitter and observed significantly different patterns of user activity. Chen et al. [8] studied cross-OSN links between Google+ and nine other OSNs. They found that over 85% of users had more than 50% matched information fields across different OSN sites. However, these studies were based on a very small sample dataset, with 500, 3,000, 30,000, and 35,000 users, respectively.

Zhong et al. [70] studied the social graphs on the source and target linking websites, analyzing the social connections and the evolution of social connections on the target website. Zhong et al. [68] also measured user behaviors on four mainstream social websites, i.e., Facebook, Twitter, LinkedIn, and Instagram, and evaluated how user profiles differed across different OSN sites. Venkatadri et al. [53] investigated the trust consistency of the same user's accounts across various OSN sites and utilized the reputation of user accounts on other websites to detect untrustworthy identities. Goga et al. [18] analyzed the geotags attached to user activities on Twitter, Flickr, and Yelp, respectively. They found that the published posts can provide enough information to correlate accounts across OSNs.

Wang et al. [58] compared a series of user activities across Foursquare, Facebook, and Twitter using 200,000 users sampled randomly. Abel et al. [1] studied the social tagging activities in Social Web systems including Flickr, Twitter, and Delicious. Also, Meo et al. [35] investigated user activities, such as tagging and friending, across different Social Sharing systems, i.e., Flickr, Delicious, and StumbleUpon. Their studies revealed the correlations and differences between the same user's activities on different sites. In our work, we have explored more on the cross-site linking function and addressed a number of important but unexplored issues, such as the distribution of different linking options, behavioral difference among users with different linking options, and user privacy. Also, based on the entire user base, we are able to calculate some metrics that cannot be computed by using sampled subsets, for example, PageRank.

6.3 Information Aggregation from Multiple Sites

Farseev et al. [15] and Chen et al. [8] explored the attribute footprint of a user by aggregating all his or her linked accounts on social websites. Yuan et al. [67] collected user-generated content on different social websites and constructed users' digital footprints. Also, since the cross-site posting function can synchronize a user's activity information on site A to a linked site B, we might refer to site B if the information on site A is not available to the public. For example, the check-in data of Foursquare is invisible to the public by default. A number of researchers, including Hu et al. [23], Silva et al. [46], Preoţiuc-Pietro et al. [42], Scellato et al. [45], and Noulas et al. [38], referred to the linked Twitter accounts to obtain the check-in data by crawling the cross-posted tweets. As we discussed in Section 3.3, the users who have linked their Foursquare accounts to Twitter are more active. Such bias should be carefully taken into consideration in the analysis.

The potential privacy concern incurred by linking services was noticed by Maheswaran et al. [33] and Yang et al. [64]. Maheswaran et al. proposed an architecture, so-called Crypto-Book, to serve as an anonymizing layer between the source websites such as Facebook and the target websites such as Foursquare. With the pseudonym generated by Crypto-Book to log in, users are

protected from being tracked by the target applications. Yang et al. analyzed the information leakage from cross-site posting and proposed a data obfuscating method without distortion to users' sharing intention. Also, Irani et al. [24] studied the cross-site information aggregation problem from the angle of personal information leakage.

6.4 Prospective Applications

The findings of these three parts are quite helpful for social-network-related applications. Since different OSN sites offer different functions, each site might record a user's activity from a certain aspect. By referring to the cross-site links, we will be able to understand a user from a more comprehensive perspective. Meo et al. [35] found that by referring to the same user's information on multiple social systems, there are a number of practical applications such as merging of user profiles, dealing with the cold-start problem for recommendation systems, and computing user similarities. Yuan et al. [67] gathered user-generated content on multiple social websites, utilizing this aggregate information to model people's life pattern. Zhong et al. [69] also made use of cross-site linking in their measurement study about user retention on interest-driven social websites. Venkatadri et al. [53] proposed the idea of interdomain trust transfer. By referring to the cross-site links, we can improve the detection of malicious accounts in newer domains such as Pinterest. Farseev et al. [14] aggregated the social medias and health sensors to understand users' wellness.

