Understanding the User Behavior of Foursquare: A Data-Driven Study on a Global Scale

Yang Chen, *Senior Member, IEEE*, Jiyao Hu, Yu Xiao, *Member, IEEE*, Xiang Li, *Senior Member, IEEE*, and Pan Hui, *Fellow, IEEE*

***Abstract*—Being a leading online service providing both local search and social networking functions, Foursquare has attracted tens of millions of users all over the world. Understanding the user behavior of Foursquare is helpful to gain insights for location-based social networks (LBSNs). Most of the existing studies focus on a biased subset of users, which cannot give a representative view of the global user base. Meanwhile, although the user-generated content (UGC) is very important to reflect user behavior, most of the existing UGC studies of Foursquare are based on the check-ins. There is a lack of a thorough study on tips, the primary type of UGC on Foursquare. In this article, by crawling and analyzing the global social graph and all published tips, we conduct the first comprehensive user behavior study of all 60+ million Foursquare users around the world. We have made the following three main contributions. First, we have found several unique and undiscovered features of the Foursquare social graph on a global scale, including a moderate level of reciprocity, a small average clustering coefficient, a giant strongly connected component, and a significant community structure. Besides the singletons, most of the Foursquare users are weakly connected with each other. Second, we undertake a thorough investigation according to all published tips on Foursquare. We start from counting the numbers of tips published by different users and then look into the tip contents from the perspectives of tip venues, temporal patterns, and sentiment. Our results provide an informative picture of the tip publishing patterns of Foursquare users. Last but not least, as a practical scenario to help third-party application providers, we propose a supervised machine learning-based approach to predict whether a user is an influential by referring to the profile and UGC, instead of relying on the social connectivity information. Our data-driven evaluation demonstrates that our approach can reach a good prediction performance with an F1-score of 0.87 and an AUC value of 0.88. Our findings provide a systematic view of the behavior of Foursquare users and are constructive for different relevant entities, including LBSN service providers, Internet service providers, and third-party application providers.**

***Index Terms*—Data-driven study, location-based social networks (LBSNs), machine learning, social graph analysis, social influence, tips.**

Manuscript received June 20, 2019; revised February 1, 2020; accepted April 18, 2020. Date of publication May 18, 2020; date of current version August 6, 2020. This work was supported by the National Natural Science Foundation of China under Grant 71731004 and Grant 61602122, in part by the CERNET Innovation Project under Grant NGII20190105, in part by the Project "PCL Future Greater-Bay Area Network Facilities for Large-Scale Experiments and Applications" under Grant LZC0019, in part by the Research Grants Council of Hong Kong under Grant 16214817, and in part by the 5GEAR project and FIT project from the Academy of Finland. *(Corresponding author: Yang Chen.)*

Yang Chen and Jiyao Hu are with the School of Computer Science, Fudan University, Shanghai 200433, China, also with the Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: chenyang@fudan.edu.cn; jyhu@cs.duke.edu).

Yu Xiao is with the Department of Communications and Networking, Aalto University, 02150 Espoo, Finland (e-mail: yu.xiao@aalto.fi).

Xiang Li is with the School of Information Science and Technology, Fudan University, Shanghai 200433, China (e-mail: lix@fudan.edu.cn).

Pan Hui is with the Department of Computer Science, University of Helsinki, 00100 Helsinki, Finland, and also with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: panhui@cs.helsinki.fi).

Digital Object Identifier 10.1109/TCSS.2020.2992294

I. Introduction

THE rapid development of mobile computing technologies and social networking services drive significant growth of location-based social networks (LBSNs), such as Foursquare [18], [42], [43], [47], [57], Yelp [23], [65], and Dianping [17], [27]. These networks not only help users interact with each other but also offer them location-centric functions. This service records a rich set of user information, including social connections between users, spatial and temporal information of user activities, and opinions expressed by users. The comprehensive LBSN data can be used to predict the movement of a massive number of users.

