World Wide Web
https://doi.org/10.1007/s11280-018-0633-7

Exploring the power of social hub services

Qingyuan Gong[1,2] · Yang Chen[1,2] · Xiaolong Yu[1,2] · Chao Xu[1,2] · Zhichun Guo[1,2] · Yu Xiao[3] · Fehmi Ben Abdesslem[4] · Xin Wang[1,2] · Pan Hui[5,6]

Received: 18 December 2017 / Revised: 19 May 2018 / Accepted: 20 August 2018 /
Published online: 25 September 2018

Abstract
Given the diverse focuses of emerging online social networks (OSNs), it is common that a user has signed up on multiple OSNs. Social hub services, a.k.a., social directory services, help each user manage and exhibit her OSN accounts on one webpage. In this work, we conduct a data-driven study by crawling over one million user profiles from about.me, a representative online social hub service. Our study aims at gaining insights on cross-OSN social influence from the crawled data. We first analyze the composition of the social hub users. For each user, we collect her social accounts from her social hub webpage, and aggregate the content generated by these accounts on different OSNs to gain a comprehensive view of this user. According to our analysis, there is a high probability that a user would provide consistent information on different OSNs. We then explore the correlation between user activities on different OSNs, based on which we propose a cross-OSN social influence prediction model. With the model, we can accurately predict a user's social influence on emerging OSNs, such as Instagram, Foursquare, and Flickr, based on her data published on well-established OSNs like Twitter.

Keywords Online social networks · Social hub services · Measurement · Social influence · Machine learning

1 Introduction

Nowadays online social networks (OSNs) are developing rapidly around the world [19]. Besides supporting social interactions among users, each OSN has its own focus. For example, Twitter acts as a well-established social media for information dissemination [11], Pinterest allows users to discover and share online life styles [14], and Foursquare supports

This article belongs to the Topical Collection: *Special Issue on Social Computing and Big Data Applications*
Guest Editors: Xiaoming Fu, Hong Huang, Gareth Tyson, Lu Zheng, and Gang Wang

✉ Yang Chen
chenyang@fudan.edu.cn

Extended author information available on the last page of the article.

location-centric activities [33]. Because of such diversity in services provided by OSNs, it is very common that a user has accounts on multiple sites [13]. Social hub services have been introduced to help a user manage and exhibit her accounts on different OSNs.

A social hub service allows a user to use a webpage to publish her basic information such as name, location, interest and the URLs of her accounts on different OSN sites. Visitors to her webpage on a social hub are able to access her profile pages on different OSNs conveniently. The popularity of social hubs opens a door to investigate a user's social behaviors across OSNs in an informative way. Aggregating the contents generated by the same user on different OSNs can help us better understand this user, and further model her interests and preferences. Given the rich information provided by social hub services, there have been several attempts to apply the aggregation to build users' social footprints. For example, Yuan et al. [42] proposed a data-driven solution called LifeSpec to study urban lifestyles based on users' social footprints across different sites. They applied a nonparametric Bayesian way to model the lifestyle spectrum of a group of users, and the lifestyle of a single user. Zhong et al. [43] made use of more than 116 thousand user profiles on a representative social hub service, and their linked profiles on Facebook, LinkedIn, Twitter and Instagram. They studied how user profiles vary across different OSNs. Farseev et al. [9] studied the user profile by referring to the user information on multiple OSN sites. They used the collected data from Twitter (textual data), Foursquare (location data) and Instagram (visual data) to demonstrate the effectiveness of their machine learning-based approach for the prediction of users' demographic attributes.

Most of the existing works about users' behaviors across OSNs have focused on the aggregation of user activities from multiple sites. When it comes to social hub services, there are still a number of important but unexplored issues. For example, what are the properties of social hub users? What motivate them to use social hub services? How do the social hub users leverage such services to advertise their profiles on different OSNs? Can we utilize social hubs to explore the correlations between users' behaviors across sites? In this paper, we introduce a data-driven study to answer these questions.

About.me is a representative online social hub service, allowing a user to add her OSN accounts to her profile page [43]. It has attracted millions of users around the world. In this work, we conduct a data-driven study to present the first comprehensive understanding of the about.me service. First, we conduct a series of analysis on 1.06 million profiles of about.me users. We study not only the demographic information of crawled users, but also the behavioral differences between OSN users who have used the social hub services and those who have not. In addition, we explore how to gain a comprehensive view of a user from a cross-OSN perspective by referring to the about.me service, and disclose the correlations of users' behaviors across OSNs. Last but not least, we formulate the cross-site prediction problem based on the correlation analysis. We propose to predict a user's social influence on emerging OSNs by referring to her data published on Twitter, a well established OSN. We reach the following key findings.

- Social hub services attract young adults, interested in technology the most, and with a wide range of occupations. They tend to list their accounts on leading OSNs, such as Twitter, Facebook and LinkedIn, on their profiles of the social hub services. This provides us an opportunity to aggregate data of the same user from multiple OSNs.
- Social hub services enable users to manage and exhibit their profiles on different OSNs, helping users advertise themselves for diverse purposes. The key purpose for social hub users is propagating personal content curation, such as personal websites, articles and music collections.

- There is a high probability that a user would provide consistent information across different OSNs. In addition, information aggregation has a great potential for better characterizing users.
- Users' behaviors on various OSNs show correlations. In our case study, users' behaviors on Twitter are highly correlated with their activities on Foursquare, Instagram and Flickr, respectively. This makes it possible to investigate a user of one OSN using her data on other OSNs.
- We formulate the problem of cross-site prediction of social influence based on social hub services. We propose to use a user's information on a well established OSN, such as Twitter, to predict the same user's social influence on an emerging site, including Foursquare, Instagram and Flickr. Our data-driven analysis shows that we can achieve a high prediction performance.

In Section 2, we present the background, data collection, and related work. In Section 3, we conduct a data-driven study of about.me user profiles from different aspects, including user composition, the motivation of using the service and how to create social footprints of users. In Section 4, we introduce the concept of social influence, and examine the correlations between a user's activities on different sites. In Section 5, we propose the idea of cross-OSN social influence prediction based on machine learning technologies. We implement and evaluate our prediction model in Section 6. We conclude the whole paper in Section 7.

2 Background, data collection and related work

About.me is a representative website providing social hub services, launched in 2009. In our work, we construct a dataset that contains the profiles of 1.06 million users on about.me. In this section, we first give an overview of the about.me website in Section 2.1, and then explain the data collection process in Section 2.2. We review the related work in Section 2.3.

2.1 Overview on the about.me website

About.me is an online website where anyone can create a webpage to manage and exhibit the URLs of her profiles on different OSNs. She can also add her name, location, occupation and interest tags to this webpage. The motivation for an individual to use the about.me service could be looking for jobs, propagating her online collections or manual products, or attracting people to visit her homepage. About.me provides the linking options to nearly all well known OSNs, covering from leading social websites such as Facebook and Twitter, to the job hunting service like LinkedIn, the city life guide applications like Foursquare, the photo sharing platforms like Instagram, and to the image management websites like Flickr. By linking her accounts from different OSNs, a user provides opportunities for other people to know better about her work, life styles, social connections and so on.

Figure 1 is the public profile page[1] of an about.me user. This user states her position as a student, shows her location in Shanghai, China, and adds several interest tags to let visitors know more about her preferences. This user creates this about.me page with a motivation of job hunting. The button in the middle of the page showing "Hire me" will guide anyone

[1] https://about.me/gongq/, December 2017.

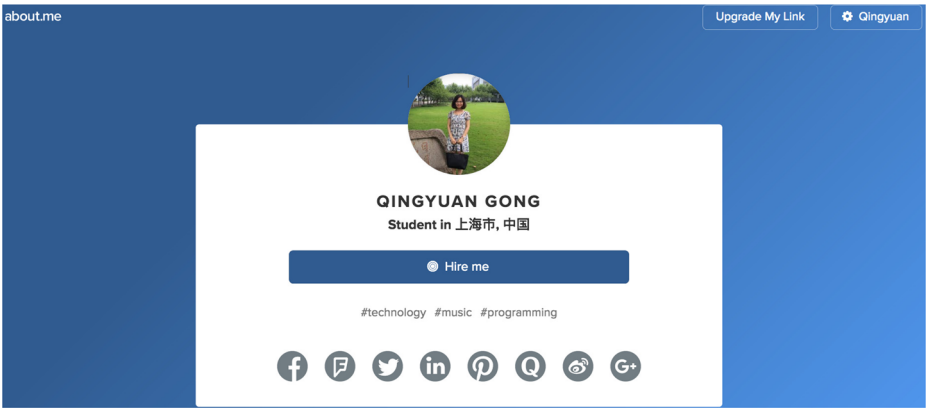


Figure 1 An about.me user's public profile page

interested in hiring this person to send her an email. If this page is created for advertising the user's homepage, this button can offer a link to her homepage. Similarly, this button can link to charitable donation, curation sharing or events advertisement. This user adds the URLs of her social profiles on Facebook, Foursquare, Twitter, LinkedIn, Pinterest, Quora, Sina Weibo and Google+. Visitors to this about.me page can be redirected to her profile pages on the listed OSNs.

2.2 Data collection

To construct a dataset of about.me user profiles, we obtained a list of 1.06 million about.me user IDs by using the "discover" function[2] recursively. Based on this list, we implemented a Python-based crawler to fetch the profile pages of all these users. We did the crawling during March 2017, using 5 virtual instances on the Vultr cloud. For each profile page, we extract the demographic information including the name, interest tags, occupation and location. In particular, we get the list of the corresponding user's linked social accounts, for example, her accounts on Facebook, Twitter, LinkedIn.