7 DISCUSSION

7.1 Privacy Concerns

The enablement of cross-site linking might lead to some privacy risks. Irani et al. [24] have introduced the concept of "unintended personal-information leakage" to represent such risks. Under the context of a single OSN, many users assume that their information will be kept within the social network's boundaries. However, as we have discussed in Section 4, a user's information fields on different sites can be aggregated by making use of cross-site links. As a result, an attacker can integrate individual pieces of information of a user from multiple sites to get a bigger picture of him or her. In addition, according to the survey results we discussed in Section 5, by putting the information from different sites together, an attacker can even recover deleted posts without the awareness of users. Under such situations, some personal information might leak unintentionally. In short, the use of cross-site linking might introduce some new risks on privacy, and many OSN users are unaware of this. Besides demonstrating the usefulness of cross-site linking, our study also shows the privacy problems caused by enabling this function. The explained privacy concerns could help users avoid some unintended privacy leaking when linking accounts across OSNs.

7.2 Practical Implications

Putting the same user's activity data from multiple OSNs together can get us a more comprehensive picture of this user. This "social footprint" can serve as a base for a number of real applications. For example, recommendation systems can take advantage of the rich information of each user collected from different sites to make the recommendation in a more accurate way. Abel et al. [1] studied the correlation of the tag-based user profiles across Flickr, Twitter, and Delicious. They proposed a user modeling strategy that achieved better performance in tag and resource recommendation on these three websites. Other applications based on cross-site information aggregation include malicious account detection [53], life pattern analysis [67], and gender-based user behavior analysis discussed in Section 4.3.

7.3 Take-Home Messages

We summarize our key take-home messages from the following three aspects. First, the cross-site linking function is quite popular. It has been adopted by nearly 60% of Foursquare users and has changed people's online activities a lot. Second, we have demonstrated the usefulness of cross-site linking from different perspectives such as cross-site content posting, social bootstrapping, and self-explanatory. Meanwhile, we also show possible privacy concerns caused by this function. Last but not least, we list several important applications based on cross-site linking. These applications can be leveraged to improve the user experience while using OSN services.

8 CONCLUSION AND FUTURE WORK

In this article, we conduct a comprehensive study on the emerging cross-site linking function on OSN sites. We first use Foursquare as an example site to conduct a measurement-based study. Using the dataset of the entire Foursquare user base, our study quantifies the popularity of this function and shows the behavioral difference of different groups of users in terms of the enablement of this function. Moreover, we formulate the cross-site information aggregation problem and explore the potential of relevant applications. Finally, we undertake a survey to collect hundreds of users' detailed opinions about this function. The major findings of this article are as follows:

- Nearly 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 creation of social connections. Also, enabling an optional information field in a Foursquare user's profile indicates a higher probability of activating the cross-site linking function.
- According to our measurement results, if a Foursquare user has linked his or her account to Facebook, he or she will have a high chance to provide consistent information to both Foursquare and Facebook, especially in the gender and name fields. The use of cross-site information aggregation helps us investigate the gender difference in Twitter, which cannot be studied directly by relying on the Twitter data only. According to our analysis, female users publish and "like" more tweets than male users, while male users subscribe to more public Twitter lists.
- We use a survey to get users' detailed opinions about the cross-site linking function and understand why they choose to or not to enable this function. We summarize the motivation and concerns of the enablement of this function.

There are a number of interesting future directions to better understand and make use of the cross-site linking function. We list them as our future work:

- To make further use of the cross-site linking function, we aim to explore the possibility of making new applications based on this function. In particular, we aim to build an integrated social footprint of a user by referring to multiple OSN sites. By aggregating the publicly accessible user profiles and activities from multiple OSNs, we can know more about a user. Such studies need at least one OSN site to support the cross-site linking function, such as Foursquare or Pinterest. We can use some well-known social hubs, such as about.me [68], to link a user's profiles and activities from up to tens of OSN sites.
- Without a user's permission, we are only able to access the publicly accessible part of his or her social network data. For example, we can view a Foursquare user's public profile, but his or her check-in history is only accessible by friends. As in [31], we plan to conduct a volunteer-based study to access and analyze the complete user activity data of the volunteers. This will result in a more comprehensive investigation into the cross-site linking function.

- As discussed by Venkatadri et al. [53], the cross-site linking function has a great potential in malicious account detection, which is a critical challenge to most of the OSN service providers. By leveraging cross-site links, we can transfer the trustworthiness of a user across OSN sites. We aim to aggregate the information gained by cross-site links into existing malicious account detection solutions, such as social-graph-based approaches [4] or machine-learning-based approaches [20, 66].
- Traditional single-layer networks are widely used to model the social graph of a selected OSN site. However, since a user might participate in different OSN sites, such single-layer networks cannot be used to represent a user's connections across multiple sites. There have been studies generalizing the network theory to cover a few connected networks called multilayer networks. Domenico et al. [10] proposed a tensorial framework as a tool to cope with multilayer network problems. Wang et al. [57] showed the difference about the disease spread within one species and in two species simultaneously. Nguyen et al. [37] modeled the "least cost influence problem" across multiple OSNs. We aim to use multiple-layer networks [2, 10] to characterize the intrasite and cross-site links among users from different OSNs. We will explore the constructed multiple-layer networks to discover influential users from the perspective of multiple OSN sites, study the topological correlations between different layers of networks, and investigate the information diffusion across different OSN sites.