Being the most popular LBSN, however, a number of important characteristics of Foursquare are still unknown, for example, how the users link with each other on a global scale. In addition, most of the existing studies rely on the data of a biased subset of Foursquare users, and accordingly the analytical results cannot represent the entire Foursquare user base. As in [26], [43], [47], and [49], researchers obtained the Foursquare data via Twitter, since some Foursquare users have chosen to automatically republish their posts on Twitter. Unfortunately, as shown by Gong *et al.* [18], Foursquare users who have linked their accounts to Twitter are more active than the other Foursquare users. In detail, the users who have linked their accounts to Twitter have more followers and followings in general, and tend to publish more tips/check-ins. As a result, if we collect the Foursquare data solely from Twitter, we can only obtain the user activity data of a set of Foursquare users who are more active, and the corresponding data set cannot reflect the user activities of the entire Foursquare population. To overcome these drawbacks, in this article, we crawl the social connections and published tips of all 60+ million Foursquare users using a distributed way, and conduct a data-driven study based on the collected data set. We not only analyze the global social graph of Foursquare but also pay attention to tips, the primary form of user-generated content (UGC) on Foursquare. Our goal is to understand the user behavior of Foursquare on a global scale by answering the following three questions: 1) what unique characteristics can we learn from the structure of the global social graph of Foursquare? 2) Can we extract useful tip publishing patterns from different angles by referring to all published tips on Foursquare? 3) From the perspective of third-party application providers, how can we leverage the user profiles and UGCs to identify the influentials without referring to the social connectivity information? We introduce a data-driven study to answer these questions, by examining the global social graph and all published tips of Foursquare. We conducted the data collection from January to February 2016, and we collected the global social graph of Foursquare covering 61.43 million users, and the detailed information of all 55.18 million tips published by these users. By analyzing the collected data set, we have made the following three key contributions.

First, we analyze the global social graph of Foursquare for the first time in Section III-A. This graph connects 60+ million users around the world. Social graphs have been used as a representative model to characterize the social connections among users, and applied in several practical applications including friend recommendation [40], data placement [29], [69], information diffusion [34], and social data transmission [64]. However, given the large user population of Foursquare, obtaining a global social graph with all users is very challenging. Therefore, most of the existing social graph-based studies on Foursquare [32], [46], [47] are based on biasedly collected subgraphs with at most hundreds of thousands of users, which cannot accurately reflect the structural properties of the global social graph. In this article, we crawl the global Foursquare social graph with 60+ million users for the first time, characterizing the social connectivity among all Foursquare users. By using different graph analysis methods, we have found several undiscovered features of this global graph, such as a moderate level of reciprocity, a small average clustering coefficient, a giant strongly connected component, and a significant community structure. We also see that most of the Foursquare users are weakly connected with each other besides the singletons. These structural characteristics are helpful for Foursquare or similar LBSN service providers to study the social interactions and information dissemination among all users on a global scale.

Second, we focus on tips, the primary form of UGC on Foursquare in Section III-B. Tips are very useful for understanding users' opinions, distributions, and movements. However, many of the existing analytical works are based on the check-in activities [42], [43], [49], [57]. Unfortunately, the check-in function is no longer supported by Foursquare since Aug. 2014, not to mention that according to the study of Zhang *et al.* [67], about 75% of all check-ins do not match the real trajectories of users. A systematic study of tips is needed. Some existing works [38], [54] about tips are based on the data set collected in a biased way, and the understanding of the tip texts is limited. To fill this gap, we propose a detailed study based on all 55.18 million published tips on Foursquare. We first get a statistical analysis of the numbers of tips published by different groups of users. We further dive into the texts of all tips and analyze the tips from the perspectives of tip venues, temporal patterns, and sentiments. Our analytical findings show the first comprehensive view of Foursquare tips.

Last but not least, as a practical application scenario to help third-party application providers, we propose the idea of social influence prediction according to a user's profile and UGC. Traditionally, many social influence metrics are based on the information of social connectivity, for example, number of followers [9] and PageRank [34]. However, nowadays, many representative online social networks (OSNs), such as Facebook, allow a user to hide her friend list. Therefore, from the perspective of third-party application providers, we might not be able to determine whether a user is an influential if the social graph information is not fully available. To remedy this problem, we first explore the relationship between social influence and users' profiles and UGCs in Section III-C. We can see that several information fields within a user's profile, and her content publishing patterns, could be used as indicators to judge whether she is an influential. Based on this intuition, we build a supervised machine learning-based model to predict whether a user is an influential by referring to her profile and UGC. According to our study, we find that our approach can uncover the influentials with high accuracy, achieving an F1-score of 0.87 and an AUC value of 0.88. Our approach provides an accurate and convenient way for third-party application providers to determine whether a Foursquare user is an influential, without relying on the social graph information.