With these publicly-available URLs directing to their homepages on other OSNs, we are able to access the content they have generated on other OSNs. Among the top 12 linked OSNs, we choose Twitter, Instagram, Foursquare and Flickr to conduct our case study. Twitter is one of the leading OSNs, while the other three are emerging OSNs of three typical services. Table 1 shows the number of users who linked each of these websites to their about.me accounts. We developed crawlers to download the public contents on these four websites, obtaining the profiles and publicly-visible activities of these users. The crawlers were deployed on 30 virtual instances on the Vultr cloud. The data collection on these linked websites lasted for one week from April 22 to April 29, 2017.

[2] https://about.me/discover, December 2017.

Table 1 Number of user profile pages crawled from the four selected OSNs

Linked OSNs	Twitter	Instagram	Foursquare	Flickr
Number of users	472,924	242,579	47,202	57,165

2.3 Related work

2.3.1 Cross-site linking in OSNs and social hub services

Given the emerging social applications, many users hold accounts on multiple OSNs since different services are provided separately. For example, users are used to looking for jobs on LinkedIn, checking in at venues on Foursquare, meeting their friends on Facebook, and receiving news on Twitter. To ease the management, emerging OSNs have started to provide the cross-site linking function [13]. This function connects the accounts of a user on well-established OSN, such as Facebook and Twitter, to her profiles of comparatively new emerging social websites such as Foursquare. URLs of the linked social profile pages on Facebook or Twitter are publicly-accessible on her Foursquare homepage [26]. Cross-site linking is a way to aggregate the content users generate across multiple OSNs [9]. Starting from the profile page of a Foursquare user, we are able to access the public content on Facebook and Twitter, if she has linked her accounts on the two well established OSNs to Foursquare. Gong et al. [44] have utilized the cross-site linking function to connect users' behavior on Medium and Twitter, exploiting the rich activity data on Twitter to predict the potential influential users on Medium.

Social hub services are quite different from the cross-site linking function. It does not rely on any specific OSN, but an independent webpage aggregating users' available social accounts. In addition, since authentication from the linked OSN site is required for cross-site linking, the number of OSNs linked is usually limited. On a social hub, users can add any of her social accounts to her profile page. By accessing the profile pages on social hubs, we can aggregate the social footprints of the users as far as they are willing to exhibit. Due to the difference between these two kinds of connections of social accounts, the users attracted and the potential applications are of different characteristics. Our work studies the user groups of social hub services, and explains the possible applications that rely on the user data collected from a number of connected OSNs.

2.3.2 Social influence in OSNs

The concept of social influence has been discussed in [1, 3, 22, 23, 29, 31, 41]. Existing studies have proposed different metrics to quantify the social influence. Cha et al. [3] proposed three metrics to evaluate the social influence, i.e., indegree influence, retweet influence and mention influence. These metrics characterize the number of followers a user has, the number of retweets related to a user, and the number of mentions containing a user's name, respectively. Similarly, Ye et al. [41] used the following three metrics to measure social influence, i.e., number of followers, number of replies and number of retweets. Kwak et al. [21] investigated two metrics, i.e., number of retweets and PageRank value. PageRank is an algorithm for the Google search engine to rank the importance of Web pages [27]. This metric considers not only the number of audience one user has, but also the influence these audience hold. PageRank contains the propagation of influence along the entire social graph of one OSN site, and also has been discussed in [23, 29, 31]. However, the calculation of

PageRank value requires the information of the entire social graph, which prevents it from being an influence metric in practical applications. In addition, PageRank does not consider the interests of the users. To remedy this, an extension of PageRank, known as TwitterRank [37] has been proposed to consider both the social structure and the topical similarity among Twitter users.

2.3.3 Machine learning-based user classification

Machine learning technologies have been widely used for classifying OSN users. One primary scenario is malicious account detection. Yang et al. [40] developed a support vector machine (SVM) classifier to uncover sybil accounts in Renren, an OSN in China. Wang et al. [36] proposed a practical sybil detection system based on a clickstream model, which can group similar user clickstreams into clusters. Gong et al. [12] proposed DeepScan, a deep learning-based approach to detect malicious accounts in location-based social networks. DeepScan applies a long short-term memory (LSTM) neural network to learn users' dynamic activities.

Also, machine learning plays an important role in different kinds of online user behavior analysis. Ma et al. [24] applied LSTM neural network to detect rumors on microblogging platforms by making use of time series event information. Similarly, Suhara et al. [30] proposed an LSTM-based predictive model to forecast people's depressed mood. They used a self-developed smartphone application to collect the historical data of users, and their evaluation demonstrated a good prediction performance.

3 Understanding the about.me service

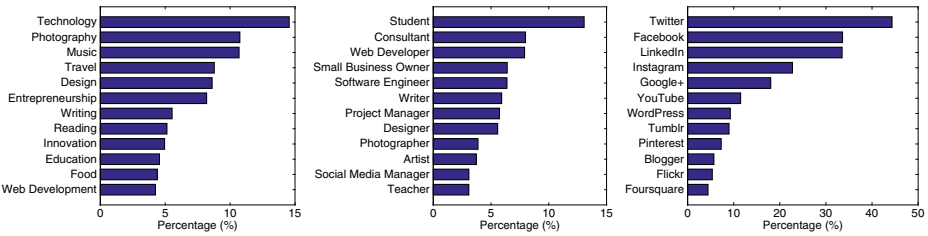
As a social hub service, about.me has a number of unique features. In this section, we conduct a series of measurements to demonstrate different characteristics of about.me users. We study the composition of the user set in Section 3.1. We further investigate the motivation of about.me users in Section 3.2. In Section 3.3, we explore the cross-OSN information aggregation enabled by the social hub services.

3.1 Composition of the about.me users

About.me provides a set of pre-defined information fields for users to introduce themselves. One about.me user can add static attributes, such as occupation, location, interest tags and a profile photo. The key function of social hub services is allowing one user to list her accounts on different OSNs. In this subsection, we study the characteristics of the about.me user set from three aspects: 1) the distribution of the users concerning the typical information fields on about.me, 2) the distribution of users concerning the demographic features, 3) the behavioral difference on Twitter between the groups of Twitter users that use about.me service and the randomly selected Twitter users.

3.1.1 Information fields analysis

Based on the crawled profiles, we study the composition of about.me users concerning their key information fields shown on their about.me homepages. First, we analyze the "interest tags" of each user. Note that a user is allowed to add multiple interest tags. The proportions



(a) Top 12 ranking of interests (b) Top 12 ranking of occupations (c) Top 12 ranking of linked OSNs

Figure 2 Top 12 interest tags, occupations and linked OSNs on about.me

of the top 12 interest tags added are shown in Figure 2a. In particular, "Technology", "Photograph" and "Music" are the three most popular interest tags. In addition, we examine the distribution of the users' occupations in Figure 2b. The top three occupations are "Student", "Consultant" and "Web Developer". Finally, we explore the corresponding distribution of linked OSN accounts, and list the top 12 linked OSN sites of about.me users in Figure 2c. The top three sites are Twitter, Facebook and LinkedIn, respectively. These three sites are also the three most popular OSN sites according to the Alexa ranking.[3]

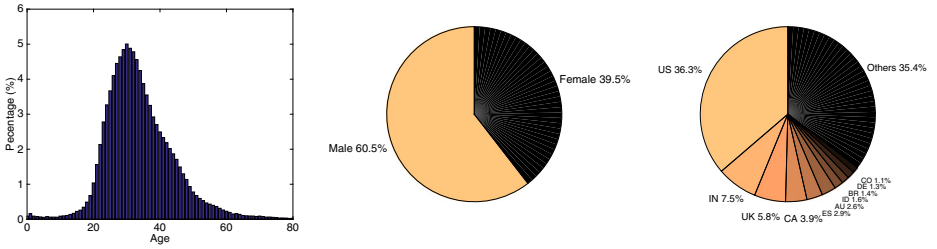
3.1.2 Demographic analysis

Each about.me user can upload an image as their profile photo. Although about.me shows limited information fields of users explicitly, more demographic characteristics such as age and gender can be deduced from users' portraits. It should be noticed that some profile photos users upload are celebrities such as movie stars, instead of their own faces. Dealt with Microsoft Azure Face API, we found that the profile photos of 309,793 users in our dataset contain a single human face. First, we have conducted an additional study to check whether most of the about.me users have used photos of celebrities (e.g. movie/music stars) as their profile photos. We selected 309,793 about.me users whose profile photos contain a single human face. We searched each of these users' profile photos online using the "search by image" function provided by Google. The search results include a number of webpages containing the same photo. If a profile photo turns out to be a photo of a celebrity, the number of relevant webpages will be large. According to our study, only 3.14% of the 309,793 users have been using photos of celebrities as their profile photos.

We further validate the correctness of the Microsoft Azure Face API. Among the about.me users who use human faces as their profile photos except for celebrities, there are 10,382 users who have disclosed their Foursquare accounts and published the gender information on Foursquare. The analysis results of 10,108 (97.36%) users are consistent with the gender information published by themselves on Foursquare based on Microsoft Azure Face API. This result verifies the recognition accuracy of the Microsoft Azure Face API.