9 DATA AVAILABILITY

We have made the anonymized dataset publicly available via https://github.com/chenyang03/CrossOSN. This dataset covers 61.39 million Foursquare users, including the public information fields shown on their Foursquare profile pages, the numbers of their check-ins and reviews, and the social graph. To keep the privacy of the users, the user profiles have been anonymized.

ACKNOWLEDGMENTS

We are grateful to Xiaoming Fu, Nishanth Sastry, Tianyi Wang, and Gang Wang for their comments and suggestions.

REFERENCES

[1] Fabian Abel, Samur Araújo, Qi Gao, and Geert-Jan Houben. 2011. Analyzing cross-system user modeling on the social web. In *Proceedings of the International Conference on Web Engineering (ICWE'11)*.
[2] Jacopo A. Baggio, Shauna B. BurnSilver, Alex Arenas, James S. Magdanz, Gary P. Kofinas, and Manlio De Domenico. 2016. Multiplex social ecological network analysis reveals how social changes affect community robustness more than resource depletion. *Proceedings of the National Academy of Sciences* 113, 48 (2016), 13708–13713.
[3] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.
[4] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. 2012. Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI'12)*.
[5] Xuezhi Cao and Yong Yu. 2016. BASS: A bootstrapping approach for aligning heterogenous social networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.
[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*.
[7] Terence Chen, Mohamed Ali Kaafar, and Roksana Boreli. 2013. The where and when of finding new friends: Analysis of a location-based social discovery networks. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'13)*.
[8] Terence Chen, Mohamed Ali Kaafar, Arik Friedman, and Roksana Boreli. 2012. Is more always merrier? A deep dive into online social footprints. In *Proceedings of the ACM Workshop on Online Social Networks (WOSN'12)*.
[9] Yang Chen, Jiyao Hu, Hao Zhao, Yu Xiao, and Pan Hui. 2018. Measurement and analysis of the swarm social network with tens of millions of nodes. *IEEE Access* 6 (2018), 4547–4559.

[10] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas. 2013. Mathematical formulation of multilayer networks. *Physical Review X* 3, 4 (Dec. 2013), 041022. DOI: http://dx.doi.org/10.1103/PhysRevX.3.041022

[11] Cong Ding, Yang Chen, and Xiaoming Fu. 2013. Crowd crawling: Towards collaborative data collection for large-scale online social networks. In *Proceedings of the ACM Conference on Online Social Networks (COSN'13)*.

[12] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.

[13] Reza Farahbakhsh, Ángel Cuevas, and Noël Crespi. 2016. Characterization of cross-posting activity for professional users across Facebook, Twitter and Google+. *Social Network Analysis and Mining* 6, 1 (2016), 33:1–33:14.

[14] Aleksandr Farseev and Tat-Seng Chua. 2017. Tweetfit: Fusing multiple social media and sensor data for wellness profile learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'17)*.

[15] Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting multiple sources for user profile learning: A big data study. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR'15)*.

[16] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874.

[17] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. 2014. Studying social networks at scale: Macroscopic anatomy of the Twitter social graph. In *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'14)*.

[18] Oana Goga, Gerald Friedland, Howard Lei, Robin Sommer, Sree Hari Krishnan, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the World Wide Web Conference (WWW'13)*.

[19] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. 2012. Evolution of social-attribute networks: Measurements, modeling, and implications using Google+. In *Proceedings of the ACM Internet Measurement Conference (IMC'12)*.

[20] Qingyuan Gong, Yang Chen, Xinlei He, Zhou Zhuang, Tianyi Wang, Hong Huang, Xin Wang, and Xiaoming Fu. 2018. DeepScan: Exploiting deep learning for malicious account detection in location-based social networks. *IEEE Communications Magazine* (2018). (In press).