Our study presents a systematic understanding of Foursquare, the representative LBSN service, including global social connectivity, content publishing behavior, and the prediction of social influence. The analytical results are constructive for different relevant entities: 1) for Foursquare itself, or similar LBSN service providers, we get a comprehensive understanding of the social connections from a global view. In other words, we construct and analyze the global Foursquare social graph of 60+ million users. This graph is helpful to study the information diffusion and social interactions of Foursquare. Meanwhile, by referring to user profiles and published tips, we know the geographic distributions of users and venues around the world. We also study the evolution of user activities. All these information are useful for LBSN service providers to schedule the resource provisioning to serve millions of users in a scalable and cost-effective way. In addition, by referring to the published tips, they can extract the opinions and movements of users. The tip information can be further applied for venue recommendations and user profiling; 2) for Internet service providers (ISPs), the understanding of the geographic distribution, content generation behavior, and interaction patterns of users from an evolutionary view can be used to characterize the traffic patterns of LBSNs. Therefore, the ISPs would be able to adjust the network resources flexibly to enhance the network performance of LBSN services; 3) for third-party application providers, the massive spatiotemporal information of Foursquare users can reflect the real-time geographic distribution of users from time to time. Such information is important for urban computing-related applications [26], [49], [62]. Also, we provide a supervised machine learning-based approach to uncover influentials conveniently for third-party application providers, without the need of referring to the social connectivity information.

II. Background and Data Collection

A. Foursquare Overview

Since 2009, Foursquare has been a leading site for the combination of location-based services (LBS) and mobile social networking. Different from traditional OSNs [30], such as Facebook, activities on Foursquare are location-centric. In the original Foursquare app released in 2009, the two key functions were conducting check-ins and leaving tips. However, in May 2014, the original Foursquare app was split into two apps, i.e., the new Foursquare and Swarm apps. These two apps share the same user database, but with a different focus. The Swarm app [11] supports the check-in function and provides a life-logging service. Differently, in the current version of Foursquare, a tip is the primary form of UGC. A tip can reflect the publisher's opinion toward a selected venue.

B. Data Crawling and Ethical Issues

The rapid development of OSNs have attracted millions of users and have produced a large amount of data for user behavior study. A number of articles have studied user behavior by crawling one or multiple OSN sites, such as Facebook [61], Twitter [34], [59], Pinterest [21], and Quora [56]. To analyze the properties of the entire Foursquare population, we aim to obtain a data set covering all Foursquare users. Getting such a data set is challenging. Similar to other OSN sites, Foursquare also employs an IP-based rate limiting policy, which prevents earlier researchers from getting a complete data set. To crawl the data quickly, we apply the crowd crawling framework [14], which allows us to use a pool of IP addresses to improve the crawling efficiency. We launched 60 virtual machines on the East US data center of the Microsoft Azure platform. Each virtual machine had an independent IP address. These virtual machines worked collaboratively to crawl the data.

Each user on Foursquare has a numeric ID, and the IDs are assigned sequentially. The earlier a user registered, the smaller the user ID the user has. Therefore, we can get the maximum ID number by registering a new account. Note that some user IDs are unassigned. To avoid the bias introduced by newly registered users, we focus on all users who have registered for more than half a year. We divide the entire Foursquare ID space evenly into 60 chunks, and each virtual machine is responsible for one chunk of IDs. We did the data collection from January 7 till February 10, 2016 by using a Python-based crawler implemented by us. Our crawling covered all Foursquare users who had registered by July 2015. We have successfully fetched the publicly visible data of 61.43 million users. For each user, we downloaded the profile and extracted the name, the number of published tips, the number of followings, and the number of followers. Also, we fetched the optional information such as the profile photograph, the gender information, the current location, the Facebook ID, the Twitter ID, and the biography from the profile [18]. In addition, we also crawled all tips published by the user, and the IDs of the user's followings and followers. For each tip, we can get the time when it was published, the text of the tip, the country of the venue, and the category information of the venue. An example tip can be represented as {user_id:12345, tip_text: "Super nice restaurant! It provides nice food, and it has a great location!" Venue_country: US, venue_category: food, date: "October 21, 2015," and time: "21:23:59"}.