We continued to use the API to analyze the age of the about.me users from their profile photos. Figure 3a shows the age distribution of about.me users, which follows a Gaussian normal distribution. 74.75% of about.me users are aged between 20 and 40 years old. The mean age is around 33. This number is larger than that of Facebook [20]. Users who have accounts on about.me are in general more senior than ordinary OSN users. We also plot the

[3] https://www.alexa.com/topsites/category/Computers/Internet/On_the_Web/Online_Communities/Social_Networking



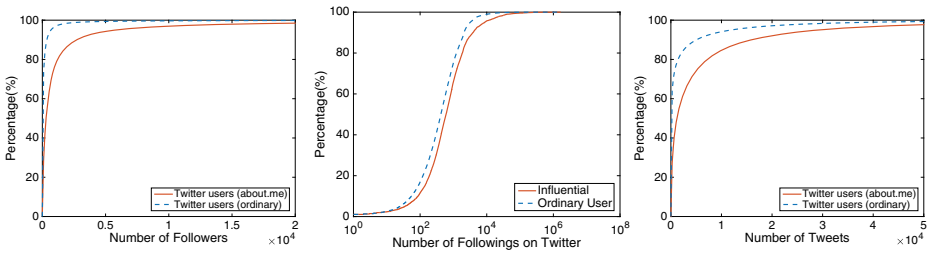
(a) Age distribution (b) Gender distribution (c) Country distribution

Figure 3 The distribution of users demographic information

distribution of male and female users in Figure 3b. We find that there are more male users on about.me than female users. In addition, Figure 3c shows the distribution of the countries where users come from. It can be seen that the USA, India, UK and Canada occupy a dominant part, reflecting the popularity of the social hub services in these countries.

3.1.3 Analysis of behaviors on Twitter

As a social hub service, about.me requires users to provide their basic personal information besides attaching social accounts on other OSNs. We observe that 44.34% of about.me users have added their Twitter accounts. To gain some insights on the activity characteristics of about.me users, we compare the behavior of Twitter users who have started to use the social hub services on about.me with randomly selected Twitter users. We call the latter "ordinary Twitter users" in this paper. Not all the self-discloded links to the Twitter accounts are valid, because some do not exist and some have been banned by Twitter. Filtering out the invalid Twitter accounts, there are 398,579 Twitter users who have about.me accounts. We compare these users with 398,579 randomly selected ordinary Twitter users. Since the primary types of activities on Twitter are related to social connections establishment and tweets spreading, we plot the cumulative distribution function (CDF) curves of the two groups of Twitter users concerning the number of followers, followings, and published tweets. Results are shown in Figure 4. We find that the Twitter users using about.me servics are better connected, and have published more tweets than the ordinary Twitter users.



(a) CDF of Followers (b) CDF of Followings (c) CDF of Tweets

Figure 4 The distribution of users demographic information on about.me

3.2 The motivations of using about.me services

About.me provides a "spotlight" button in the middle of the homepage to help users disclose their purposes of using such a social hub service. Intuitively, we believe that most of the OSN users sign up for social hub services with a purpose of advertising themselves. On about.me, a user is allowed to pick from a number of options to highlight a motivation of using such service, such as "visit my website", "view my portfolio", "hire me", "read my blog" and "visit my company website". Accordingly, the spotlight button could be enabled for opening a pre-defined URL and be used for sending a contact message to the user's registered email address automatically.

In our about.me dataset, there are 513,191 (48.12%) users who have activated this spotlight button. We summarize the primary purposes of enabling this button and list the top 10 in Figure 5. Most of these highly ranked purposes are attracting visitors to view their personal sites, such as personal pages, portfolios, blogs, advertising companies, and online curations of photos, videos or musics. The highest ranked purpose is "visit my website", which reflects users' inclination to attract visitors to their personal websites. Visitor can be redirected to the personal website of an about.me user by a single click at the spotlight button. Other highly ranked purposes are also in accordance with the distributions of the interests and occupations shown in Figure 2a and b, respectively.

3.3 Aggregation of user data across OSNs

About.me provides us an opportunity to access each user's linked social accounts on multiple OSN sites. Therefore, we are able to further crawl a user's social accounts on different sites to get a comprehensive view of this user. The profile fields required by each OSN often vary with the services it provides, as listed in Table 2. We can see the usefulness of the cross-OSN information aggregation since a user might disclose different pieces of information on different OSNs.

We focus on two aspects, i.e., *cross-OSN information consistency* and *cross-OSN information aggregation*. For consistency, we aim to know whether a user prefers to disclose consistent information on different OSNs or not. For aggregation, we are interested in the

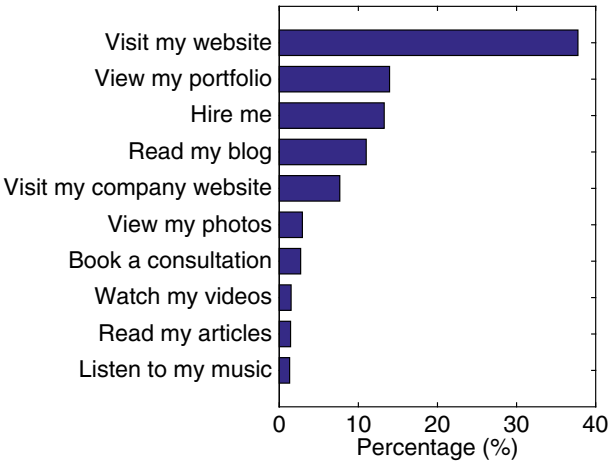


Figure 5 Distribution of motivations of about.me users

Table 2 Profile fields on leading OSN sites

OSN site	about.me	Facebook	Google+	Foursquare	Blogger	Instagram	Flickr
Name	√	√	√	√	√	√	√
Profile Image	√	√	√	√	√		
Gender	–	√	√	√	√	–	–
Work/Education	√	√	√	-	√	–	√
Location	√	√	√	√	√	√	√
Interests	√	–	–	–	√	–	–
Relationship	–	√	–	–	–	–	–
Birthday	–	√	–	–	–	√	–
Bio	√	–	√	√	√	√	√
Contact	–	–	√	–	–	–	√
Social Links	√	√	√	√	√	√	√
Sexual Orientation	–	√	–	–	–	–	–
Social Graph	–	√	–	√	–	√	√
User-Generated Content (UGC)	–	√	–	√	√	√	√
Total	7	11	8	8	9	7	9

usefulness of aggregating a user's information on different OSN sites together. We randomly select a subset of 30,000 about.me accounts and further crawl their public profiles on Facebook, Foursquare, Google+ and Blogger.

3.3.1 Cross-OSN information consistency

To evaluate the cross-OSN information consistency, we use the gender and location fields, which are two common information fields of the profiles on all these four OSNs. For any two sites, we define the "consistency" metric as the fraction of users that set the same values for a certain information field. In particular, for the location field, we evaluate the consistency at the country level. For each pair of the four selected OSNs, we use a two-tuple to represent the consistency in terms of gender and country information respectively. Since Twitter is one of the most popular OSNs, we also compare users' public information between each of the selected OSNs and Twitter. We focus on the consistency of location information, since Twitter does not allow users to show their gender. As shown in Table 3, there exist high consistency of gender and location information between pairs of OSNs.

In short, we can see that users tend to provide consistent information across different OSNs.

Table 3 Consistency of each OSN pair (Gender, Location)

Websites	Facebook	Foursquare	Google+	Blogger	Twitter
Facebook	\	(0.99, 0.92)	(0.99, 0.94)	(0.98, 0.89)	(– , 0.92)
Foursquare	(0.99, 0.92)	\	(0.99, 0.93)	(0.99, 0.91)	(– , 0.89)
Google+	(0.99, 0.94)	(0.99, 0.93)	\	(0.98, 0.89)	(– , 0.91)
Blogger	(0.98, 0.89)	(0.99, 0.91)	(0.98,0.89)	\	(– , 0.82)
Twitter	(–, 0.92)	(– , 0.89)	(– , 0.91)	(– , 0.82)	\

3.3.2 Cross-OSN information aggregation

Aggregating the information generated by the same user on multiple OSNs is crucial for studying users' behaviors across OSNs. As each OSN typically focuses on a specific type of function or service, the user-generated contents on a single OSN can hardly provide a comprehensive view of a user's online behavior. Therefore, it has become important to study the characteristics of users' behavior across multiple OSNs. In fact, users' aggregated information has been applied to uncover malicious accounts, to study user behaviors by referring to different OSNs providing various functions, or to align the accounts on different OSNs that belong to the same person. For example, Venkatadri et al. [35] proposed to improve the detection of malicious accounts in newer domains such as Pinterest by referring to the users' reputation on Twitter. They called it inter-domain trust transfer. Yuan et al. [42] gathered user generated content on multiple social websites, and utilized the aggregated information to model people's life pattern. Jain et al. [17] observed that a section of users' information fields evolved over time. They compared the evolution of the attributes of the same person on different OSNs.