[21] Wanqiu Guan, Haoyu Gao, Mingmin Yang, Yuan Li, Haixin Ma, Weining Qian, Zhigang Cao, and Xiaoguang Yang. 2014. Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events. *Physica A: Statistical Mechanics and Its Applications* 395, 0 (2014), 340–351.

[22] Jinyoung Han, Daejin Choi, Byung-Gon Chun, Ted Kwon, Hyun-chul Kim, and Yanghee Choi. 2014. Collecting, organizing, and sharing pins in Pinterest: Interest-driven or social-driven? In *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'14)*.

[23] Tianran Hu, Eric Bigelow, Jiebo Luo, and Henry Kautz. 2017. Tales of two cities: Using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. *IEEE Transactions on Big Data* 3, 1 (2017), 55–66.

[24] Danesh Irani, Steve Webb, Kang Li, and Calton Pu. 2011. Modeling unintended personal-information leakage from multiple online social networks. *IEEE Internet Computing* 15, 3 (2011), 13–19.

[25] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. 2016. Other times, other values: Leveraging attribute history to link user profiles across online social networks. *Social Network Analysis and Mining* 6, 1 (2016), 85.

[26] Long Jin, Yang Chen, Tianyi Wang, Pan Hui, and Athanasios V. Vasilakos. 2013. Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine* 51, 9 (2013), 144–150.

[27] Xiangnan Kong, Jiawei Zhang, and Philip S. Yu. 2013. Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'13)*.

[28] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. 2012. Geographic dissection of the Twitter network. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'12)*.

[29] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the World Wide Web Conference (WWW'10)*.

[30] Roy Ka-Wei Lee, Tuan-Anh Hoang, and Ee-Peng Lim. 2017. On analyzing user topic-specific platform preferences across multiple social media sites. In *Proceedings of the World Wide Web Conference (WWW'17)*.

[31] Shihan Lin, Rong Xie, Qinge Xie, Hao Zhao, and Yang Chen. 2017. Understanding user activity patterns of the swarm app: A data-driven study. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp/ISWC'17)*.

[32] Siyuan Liu, Shuhui Wang, Feida Zhu, Jinbo Zhang, and Ramayya Krishnan. 2014. HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'14)*.

[33] John Maheswaran, Daniel Jackowitz, Ennan Zhai, David Isaac Wolinsky, and Bryan Ford. 2016. Building privacy-preserving cryptographic credentials from federated online identities. In *Proceedings of the ACM Conference on Data and Application Security and Privacy (CODASPY'16)*.