Note that we respect the privacy of Foursquare users, and only crawled the publicly visible information for our study using the official Foursquare API. Before undertaking the analysis, we anonymized the IDs of all users. In addition, we stored and analyzed the anonymized data in an offline environment. The ethical assessment of this article has been reviewed and approved by the Research Department of Fudan University.

Among all the 61.43 million users, 67.22% of them had uploaded a profile photograph, while the rest 32.78% had not. Regarding the gender information, we can see that 51.31% of them were male; 42.16% were female, and the rest 6.53% did not want to disclose their gender information. For the users' home countries, 87.75% had added some information to the "location" field. We used the Google Maps Geocoding API to infer each user's home country. In our data set, the top four countries with the biggest numbers of Foursquare users were USA (30.36%), Turkey (13.06%), Indonesia (9.76%), and Brazil (6.26%).

III. Data-Driven User Behavior Analysis of Foursquare

In this section, we study the crawled Foursquare data and extract the user behavioral patterns from different aspects. In Section III-A, we analyze the Foursquare social graph of 61.43 million users on a global scale and extract a number of undiscovered characteristics of this massive graph. In Section III-B, we study the tip publishing behavior of all Foursquare users. We first study the numbers of tips published by different users, and then dive into the contents of tips, by referring to the aspects of tip venues, temporal patterns, and sentiment. In Section III-C, as a practical scenario to help third-party application providers, we propose a supervised machine learning-based approach to predict whether a Foursquare user is an influential based on the profile and UGC, instead of relying on the social connectivity information.

A. Social Graph Analysis

Social graphs have been widely used to describe the connections between OSN users. The asymmetric "following" relationship among Foursquare users can be modeled as a directed graph $G = (V, E)$. A node $v \in V$ represents a Foursquare user. When user A follows user B, there will be a directed edge $(v_A, v_B) \in E$. There are 61.43 million nodes and 2.67 billion edges in this graph, characterizing social connections among Foursquare users. In this section, we present the first comprehensive analysis of the global