Each OSN site has its own set of available information fields. Some of them are mandatory and some are optional. Intuitively, putting the information fields of different OSN sites together will provide a more detailed view of a user's online footprints [5]. Note that on different OSN sites, the same information field might have different names, such as "work" on Facebook and "occupation" on Blogger. We manually examine all information fields of

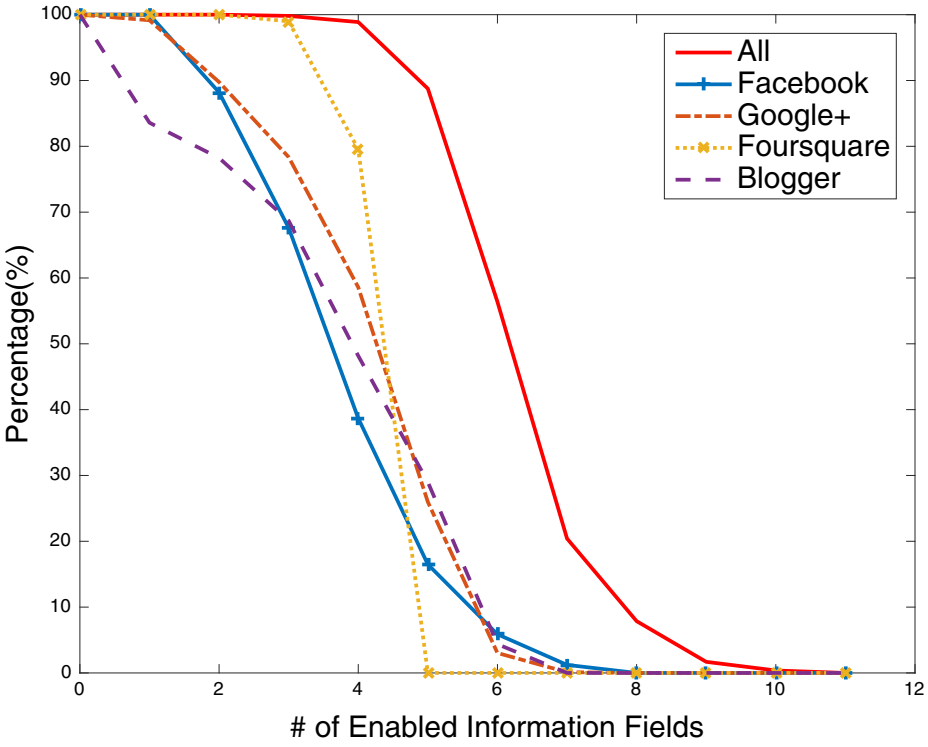


Figure 6 Cross-OSN information aggregation

the four selected OSN sites and merge these fields. Finally, we get 11 different information fields across all four sites. In Figure 6, we see the complementary cumulative distribution function (CCDF) of the number of activated information fields on each of the four OSN sites. In addition, an aggregated set of a user's enabled information fields on different OSN sites are shown as "All". We can see that through the cross-OSN information aggregation by referring to profiles on about.me, the number of activated information fields of each user increases significantly. In other words, putting a user's disclosed information on different OSN sites together, we can know more about this user. This demonstrates the great potential of the social hub services.

4 Relating cross-site user behaviors via social hub services: social influence analysis

Although it is common that one person maintains accounts on multiple OSNs, correlations between the same user's activities across OSNs have not been fully understood yet. In this section, we take the social influence as an example to study the correlations between users activities across OSNs. We first summarize the general metrics of social influence in Section 4.1. Then we disclose the correlations of user behaviors between users' social influence on emerging OSNs and their behaviors on established OSNs in Section 4.2.

4.1 Definition of social influence

4.1.1 Formulization of social influence

Social influence [1, 3, 41] is a metric used to quantify the potential power one user owns to affect others, reflecting the importance of one user on a selected OSN. Knowing more about the users' social influence is beneficial to understand users' power of propagating online issues or delivering specific opinions. Also, users' social influence is crucial for studying the spreading topics online, including hot news, rumors and advertisements.

The measurement of the social influence one user u owns can be expressed as a step function. We denote the value of one user's social influence as I_u, which could be calculated by different definitions. If a user's social influence I_u is beyond a pre-defined threshold t, she can be considered as an *influential*. Otherwise, she will be defined as an *ordinary user*. The formulization of influentials and ordinary users on an OSN site is shown as follows.

$$I_u \begin{cases} > t & \text{u is an influential.} \\ \leq t & \text{u is an ordinary user.} \end{cases} \tag{1}$$

There are several definitions of social influence. As in [1, 3, 41], social influence of a user on Twitter is defined as the number of followers a user has, the number of tweets that have been retweeted, and the number of tweets mentioning the user's name, respectively. There are also a number of social graph-based definitions, such as the PageRank value of a user in a social graph. However, as discussed in [38], a social connection does not indicate an essential social interaction between users. Instead, the reactions from other users are more reflective of the power one user owns for information propagation. In this work, we focus on the "like" function, which can better characterize the reactions of users to the published posts.

The value of the threshold t can be determined according to the specific requirements to the objective influentials and applications. For example, for a product promotion with a limited advertisement budget, the advertiser might wish to approach a small set of users who have higher influence within the social networks. These users help diffuse information over the entire network. To obtain t, we can specify a percentage $p\%$ and denote all users with social influence values ranked among the top $p\%$ as influentials and the rest of users as ordinary users.

4.1.2 Calculation of social influence

The "like" function is a widely-used function on most mainstream OSNs, allowing users to show their interests or favor on published contents. This function is supported by a number of popular social networks, such as Facebook [7], Twitter [13] and Instagram [18]. A "like" operation is often implemented by lighting a small heart icon, showing the appreciation of the user for the published content, such as the "like" icon under every tweet in Twitter. It is obvious that a user with a higher social influence will receive more "likes" from other users.

We obtain the value of the social influence by referring to the $\mathcal{H}$-index based definition [15]. Instead of using the total number of "likes", which might be biased to a very popular post, we aim to uncover users who can publish a set of popular posts, indicating a constant influence over time. The $\mathcal{H}$-index metric has been widely used to evaluate the scientific impact of a scholar. The operator $\mathcal{H}$ acts on a set of real numbers $(x_1, x_2, ..., x_n)$. It returns the integer y, indicating that y elements within the range of $(x_1, x_2, ..., x_n)$ are larger or equal to y, while the rest $n - y$ elements are smaller than y.

Using $\mathcal{H}$-index, we define the value of the social influence I_u of one user u as follows.

$$I_u = \mathcal{H}(k_{u1}, k_{u2}, ..., k_{un}) \tag{2}$$

In (2), n denotes the total number of posts this user has published, while $k_{um} (1 \leq m \leq n)$ represents the number of "likes" received by her m-th post. We take the number of "likes" received by each published post $(k_{u1}, k_{u2}, ..., k_{un})$ as an "likes vector" of the user u.

The defined social influence is compatible with different OSNs. For each site, a "post" can be implemented in the primary form of the user-generated contents on this site. In our study, we select three emerging OSNs of different types for case studies, including the city life app Foursquare [34], the photo sharing app Instagram [16] and the photo/video management service Flickr [4]. By crawling the demographic information and generated content of users on these websites, we construct the likes vectors for users on each OSNs, concerning the behaviors that are the most representative of influence.

- Foursquare: As a city life application, Foursquare allows users to show their opinions to a point of interest (POI) by posting tips, which are public to other users. Anyone can signify their "like" or "unlike" when viewing any tips. The "likes" obtained by the user's tips are extracted to construct the likes vector that are used to estimate this user's influence.
- Instagram: The typical activity of Instagram users is to publish photos tagged with their opinions. Similarly, these posts are publicly available. The "likes" obtained by a user's posted photos form the likes vector of her.
- Flickr: The goal of Flickr is to help people share online albums conveniently. Users of Flickr can build albums to manage their photos. They can also view other users' albums and show "likes" to the ones that they think good. The numbers of "likes" received by the photos in a user's album construct her likes vector.

Based on the likes vector of each user, we are able to calculate the values of her social influence.

4.2 Correlations between emerging and established OSNs

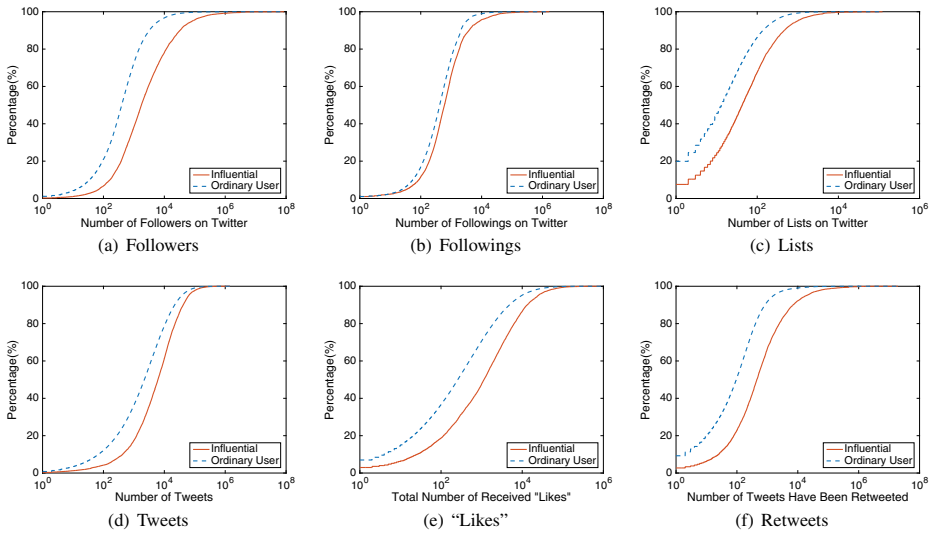
Nowadays, a number of social applications are emerging fast. These emerging OSNs often focus on a specific type of function, supporting social services as a secondary function for better user experience and retention.

A large number of users are more active on well established OSNs than on emerging OSNs. We are trying to exploit the rich behavior trajectories of users on established OSNs to predict their behavior characteristics on emerging OSNs. In this section, we conduct a case study of users' social influence and explore the correlations of users' behaviors between emerging and established OSNs.

As a representative OSN, Twitter ranks the first of the linked OSN sites on about.me, with 44.34% of about.me users exhibiting their Twitter accounts on their about.me homepages. It means that many users own accounts on Twitter and tend to show it on the social hub. To evaluate the correlations between the emerging and established OSNs, we pick the users who show both their accounts on Twitter and an emerging OSN, for example, Instagram, Foursquare or Flickr. The correlation between users' social influence on each of these OSNs and their behaviors on Twitter is evaluated respectively. Among the three emerging sites, we take Instagram as an example to show the results. About 11.23% of about.me users add both their Twitter and Instagram accounts. We select these 119,762 about.me users, and have crawled the contents they published both on Instagram and on Twitter. Based on the values of the $\mathcal{H}$-index of the "likes" the users have received for their posts, we obtain the social influence rankings of these users on Instagram. Taking users of the top 5% as the influentials concerning social influence values, we classify these users into two groups, namely, influentials and ordinary users on Instagram.