[34] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (1947), 153–157.
[35] Pasquale De Meo, Emilio Ferrara, Fabian Abel, Lora Aroyo, and Geert-Jan Houben. 2013. Analyzing user behavior across social sharing environments. *ACM Transactions on Intelligent Systems and Technology* 5, 1 (2013), 14:1–14:31.
[36] Tehila Minkus, Kelvin Liu, and Keith W. Ross. 2015. Children seen but not heard: When parents compromise children's online privacy. In *Proceedings of the World Wide Web Conference (WWW'15)*.
[37] Dung T. Nguyen, Huiyuan Zhang, Soham Das, My T. Thai, and Thang N. Dinh. 2013. Least cost influence in multiplex social networks: Model representation and analysis. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'13)*.
[38] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'11)*.
[39] Neil O'Hare and Vanessa Murdock. 2012. Gender-based models of location from flickr. In *Proceedings of the ACM Workshop on Geotagging and Its Applications in Multimedia (GeoMM'12)*.
[40] Raphael Ottoni, Diego de Las Casas, João Paulo Pesce, Wagner Meira Jr., Christo Wilson, Alan Mislove, and Virgílio Almeida. 2014. Of pins and tweets: Investigating how users behave across image- and text-based social networks. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'14)*.
[41] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The Pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*.
[42] Daniel Preoţiuc-Pietro and Trevor Cohn. 2013. Mining user behaviours: A study of check-in patterns in location based social networks. In *Proceedings of the ACM Web Science Conference (WebSci'13)*.
[43] Daniele Quercia, Mansoureh Bodaghi, and Jon Crowcroft. 2012. Loosing "friends" on facebook. In *Proceedings of the ACM Web Science Conference (WebSci'12)*.
[44] J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco.
[45] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. 2011. Socio-spatial properties of online location-based social networks. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'11)*.
[46] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. 2014. Revealing the city that we cannot see. *ACM Transactions on Internet Technology* 14, 4 (Dec. 2014), 26:1–26:23.
[47] Xiaodan Song, Yun Chi, Koji Hino, and Belle Tseng. 2007. Identifying opinion leaders in the blogosphere. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'07)*.
[48] Jie Tang, Tiancheng Lou, and Jon Kleinberg. 2012. Inferring social ties across heterogenous networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM'12)*.
[49] Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. 2017. Gender bias in the job market: A longitudinal analysis. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, Article 99 (Dec. 2017), 19 pages.
[50] Mike Thelwall. 2008. Social networks, gender and friending: An analysis of MySpace member profiles. *Journal of the American Society for Information Science and Technology* 59, 8 (2008), 1321–1330.
[51] Asimina Vasalou, Adam N. Joinson, and Delphine Courvoisier. 2010. Cultural differences, experience with social networks and the nature of "true commitment" in Facebook. *International Journal of Human-Computer Studies* 68, 10 (2010), 719–728.
[52] Marisa Affonso Vasconcelos, Saulo Ricci, Jussara Almeida, Fabrício Benevenuto, and Virgílio Almeida. 2012. Tips, dones and to-dos: Uncovering user profiles in foursquare. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM'12)*.
[53] Giridhari Venkatadri, Oana Goga, Changtao Zhong, Bimal Viswanath, Krishna P. Gummadi, and Nishanth Sastry. 2016. Strengthening weak identities through inter-domain trust transfer. In *Proceedings of the World Wide Web Conference (WWW'16)*.
[54] King wa Fu and Michael Chau. 2013. Reality check for the Chinese microblog space: A random sampling approach. *PLoS ONE* 8, 3 (2013), e58356.
[55] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao. 2013. Wisdom in the social crowd: An analysis of quora. In *Proceedings of the World Wide Web Conference (WWW'13)*.
[56] Gang Wang, Sarita Y. Schoenebeck, Haitao Zheng, and Ben Y. Zhao. 2016. "Will check-in for badges": Understanding bias and misbehavior on location-based social networks. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'16)*.
[57] Huijuan Wang, Qian Li, G. D'Agostino, Shlomo Havlin, H. Stanley, and Piet Van Mieghem. 2013. Effect of the interconnected network structure on the epidemic threshold. *Physical Review E* 88, 2 (2013), 022801.

[58] Pinghui Wang, Wenbo He, and Junzhou Zhao. 2014. A tale of three social networks: User activity comparisons across Facebook, Twitter, and Foursquare. *IEEE Internet Computing* 18, 2 (2014), 10–15.

[59] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in english on Twitter. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'14)*.

[60] Yi-Chia Wang, Moira Burke, and Robert E. Kraut. 2013. Gender, topic, and audience response: An analysis of user-generated content on Facebook. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*.

[61] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: Finding topic-sensitive influential Twitterers. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM'10)*.

[62] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. 2009. User interactions in social networks and their implications. In *Proceedings of the ACM European Conference on Computer Systems (EuroSys'09)*.

[63] Chunjing Xiao, Ling Su, Juan Bi, Yuxia Xue, and Aleksandar Kuzmanovic. 2012. Selective behavior in online social networks. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'12)*.

[64] Dingqi Yang, Daqing Zhang, Bingqing Qu, and Philippe Cudré-Mauroux. 2016. PrivCheck: Privacy-preserving check-in data publishing for personalized location based services. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16)*.

[65] Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the International Conference on Machine Learning (ICML'97)*.

[66] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, and Yafei Dai. 2014. Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data* 8, 1 (2014), 2:1–2:29.

[67] Nicholas Jing Yuan, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie. 2013. We know how you live: Exploring the spectrum of urban lifestyles. In *Proceedings of the ACM Conference on Online Social Networks (COSN'13)*.

[68] Changtao Zhong, Hau-wen Chang, Dmytro Karamshuk, Dongwon Lee, and Nishanth Sastry. 2017. Wearing many (social) hats: How different are your different social network personae? In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'17)*.

[69] Changtao Zhong, Nicolas Kourtellis, and Nishanth Sastry. 2016. Pinning alone? A study of the role of social ties on Pinterest. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'16)*.

[70] Changtao Zhong, Mostafa Salehi, Sunil Shah, Marius Cobzarenco, Nishanth Sastry, and Meeyoung Cha. 2014. Social bootstrapping: How Pinterest and last.fm social communities benefit by borrowing links from Facebook. In *Proceedings of the World Wide Web Conference (WWW'14)*.

Received July 2017; revised March 2018; accepted March 2018