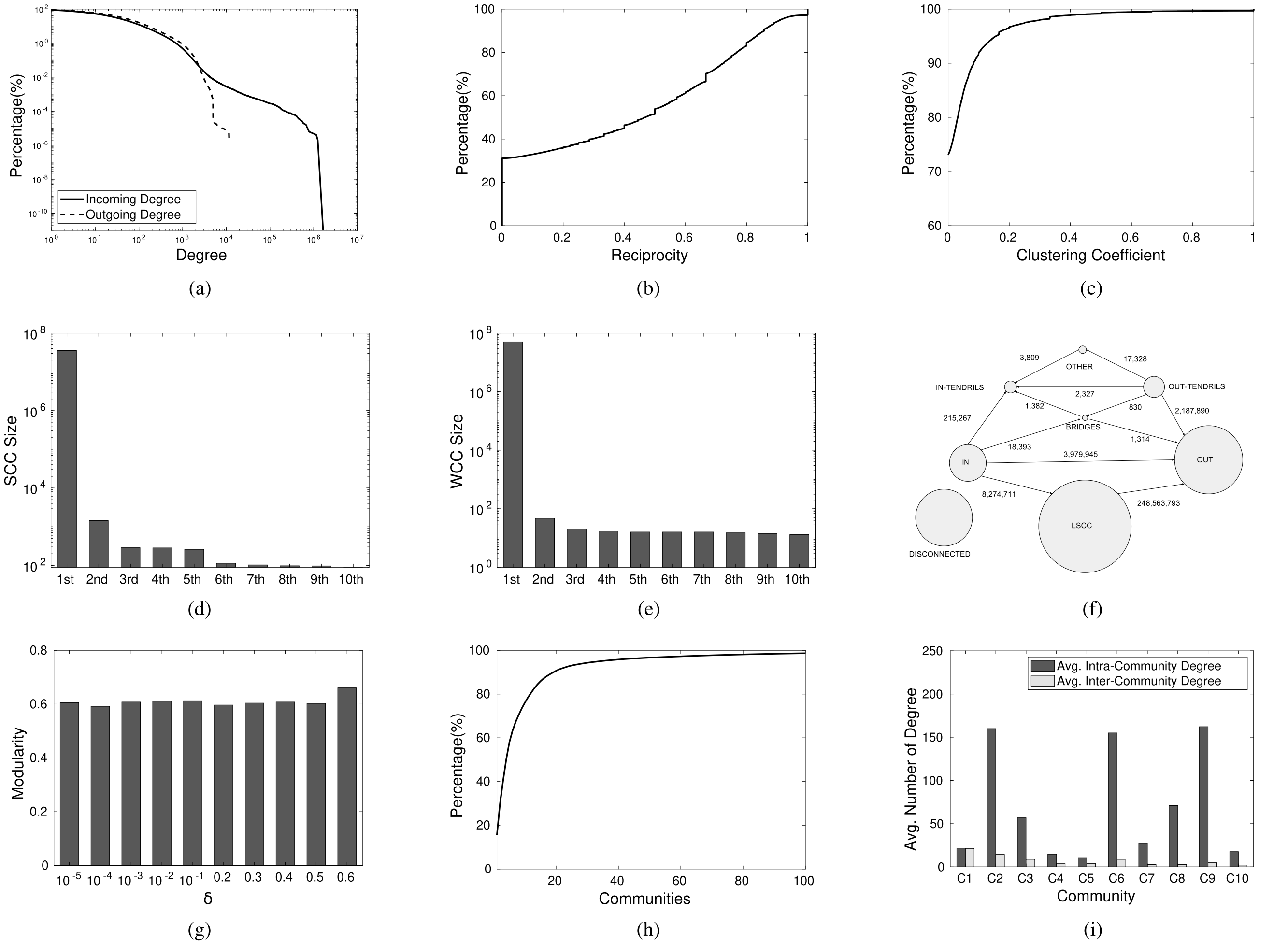


Fig. 1. Social graph analysis. (a) Degree. (b) Reciprocity. (c) Clustering coefficient. (d) Strongly connected components. (e) Weakly connected components. (f) Macrostructure. (g) Modularity. (h) Coverage of top 100 communities. (i) Intra-/Inter-community edges.

Foursquare social graph of all 60+ million users. A number of classic graph metrics are examined.

1) Degree: The incoming degree of a node is defined as the number of followers the corresponding user has. The outgoing degree of a node is denoted as the number of users the corresponding user follows. We also call it the number of followings. Fig. 1(a) shows the complementary cumulative distribution function (CCDF) of the incoming and outgoing degrees of the Foursquare social graph. We can see that very few nodes can have more than one million followers. Regarding the number of followings, the maximal number is only 13 830. Therefore, normally a user does not follow a large number of other users, while a users' total number of followers can exceed one million. In other words, the distributions of the numbers of followings and that of followers are quite different.

As in (1), we calculate the *balance* metric for each user [58]. It defines the ratio of the number of followers to the number of followings:

$$B(u) = \frac{\mid \text{follower}(u) \mid}{\mid \text{following}(u) \mid}. \tag{1}$$

For a given user u, if $B(u)$ is between 0.5 and 2, we regard u as a *well balanced user*. Looking at the entire network, about 57.45% of users are well balanced.

We also evaluate the *reciprocity* metric [41], another metric to quantify the symmetric relationship between users. By default, the reciprocity metric is defined at the level of the entire graph. The reciprocity of graph G, i.e., r_G, can be calculated as the fraction of the number of edges pointing in both directions to the total number of edges in the graph. For the Foursquare social graph G, the value of r_G is 0.42.

The reciprocity metric can also be defined at the per-user level. As shown in the following equation, for user u, $r_G(u)$ is defined as the ratio of the number of edges in both directions and the sum of u's numbers of followings and followers:

$$r_G(u) = \frac{2\times \mid \text{following}(u) \cap \text{follower}(u) \mid}{\mid \text{following}(u) \mid + \mid \text{follower}(u) \mid}. \tag{2}$$

According to Fig. 1(b), i.e., the cumulative distribution function (CDF) of reciprocity, we can see that about 48.51% of users have a reciprocity value larger than or equal to 0.5.