By crawling the profile page of a user on Twitter, we obtain all her demographic information, her published tweets, and her numbers of followers and followings. One Twitter user can post her original post directly, or repost a published tweet by using the "retweet" function. She can follow other users to subscribe for their tweets, and click the "like" icon represented by a small heart, showing the appreciation for the tweet. Considering these functions, we pay special attention to the following three aspects of contents generated by Instagram users on Twitter, i.e., social connections including the numbers of followings and followers, curation habits of subscriptions to tweets through lists, and the properties of published tweets. We compare the behaviors on Twitter between influentials and ordinary users of Instagram in Figure 7. Summarizing the distributions of the influentials and ordinary users on Instagram, concerning each selected metric on Twitter, we can reach the following conclusions.

- Influentials on Instagram often have more followers than followings on Twitter. For both influentials and ordinary users on Instagram, we examine the numbers of followers and followings on Twitter. The distributions are shown in Figure 7a and b, respectively. It is obvious that influentials on Instagram have both more followers and followings on Twitter than ordinary users. A larger gap appears when it comes to the distributions of the number of followers.
- Influentials on Instagram tend to subscribe to more lists on Twitter. Figure 7c compares the number of lists the influential and ordinary Instagram users have subscribed to on Twitter. Subscribing to a "list" is a convenient way for users to focus on the users they



(a) Followers (b) Followings (c) Lists

(d) Tweets (e) "Likes" (f) Retweets

Figure 7 Comparison of influentials and ordinary users on Instagram based on their selected behaviors on Twitter

are interested in. The results reflect that influentials are more professional or devoted to content curation on Twitter.

– Influentials on Instagram publish more tweets, and attract more "likes" or retweets. Figure 7d shows that influentials on Instagram generate more tweets on Twitter than ordinary users. Also, the tweets generated by influentials receive more "likes" or retweets by other Twitter users, demonstrated in Figure 7e and f.

The correlations between Twitter and Foursquare/Flickr show similar characteristics. We also find obvious difference between the influentials and ordinary users on Foursquare and Flickr, respectively, concerning their behaviors on Twitter. The comparisons between the influentials and ordinary users on these three OSNs concerning six typical behavioral metrics on Twitter are shown in Table 4. All the above analysis shows that the social influence of users on emerging OSNs is related to their activities on Twitter, one of the leading social networks. This reveals the new opportunity to model the behavior of social users on emerging OSNs, with the help of the content they generate on Twitter.

Table 4 Behavior characteristics of Instagram, Foursquare and Flickr users on Twitter (Infl.: Influentials, Ord.: Ordinary users)

Median	Followers		Followings		Lists		Tweets		"Likes"		Retweets	
	Infl.	Ord.	Infl.	Ord.	Infl.	Ord.	Infl.	Ord.	Infl.	Ord.	Infl.	Ord.
Instagram	1780	284	591	340	43	9	3197	1160	1128	117	455	56
Foursquare	1845	288	957	342	76	9	3211	1178	1119	120	323	58
Flickr	650	289	523	342	35	9	3199	1179	495	120	178	58

5 Cross-OSN social influence prediction

According to the studies in Section 4, we can see that a user's activities on Twitter are correlated to her social influence on emerging OSNs such as Instagram, Foursquare and Flickr. In this section, we propose the concept of *cross-OSN social influence prediction*. We aim to use user-generated content on Twitter to identify potential influentials on an emerging OSN. In this way, we can quickly identify influentials on an emerging OSN without referring to the users' activity data of this site, and can predict whether one Twitter user will be an influential on the emerging OSN even when she has not generated any behavior trajectories on this OSN. In Section 5.1, we discuss the workflow of the machine learning-based framework. In Section 5.2, we introduce the design of the cross-OSN influence prediction system.

5.1 Machine learning-based framework

Machine learning frameworks have been introduced to study user behaviors in OSNs. Figure 8 illustrates the workflow of the machine learning based framework for social influence prediction. The key components include datasets construction, feature extraction, parameter tuning and performance evaluation.

- Dataset preparation: The first step of using machine learning techniques is to construct the datasets needed. A training and validation dataset should be prepared to train and validate the prediction model, while a test dataset is needed to evaluate the prediction performance of the trained model. User instances in both datasets should be labeled in advance, signaling whether one user instance is an influential or an ordinary user.
- Feature extraction: This step is to select the features of the user instances which are the most relevant to the classification objective. The subsets of features are exacted by considering the behaviors users can conduct on OSNs, often from three aspects, including demographic features, social connections features and content generation features. The extracted features are fed into the classification models in the form of vectors.
- Parameter tuning: This step is conducted over the training and validation dataset, aiming to find an optimal set of parameters that can achieve the best prediction performance when generalizing to the test dataset. Parameters are often tuned by grid search, i.e., going through the parameters space and recording the parameters that could enable the

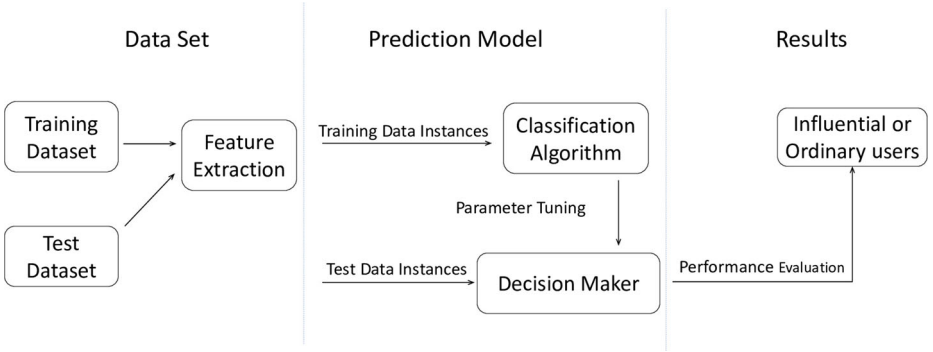


Figure 8 Machine learning based framework for prediction of social influence

prediction model to achieve the highest value concerning one chosen metric. Once the set of parameters is fixed, the classification algorithm can work as a decision maker.

– Performance evaluation: Four classic metrics are often used to examine the prediction performance of a trained decision maker, i.e., precision, recall, F1-score and AUC value [10]. Precision is the percentage of influentials that are predicted as influentials. Recall measures the number of user instances that are correctly judged as influentials or ordinary users. F1-score is the harmonic mean of the two metrics. AUC is the possibility that a classifier will rank a randomly chosen influential higher than a randomly chosen ordinary user. A statistical significance test, for example, McNemar's test [25], is needed when we compare the performance of different classification algorithms.

5.2 Predicting social influence on emerging OSNs

Section 4.2 shows the correlations between users' activities on a well-established OSN such as Twitter and their social influence on emerging OSNs such as Instagram, Foursquare and Flickr. In this section, we explore to predict users' social influence across OSNs, implementing the machine learning framework demonstrated in Figure 8. Considering a user on an emerging OSN such as Instagram, our objective is to detect whether she is an influential or an ordinary user. The system design is shown in Figure 9, using Instagram as an example for explanation. Based on the workflow of the machine learning framework in Section 5.1, we implement and evaluate the social influence prediction system for Instagram users in the following steps.

5.2.1 Datasets construction

To construct the datasets, we select the 119,762 about.me users showing both Instagram and Twitter accounts. For each user in the training and validation dataset and the test dataset, we need to label her as either an influential or an ordinary user. Based on our definition of social influence in Section 4.1, the social influence value of a user can be calculated using the $\mathcal{H}$-index operator on the "likes" vector of each Instagram user. Moreover, our system is compatible with other definitions of social influence. Based on the social influence definitions, we can label user instances as influentials or ordinary users in the datasets. The

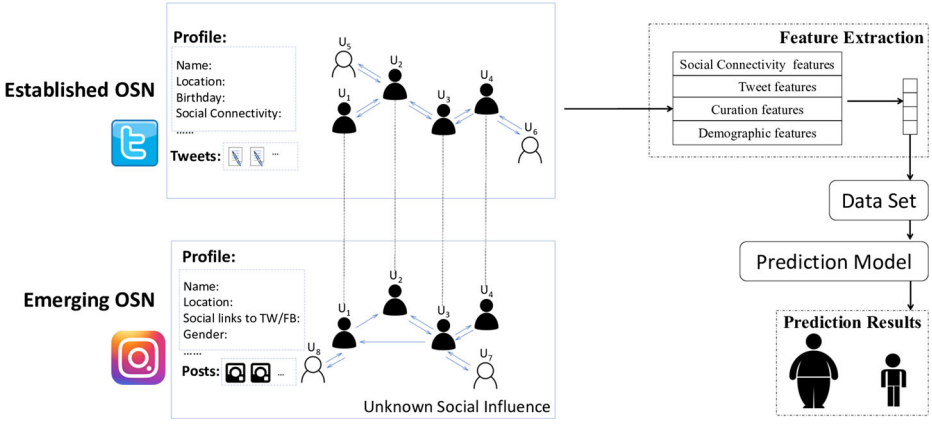


Figure 9 Prediction on users' social influence on emerging OSNs based on selected features from Twitter

threshold can be decided according to the requirement to the influentials in corresponding application scenarios. In this part of experiment, we take the top 10% users as influentials concerning the ranking of the social influence values, and other users as ordinary users.

5.2.2 Feature extraction

Referring to the profile page on the social hub website of an Instagram user, we can access her public profile page on any selected OSN, i.e., Twitter. Obtaining the content generated on Twitter by the same users who also have accounts on other emerging OSNs, including the demographic profiles, published tweets and social connections, we aim to exact informative features that are helpful for distinguishing the influentials on emerging OSNs. Through carefully studying the footprints of users, we extract the following sets of features that could describe users' behaviors on Twitter.

Social connectivity features: This set of features capture the sociality related activeness of the users. On Twitter, the social graph is constructed by the following relations among users. Thus, we take the number of followers, and the number of followings of one user as her social connectivity features.

Tweet publishing features: Twitter users can either publish original tweets, or forward existing tweets by retweeting them. We extract users' tweet publishing features from the following four aspects, including the number of total tweets and retweets, the median interval between original tweets, the median length of the original tweets, the $\mathcal{H}$-index of the retweets, and the total number of "like" she receives.

Linguistic features: We use the Linguistic Inquiry and Word Count (LIWC) tool [32] to study the linguistic characteristics of the published tweets. We take the first 200 tweets of each user to study the users' linguistic features, including the property of the word, the number of the words in each sentence, or the categories of vocabularies.

Content curation feature: Twitter users can create lists. A list covers a set of Twitter accounts, and aggregates the tweets published by accounts within this list only. The curation feature is the number of lists a user has subscribed to.

Demographic features: We consider the demographic information of a user, such as the time zone offset to the Greenwich Mean Time (GMT), whether the location tag for tweets is enabled, whether the user uses the default profile photo, if the URLs of her other social accounts are attached, and the registration time of her account.

5.2.3 Parameter tuning and performance evaluation

Feeding the classification algorithm with the extracted features and the labels of the user instances in the training and validation dataset, parameters of the classifier can be tuned to generate a prediction model.

We employ different supervised machine learning algorithms to undertake the classification. The candidate algorithms include XGBoost [6], C4.5 decision tree (J48) [28], Random Forest [2], and linear support vector machine (LIBLINEAR) [8]. It is noteworthy that XGBoost is a tree-boosting system that has been widely used in machine learning challenges such as Kaggle. Once the feature set and a classification algorithm are selected, we introduce grid search to sweep a grid of parameters and record an optimal set of parameters, which could achieve the highest prediction performance in terms of F1-score. After that, we evaluate the performance of the obtained model on the test dataset. We show the comparison among evaluated algorithms in Table 5. From the table we see that, XGBoost

Table 5 Prediction of Instagram influentials with various classifications

Algorithm	Parameter	Precision	Recall (TPR)	F1-score	FPR	AUC
XGBoost	Refer to Table 6	0.738	0.712	0.725	0.253	0.789
RF	20 trees, 40 features/tree	0.714	0.714	0.714	0.286	0.714
J48	Confidence factor =0.01, Instance/leaf M=3	0.665	0.665	0.665	0.335	0.665
LIBLINEAR	L2-regularized L2-loss support vector classification, Cost = 7	0.507	0.447	0.475	0.430	0.509
Bayes Network	Default	0.654	0.654	0.654	0.346	0.654

outperforms the other classification algorithms. We conduct McNemar's test [25] for each pair of the classification algorithms, and learn that there are significant differences between any two of these classification algorithms (p-value < 0.001, McNemar's test). We then choose XGBoost as the classification algorithm to construct the decision maker in the following experiments.

In the parameter tuning process of XGBoost, the meaning of the parameters used are explained as follows. The parameters "alpha" and "lambda" are L1 and L2 regularization terms on weights, respectively, controlling the regulation part of XGBoost. The parameter "gamma" controls the split of the leaf nodes of the tree. The three parameters could make the algorithm conservative. In addition, "$learning\_rate$", "min_child_weight", "max_depth", "$subsample$" and "$colsample\_bytree$" are used to represent the gradient descent rate (to prevent overfitting), the minimum sum of the weights concerning all observations in a child, the maximum depth of the tree, subsample fraction of the training instance, subsample fraction of columns when constructing each tree, respectively. The parameter "objective" signifies the learning objective of this model. The parameter "seed" acts as the seed used by the random number generator. Table 6 shows the values of the parameters used in the XGBoost algorithm for the social influence prediction in this part.

To measure the correlations between the features used and the judgement the machine learning model makes, we conduct χ^2 (Chi Square) static analysis over all features [39]. The ten most discriminative features are listed in Table 7. From the table we see that the number of followers on Twitter is of the greatest relevance to the user's influence on Instagram, while the age of the account ranks the second. The tweet publishing behavior characteristics are also important features for recognizing influentials on Instagram.

Table 6 Parameters set of XGBoost for social influence prediction on Instagram

Parameter	Value
Learning_rate	0.05
Min_child_weight	2
Max_depth	10
Gamma	0.2
Subsample	0.7
Colsample_bytree	0.8
Alpha	0.001
Objective	multi:softprob
Seed	7
Lambda	1

Table 7 χ^2 statistic

Rank	χ^2	Feature	Feature Category
1	2428.9360	Number of followers	Social Connectivity
2	2380.4440	Registration time of the account	Demographic
3	2103.6384	$\mathcal{H}$-index of "like" received	Tweet publishing
4	1340.9825	Number of "like" received	Tweet publishing
5	1340.98251	Number of tweets retweeted	Tweet publishing
6	1258.9897	Number of lists subscribed	Content curation
7	1227.3252	Number of tweets	Tweet publishing
8	539.83838	Number of followings	Social Connectivity
9	531.65589	Average word count of tweets	Linguistic
10	531.65589	Frequency of the word "work" used in the tweets	Linguistic

6 Evaluation

In this part of evaluation, we implement our model to distinguish the influentials on Instagram, Foursquare and Flickr individually. We evaluate the prediction performance of our system from two aspects: the different criteria to label the influentials and the different definition of the social influence. For each definition of social influence, the corresponding criterion decides the numbers of user instances in the training and validation dataset and the test dataset, respectively. To test the performance when applying different labeling criteria, we construct the datasets used in the experiment by taking the top 1%, 5% and 10% as the influentials respectively, according to the ranking of the social influence values. To test the compatibility to different definitions of social influence, we compute the social influence based on the $\mathcal{H}$-index of the number of "likes", and the $\mathcal{H}$-index of the number of comments received by the posts published by each user.

In the evaluation, we take Instagram, Foursquare and Flickr as the examples of emerging OSNs to conduct case studies. Obtaining accounts of users on multiple OSNs through their about.me profile pages, we implement the cross-OSN social influence prediction models for Instagram, Foursquare and Flickr users, based on the content generated by the same user on Twitter. We evaluate our prediction model on two categories of about.me user groups. The first category includes the about.me users who have accounts on both Twitter and either one of Instagram, Foursquare and Flickr. We call it Model1. The second category contains the about.me users who have accounts on Twitter and all the three emerging OSNs, which is the so-called Model2.

The sizes of user datasets covered in our experiment are listed as follows.

- **Instagram-Twitter:** 119,762
- **Foursquare-Twitter:** 31,372
- **Flickr-Twitter:** 41,231
- **Instagram-Foursqaure-Flickr-Twitter:** 5899

Our experiments are divided into two parts, defining the value of social influence of each user based on the $\mathcal{H}$-index of the number of "likes" and the $\mathcal{H}$-index of the number of comments received by her published posts. In both parts of the experiments, we use different thresholds to label the influentials, i.e., the top 1%, 5% and 10% in the descending ranking

Table 8 The number of user instances in Training/Validation Dataset and Test Dataset (Model1: Datasets of the users that have accounts on both Twitter and either one of Instagram, Foursquare and Flickr; Model2: Dataset of the users that have accounts on all these four OSNs)

Model	Training/Validation dataset	Test dataset
Model1 Instagram (1%)	1916	480
Model1 Foursquare (1%)	502	126
Model1 Flickr (1%)	660	166
Model1 Instagram (5%)	9682	2396
Model1 Foursquare (5%)	2510	628
Model1 Flickr (5%)	3300	824
Model1 Instagram (10%)	19164	4790
Model1 Foursquare (10%)	5020	1256
Model1 Flickr (10%)	6598	1650
Model2 (5%)	472	118
Model2 (10%)	944	236

of the social influence values. For each labeling threshold, the number of user instances in the training and validation dataset is shown in the Table 8, as well as that in the test dataset.

During the training phase of each model, we use 5-fold cross-validation. The training/validation dataset is divided into five subsets with equal size. Among the five subsets, one is retained for validation, and the other four are used for training. The cross-validation process is repeated for 5 times, with each of the five subsets serving as the validation dataset. Each round of validation trainings can be measured by precision, recall, FPR, AUC and F1-score. We determine the parameters of the machine learning model based on cross validation that achieves the highest F1-score. Using the trained model, we test its performance using the user instances in the test dataset. The cross-site prediction framework is compatible with different emerging OSNs, and accordingly can be used to discover influentials on different sites by referring to the generated data on Twitter.

6.1 Social influence definition based on the $\mathcal{H}$-index of the number of "likes"

In this part, the influentials in the training and test datasets are identified based on the top 1%, 5% and 10% of social influence values computed by the $\mathcal{H}$-index of the number of "likes". We first use the **Instagram-Twitter**, **Foursquare-Twitter** and **Flickr-Twitter** datasets to examine the prediction performance for the users that add their accounts on each of the emerging OSN (i.e., Instagram, Foursquare, or Flickr) and Twitter on about.me. Using features extracted from Twitter and the labels of influentials in each dataset, prediction models are trained separately. From Table 9 we can conclude that users' features extracted from Twitter are useful for predicting her social influence on other emerging OSNs. In general, the prediction performance for Instagram is better than the other two OSNs, with the highest F1-score as 0.863 when the influentials are labeled as the top 1% in the Instagram-Twitter dataset. Parameters of XGBoost are tuned independently to obtain different prediction models for each of the three emerging OSNs, with an objective of optimizing the F1-score. The corresponding parameters for XGBoost are shown in Table 10, with the user instances in the training and validation dataset and the corresponding labels fed into the classification algorithm.