2) Clustering Coefficient: The clustering coefficient (CC) of a node is a measure of the degree to which nodes in the social graph G tend to cluster together. For a node v_i, its neighborhood N_i is defined as immediately connected nodes, i.e., $N_i = \{j \in V \mid i \neq j, (v_i, v_j) \in E \; or \; (v_j, v_i) \in E\}$. Since G is a directed network, (v_i, v_j) is different from (v_j, v_i). The clustering coefficient of node i is defined as follows:

$$C_i = \begin{cases} \frac{|N_i|}{k_i(k_i-1)}, & k_i > 1 \\ 0, & k_i \in \{0, 1\} \end{cases} \tag{3}$$

where k_i is the number of neighbors of node v_i. The CC value of a node is between 0 and 1. Fig. 1(c) shows the CDF of the clustering coefficient of all nodes. The average CC of the entire network is only 0.065, indicating that the Foursquare social graph is loosely connected.

3) Strongly Connected Components and Weakly Connected Components: A strongly connected component of a directed graph is a subgraph that all nodes are strongly connected. For any node pair (u, v) in this subgraph, there is a directed path from u to v, and a directed path from v to u. Meanwhile, no additional edges or nodes can be added to this subgraph without breaking its property of being strongly connected. According to Fig. 1(d), the sizes of the two largest strongly connected components are 35 583 350, and 1440, respectively. The largest strongly connected component (LSCC) covers 57.92% of all nodes.

If we convert all edges of a directed graph into undirected edges, we can define a weakly connected component if there is a path between any node pair in this subgraph, and no additional edges or nodes can be added to this subgraph without breaking the weakly connected property. According to Fig. 1(e), the sizes of the two largest weakly connected components are 50 568 619, and 47, respectively. The largest weakly connected component (LWCC) covers 82.32% of all nodes.

The LWCC is much larger than the LSCC, while the second-largest WCC is much smaller than the second-largest SCC. Among all nodes, 17.50% of them are singletons, i.e., they do not have any following or followers. In other words, most of the Foursquare users, besides the singletons, are weakly connected with each other.

4) Macrostructure: To abstract the social graph from a high-level view, we study the macrostructure of the global social graph. The method we adopt was proposed by Broder *et al.* [5] to study the structure of web pages, and was further improved by Gabielkov *et al.* [16] to analyze the social graph of Twitter. As shown in Fig. 1(f), we divide the entire social graph into 8 components, i.e., LSCC, IN, OUT, IN-TENDRILS, OUT-TENDRILS, BRIDGE, OTHER, and DISCONNECTED. Each node belongs to one of them. As mentioned earlier, LSCC stands for the largest strongly connected component. IN covers the nodes with a directed path to any node in LSCC, and OUT covers the nodes with a directed path from any node in LSCC. Afterward, if we run a breadth-first search (BFS) starting from a node in the IN component, and a reverse starting from the OUT component, reachable nodes, besides the ones in the LSCC, IN, or OUT components, are chosen as IN-TENDRILS and OUT-TENDRILS, respectively. The BRIDGE component contains a set of nodes connecting the IN and OUT components bypassing the LSCC. In Fig. 1(f), the sizes of different components are positively correlated with the number of users in each of them. Meanwhile, the numbers on arrows indicate the numbers of links between components. The numbers and percentages of users in each component are shown in Table I. The LSCC, IN, and OUT components cover 81.04% of nodes.

TABLE I

SIZES OF THE COMPONENTS (MACROSTRUCTURE)

Component	Number of Users	Percentage (%)
LSCC	35.58 million	57.92%
IN	1.48 million	2.41%
OUT	12.72 million	20.71%
DISCONNECTED	10.75 million	17.50%
IN-TENDRILS	193678	0.32%
OUT-TENDRILS	553793	0.90%
BRIDGES	16910	0.03%
OTHER	132914	0.22%