Table 9 Prediction of social influence for users that have accounts on Twitter and each of emerging OSNs

Top 1% of "Likes" $\mathcal{H}$-index labeled as influentials					
Cross-OSN sites	Parameter	Precision	Recall	F1-score	AUC
Instagram -Twitter	Model1 Instagram (1%)	0.839	0.766	0.800	0.881
Foursquare -Twitter	Model1 Foursquare (1%)	0.833	0.896	0.863	0.897
Flickr-Twitter	Model1 Flickr (1%)	0.747	0.826	0.785	0.784
Top 5% of "Likes" $\mathcal{H}$-index labeled as influentials					
OSN site	Parameter	Precision	Recall	F1-score	AUC
Instagram -Twitter	Model1 Instagram (5%)	0.776	0.776	0.751	0.834
Foursquare -Twitter	Model1 Foursquare (5%)	0.778	0.766	0.772	0.823
Flickr -Twitter	Model1 Flickr (5%)	0.675	0.712	0.693	0.732
Top 10% of "Likes" $\mathcal{H}$-index labeled as influentials					
OSN site	Parameter	Precision	Recall	F1-score	AUC
Instagram -Twitter	Model1 Instagram (10%)	0.735	0.714	0.724	0.802
Foursquare -Twitter	Model1 Foursquare (10%)	0.709	0.713	0.711	0.746
Flickr -Twitter	Model1 Flickr (10%)	0.670	0.674	0.672	0.721
Random Guess	–	0.5	0.5	0.5	0.5

We further examine the prediction performance for about.me users who have accounts on all the four OSNs. Using the Instagram-Foursqaure-Flickr-Twitter dataset, we predict their social influence on the three emerging OSNs. In this part, prediction models are also trained for the three emerging OSNs, respectively, with the feature sets extracted from Twitter. Considering the limited number of users in the dataset, we only take two thresholds to label influentials of the top 5% and the top 10%. Distinct training and validation dataset is constructed for each emerging OSN based on the label of the social influence on different emerging OSNs respectively. Prediction performance on the test datasets is shown in Table 11. The tuned parameters for XGBoost algorithms used are shown in Table 12. From

Table 10 Parameters set for XGBoost in Table 9

Parameter	Model1 Instagram			Model1 Foursquare			Model1 Flickr		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
Learning_rate	0.05	0.05	0.05	0.1	0.1	0.1	0.15	0.15	0.05
Min_child_weight	2	2	2	2	2	2	1	2	2
Max_depth	10	10	10	10	9	9	10	10	9
Gamma	0.2	0.2	0.2	0	0.2	0	0	0	0.2
Subsample	0.9	0.9	0.9	0.8	0.6	0.7	0.6	0.7	0.8
Colsample_bytree	0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.8	0.9
Alpha	0.001	0.001	0.001	0.001	0	0	0.001	0.001	0
Objective	multi:softprob								
Seed	7								
Lambda	1								

Table 11 Prediction of social influence for users have accounts on Twitter and all three emerging OSNs

Top 5% of "Likes" $\mathcal{H}$-index labeled as influentials					
Cross-OSN sites	Parameter	Precision	Recall	F1-score	AUC
Instagram	Model2 Instagram (5%)	0.742	0.831	0.784	0.799
Foursquare	Model2 Foursquare (5%)	0.754	0.648	0.697	0.702
Flickr	Model2 Flickr (5%)	0.707	0.757	0.731	0.707
Top 10% of "Likes" $\mathcal{H}$-index labeled as influentials					
Cross-OSN sites	Parameter	Precision	Recall	F1-score	AUC
Instagram	Model2 Instagram (10%)	0.756	0.775	0.765	0.775
Foursquare	Model2 Foursquare (10%)	0.745	0.760	0.753	0.790
Flickr	Model2 Flickr (10%)	0.686	0.793	0.736	0.745
Random Guess	–	0.5	0.5	0.5	0.5

the results we see that the behavior data generated on Twitter are useful to predict users' social influence on other emerging OSNs.

We can only obtain precision, recall, F1-score and AUC as 0.5 based on random guess. Our prediction models achieve a better performance by including the same set of features extracted from established OSNs. Twitter features demonstrate great efficiency to predict whether a user is an influential on other emerging OSNs.

6.2 Social influence definition based on the $\mathcal{H}$-index of the number of comments

We evaluate the prediction model against other definitions of social influence. In this subsection, we take the $\mathcal{H}$-index of the number of comments received by the user on emerging OSNs as the metric of social influence, as the comment function can also reflect the social interactions. This part of evaluations are also conducted on the two categories of about.me

Table 12 Parameters set for XGBoost in Table 11

Parameter	Model2 Instagram		Model2 Foursquare		Model2 Flickr	
	5%	10%	5%	10%	5%	10%
Learning_rate	0.1	0.1	0.15	0.15	0.1	0.1
Min_child_weight	1	1	2	1	2	1
Max_depth	9	9	9	10	8	8
Gamma	0.2	0	0	0.1	0.1	0.1
Subsample	0.6	0.8	0.7	0.7	0.8	0.8
Colsample_bytree	0.8	0.8	0.7	0.8	0.9	0.7
Alpha	0	0.001	0	0.0001	0.0001	0.001
Objective	multi:softprob					
Seed	7					
Lambda	1					

Table 13 Prediction of social influence for users having accounts on Twitter and one of the emerging OSNs

Top 1% of comments $\mathcal{H}$-index labeled as influentials					
Cross-OSN sites	Parameter	Precision	Recall	F1-score	AUC
Instagram -Twitter	Model1 Instagram (1%)	0.837	0.813	0.825	0.883
Foursquare -Twitter	Model1 Foursquare (1%)	0.851	0.875	0.863	0.866
Flickr-Twitter	Model1 Flickr (1%)	0.729	0.843	0.782	0.767
Top 5% of comments $\mathcal{H}$-index labeled as influentials					
OSN site	Parameter	Precision	Recall	F1-score	AUC
Instagram -Twitter	Model1 Instagram (5%)	0.772	0.741	0.756	0.827
Foursquare -Twitter	Model1 Foursquare (5%)	0.720	0.791	0.754	0.782
Flickr -Twitter	Model1 Flickr (5%)	0.669	0.666	0.667	0.700
Top 10% of comments $\mathcal{H}$-index labeled as influentials					
OSN site	Parameter	Precision	Recall	F1-score	AUC
Instagram -Twitter	Model1 Instagram (10%)	0.741	0.715	0.728	0.798
Foursquare -Twitter	Model1 Foursquare (10%)	0.710	0.767	0.738	0.777
Flickr -Twitter	Model1 Flickr (10%)	0.696	0.676	0.686	0.741
Random Guess	–	0.5	0.5	0.5	0.5

users: having accounts on both Twitter and either one of Instagram, Foursquare and Flickr, and having accounts on Twitter and all the three emerging OSNs. The experiments are conducted in the same way as described in the previous subsection. When identifying the influentials in the user datasets of the first category, the thresholds are set as the top 1%, 5% and 10%, respectively. Prediction models are constructed in each scenario. Results and the corresponding model parameters are shown in Table 13 and Table 14. The prediction models also perform well when we apply another definition of social influence. Similarly, we conduct the experiments on the **Instagram-Foursqaure-Flickr-Twitter** dataset, with results and the parameters shown in Table 15 and Table 16. The overall performance is a bit

Table 14 Parameters set of XGBoost in Table 13

Parameter	Model1 Instagram			Model1 Foursquare			Model1 Flickr		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
Learning_rate	0.1	0.05	0.05	0.05	0.1	0.15	0.1	0.15	0.1
Min_child_weight	2	1	2	1	2	2	2	2	1
Max_depth	9	9	8	9	8	10	10	9	8
Gamma	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2
Subsample	0.7	0.7	0.8	0.8	0.7	0.8	0.7	0.9	0.8
Colsample_bytree	0.9	0.7	0.7	0.7	0.6	0.8	0.8	0.6	0.9
Alpha	0.001	0.001	0	0.001	0.001	0.001	0	0.001	0.001
Objective	multi:softprob								
Seed	7								
Lambda	1								

Table 15 Prediction of social influence for users having accounts on Twitter, Instagram, Foursquare and Flickr

Top 5% of comments $\mathcal{H}$-index labeled as influentials					
Cross-OSN sites	Parameter	Precision	Recall	F1-score	AUC
Instagram	Model2 Instagram (5%)	0.760	0.760	0.760	0.779
Foursquare	Model2 Foursquare (5%)	0.770	0.839	0.803	0.833
Flickr	Model2 Flickr (5%)	0.647	0.743	0.692	0.658
Top 10% of comments $\mathcal{H}$-index labeled as influentials					
Cross-OSN sites	Parameter	Precision	Recall	F1-score	AUC
Instagram	Model2 Instagram (10%)	0.783	0.761	0.771	0.791
Foursquare	Model2 Foursquare (10%)	0.714	0.776	0.744	0.778
Flickr	Model2 Flickr (10%)	0.729	0.720	0.724	0.759
Random Guess	–	0.5	0.5	0.5	0.5

inferior, compared to Table 13, which we believe accounts for the limited number of user instances in the training and validation dataset.

7 Conclusion and future work

In this work, we undertake a measurement-based analysis on about.me, a representative social hub service. We explore the user groups it attracts, analyze the aggregated information of users from their linked OSNs, and then analyze the correlations between users' behavior on emerging OSNs and that on well established OSNs to explore the possible applications.

We study both cross-OSN consistency and cross-OSN aggregation problems. Our findings reveal the great power of the linkage of a user's accounts on multiple OSN sites. We come to the conclusion that users' behaviors on emerging and established OSNs are correlated, with the social influence analysis of users as a case study. Based on the conclusions, we propose a machine learning-based model to predict users' social influence on emerging

Table 16 Parameters set for XGBoost in Table 15

Parameter	Model2 Instagram		Model2 Foursquare		Model2 Flickr	
	5%	10%	5%	10%	5%	10%
Learning_rate	0.15	0.1	0.1	0.15	0.15	0.15
Min_child_weight	1	1	2	1	2	2
Max_depth	8	9	9	9	10	8
Gamma	0	0	0	0.2	0	0.2
Subsample	0.9	0.7	0.6	0.6	0.8	0.8
Colsample_bytree	0.6	0.7	0.7	0.9	0.8	0.8
Alpha	0	0.001	0.001	0.0001	0.0001	0.001
Objective	multi:softprob					
Seed	7					
Lambda	1					

OSNs using the selected features on Twitter. Our cross-OSN prediction model is efficient in the scenarios where the predictive objective is correlated to the informative clues such as the behavior characteristics on other OSNs. We have also verified the performance of the cross-site predictive model based on the cross-site linking function. The system can predict the potential influential users on Medium using the features extracted from the users' Twitter accounts [44].

For future work, we aim to make good use of the linkage between user accounts on different OSNs listed on the social hub websites. First, by aggregating a user's information across multiple OSNs, we will achieve a more informative view by getting a better "social footprint" of a user. We can harvest the comprehensive demographic information of activities generated by users. Second, the social footprint of users can be used to analyze the characteristics of user behaviors across OSNs. Since each OSN usually has its own focus of services, explorations on the consistency and difference of the user activities across OSNs are useful for better understanding a user. Last but not least, we plan to connect a user's activities from both online and offline perspectives. We can leverage the behavioral data from different types of OSNs, for example, connecting the user's offline mobility by referring to a user's location information on location-based social networks (LBSNs) such as Foursquare, and the content they generate online by using online services such as tweeting on Twitter.

Acknowledgments This work is sponsored by National Natural Science Foundation of China (No. 61602122, No. 71731004), Natural Science Foundation of Shanghai (No. 16ZR1402200), Shanghai Pujiang Program (No. 16PJ1400700).

References

1. Bakshy, E., Mason, W.A., Hofman, J.M., Watts, D.J.: Everyone is an influencer: quantifying influence on Twitter. In: Proc. of ACM WSDM (2011)
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in Twitter: The million follower fallacy. In: Proc. of AAAI ICWSM (2010)
4. Cha, M., Benevenuto, F., Ahn, Y., Gummadi, P.K.: Delayed information cascades in Flickr: Measurement, analysis, and modeling. Comput. Netw. **56**(3), 1066–1076 (2012)
5. Chen, T., Kaafar, M.A., et al.: Is more always merrier? A deep dive into online social footprints. In: Proc. of ACM WOSN (2012)
6. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. of ACM KDD (2016)
7. Cristofaro, E.D., Friedman, A., Jourjon, G., Kaafar, M.A., Shafiq, M.Z.: Paying for likes?: Understanding Facebook like fraud using honeypots. In: Proceedings of the 2014 Internet Measurement Conference, IMC 2014, Vancouver, BC, Canada, November 5-7, 2014, pp. 129–136 (2014)
8. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
9. Farseev, A., Nie, L., Akbari, M., Chua, T.: Harvesting multiple sources for user profile learning: A big data study. In: Proc. of ACM ICMR (2015)
10. Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
11. Gabielkov, M., Rao, A., Legout, A.: Studying social networks at scale: Macroscopic anatomy of the Twitter social graph. In: Proc. of ACM SIGMETRICS (2014)
12. Gong, Q., Chen, Y., He, X., Zhuang, Z., Wang, T., Huang, H., Wang, X., Fu, X.: DeepScan: Exploiting deep learning for malicious account detection in location-based social networks. IEEE Communications Magazine (2018)

13. Gong, Q., Chen, Y., Hu, J., Cao, Q., Hui, P., Wang, X.: Understanding cross-site linking in online social networks to appear in ACM Transactions on the Web (2018)
14. Han, J., Choi, D., Chun, B.G., Kwon, T., Kim, H.C., Choi, Y.: Collecting, organizing, and sharing pins in Pinterest: Interest-driven or social-driven? In: Proc. of ACM SIGMETRICS (2014)
15. Hirsch, J.E.: An index to quantify an individual's scientific research output. PNAS **102**(46), 16,569–16,572 (2005)
16. Hu, Y., Manikonda, L., Kambhampati, S.: What we Instagram: A first analysis of Instagram photo content and user types. In: Proc. of AAAI ICWSM (2014)
17. Jain, P., Kumaraguru, P., Joshi, A.: Other times, other values: Leveraging attribute history to link user profiles across online social networks. Soc. Netw. Anal. Min. **6**(1), 85 (2016)
18. Jang, J.Y., Han, K., Shih, P.C., Lee, D.: Generation like: Comparative characteristics in Instagram. In: Proc. of the ACM CHI, pp. 4039–4042 (2015)
19. Jin, L., Chen, Y., Wang, T., Hui, P., Vasilakos, A.V.: Understanding user behavior in online social networks: A survey. Commun. Mag. IEEE **51**(9), 144–150 (2013)
20. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. **110**(15), 5802–5805 (2013)
21. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proc of WWW (2010)
22. Li, J., Peng, W., Li, T., Sun, T., Li, Q., Xu, J.: Social network user influence sense-making and dynamics prediction. Expert Syst. Appl. **41**(11), 5115–5124 (2014)
23. Liu, Q., Xiang, B., Yuan, N.J., Chen, E., Xiong, H., Zheng, Y., Yang, Y.: An influence propagation view of pagerank. ACM Trans. Knowl. Discov. Data **11**(3), 30:1–30:30 (2017)
24. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: Proc. of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), pp. 3818–3824 (2016)
25. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
26. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in foursquare. In: Proc. of AAAI ICWSM (2011)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the Web Stanford InfoLab (1999)
28. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
29. Song, X., Chi, Y., Hino, K., Tseng, B.: Identifying opinion leaders in the blogosphere. In: Proc. of ACM CIKM (2007)
30. Suhara, Y., Xu, Y., Pentland, A.S.: DeepMood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In: Proc. of the 26th International Conference on World Wide Web (WWW), pp. 715–724 (2017)
31. Tang, J., Lou, T., Kleinberg, J.: Inferring social ties across heterogenous networks. In: Proc. of ACM WSDM (2012)
32. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
33. Vasconcelos, M.A., Ricci, S., et al.: Tips, dones and to-dos: Uncovering user profiles in foursquare. In: Proc. of ACM WSDM (2012)
34. Vasconcelos, M.A., Ricci, S.M.R., Almeida, J.M., Benevenuto, F., Almeida, V.A.F.: Tips, dones and todos: Uncovering user profiles in foursquare. In: Proc. of ACM WSDM (2012)
35. Venkatadri, G., Goga, O., Zhong, C., Viswanath, B., Gummadi, K.P., Sastry, N.: Strengthening weak identities through inter-domain trust transfer. In: Proc. of WWW (2016)
36. Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., Zhao, B.Y.: You are how you click: Clickstream analysis for Sybil detection. In: Proc. of USENIX Security (2013)
37. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: Finding topic-sensitive influential Twitterers. In: Proc. of ACM WSDM (2010)
38. Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P., Zhao, B.Y.: User interactions in social networks and their implications. In: Proc. of ACM EuroSys (2009)
39. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proc. of ICML (1997)
40. Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B.Y., Dai, Y.: Uncovering social network Sybils in the wild. ACM Trans. Knowl. Discov. Data **8**(1), 2:1–2:29 (2014)
41. Ye, S., Wu, S.F.: Measuring message propagation and social influence on Twitter.com. Int. J. Commun. Netw. Distrib. Syst. **11**(1), 59–76 (2013)

42. Yuan, N.J., Zhang, F., Lian, D., Zheng, K., Yu, S., Xie, X.: We know how you live: Exploring the spectrum of urban lifestyles. In: Proc. of ACM COSN (2013)
43. Zhong, C., Chang, H.w., Karamshuk, D., Lee, D., Sastry, N.: Wearing many (social) hats: How different are your different social network personae? In: Proc. of AAAI ICWSM (2017)
44. Qingyuan, G., Yang, C., Xinlei, H., Fei, L., Xin, W., Yu, X., Xiaoming, F., Pan, H.: Identification of Influential Users in Emerging Online Social Networks Using Cross-Site Linking. In: Proc. of ChineseCSCW (2018)

Affiliations

Qingyuan Gong[1,2] · Yang Chen[1,2] · Xiaolong Yu[1,2] · Chao Xu[1,2] · Zhichun Guo[1,2] · Yu Xiao[3] · Fehmi Ben Abdesslem[4] · Xin Wang[1,2] · Pan Hui[5,6]

Yu Xiao
yu.xiao@aalto.fi

Fehmi Ben Abdesslem
fehmi.ben.abdesslem@ri.se

Xin Wang
xinw@fudan.edu.cn

Pan Hui
panhui@cs.helsinki.fi

1 School of Computer Science, Fudan University, Shanghai, China

2 Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China

3 Department of Communications and Networking, Aalto University, Espoo, Finland

4 RISE SICS, Kista, Sweden

5 Department of Computer Science, University of Helsinki, Helsinki, Finland

6 CSE Department, Hong Kong University of Science and Technology, Hong Kong, Hong Kong