TABLE II

PERCENTAGE OF USERS IN TOP FOUR COUNTRIES PER COMMUNITY

Community	Countries (% of Users)			
$C1$	US(36.05%)	ID(12.94%)	TH(4.76%)	MY(4.52%)
$C2$	US(38.59%)	ID(7.89%)	TR(6.20%)	IN(3.24%)
$C3$	TR(80.09%)	US(7.50%)	ID(2.77%)	BR(1.15%)
$C4$	BR(66.92%)	US(13.48%)	ID(2.97%)	PT(2.61%)
$C5$	RU(40.09%)	US(20.78%)	UA(11.97%)	ID(4.11%)
$C6$	US(39.54%)	ID(6.11%)	MX(4.22%)	GB(3.08%)
$C7$	TR(81.58%)	US(7.16%)	ID(2.36%)	BR(1.03%)
$C8$	SA(28.83%)	US(18.64%)	KW(9.76%)	AE(4.82%)
$C9$	BE(64.72%)	US(13.59%)	ID(2.54%)	DE(2.18%)
$C10$	TR(82.85%)	US(6.53%)	ID(2.39%)	BR(0.98%)

5) Communities: It is quite common for users of an OSN to exhibit a community structure. A community is composed of a number of nodes, and these nodes are densely connected internally. Meanwhile, there are fewer intercommunity connections.

We apply the widely used Louvain algorithm [3] to group users into different communities. This algorithm was designed to process undirected networks. Following the earlier approaches in [25], [44], and [70], we convert the social graph into an undirected graph. Louvain algorithm optimizes the *modularity* metric. If there are c communities, the modularity Q is defined as follows:

$$Q = \sum_{i=1}^{c} \left(e_{ii} - a_i^2\right). \tag{4}$$

In (4), e_{ii} denotes the fraction of edges with both ends in the ith community, and a_i means the fraction of edges with at least one end in the ith community. The value of Q is between -1 and 1. A value of Q larger than 0.3 means that the network has a significant community structure [33]. After running the algorithm, each node will be assigned to a selected community. As discussed in [68], there is a δ parameter, which is a critical tuning parameter for the Louvain algorithm. In Fig. 1(g), we can see that different choices of δ will lead to a similar Q value around 0.6. Therefore, the Foursquare social graph demonstrates a viable community structure.

According to Fig. 1(h), top ten communities cover 76.14% of all nonsingleton users, and top 30 communities cover 94.32% of nonsingleton users. Therefore, although the nonsingleton users form 55 396 communities, most of them belong to the top few ones. We show the average intra-community degree and inter-community degree of each of the top ten communities in Fig. 1(i). We find that only for the largest community, the average intra-community degree is slightly larger than the average inter-community degree. For each of the other communities, the average intra-community degree is much larger than the average inter-community degree. We also look at the user composition of the top ten communities, in terms of their home countries in Table II. Many of them have one or two dominant countries. For C3, C4, C7, C9, and C10, each of them has a country which covers more than 60% of the users.

6) Summary and Discussion: In this section, we use a series of graph metrics to understand the structural properties of the global Foursquare social graph, which has never been reported in literatures before. According to our studies of the followings and followers of all users, we find that the entire Foursquare social graph has a reciprocity value of 0.42. This is a moderate value, which is larger than that of some existing Twitter data sets, while Twitter is a representative directed social network. For example, the Twitter data set collected by Kwak *et al.* [34] in July 2009 (41.7 million users and 1.47 billion social relationships) has a reciprocity value of 0.22. The Twitter data set collected by Watanabe *et al.* [59] from July 2012 to October 2012 (469.9 million users and 28.7 billion social relationships) has a reciprocity value of 0.19. The average clustering coefficient is only 0.065, indicating the global social graph of Foursquare is loosely connected. Regarding the distributions of strongly/weakly connected components, there is a huge LSCC covering nearly 60% of users. Also, besides the singletons, most of the Foursquare users are weakly connected. The social graph reveals a clear community structure, with a Q value of about 0.6.

B. Tip Publishing Behavior

A tip is the primary form of UGC on Foursquare. A tip records the detailed opinion of a user for a selected venue. However, many of the existing measurement works on Foursquare [42], [43], [49], [57] are based on the check-in data. In this subsection, we explore the tip publishing behavior from different aspects to learn the characteristics of UGC on Foursquare. We first count the numbers of tips published by different users, and do comparative study among user groups. Then we look into the contents of tips and perform analysis from the angles of tips venues, temporal patterns, and sentiment. Our results provide an informative picture of tip publishing behavior from a global view.

1) Numbers of Tips Published by Different Users: In this subsection, we first consider the number of tips each user has published and analyzed the distribution of each user's total number of published tips. In Fig. 2, we show the CDF of the number of tips published by each user. 83.47% users have never published any tip. In other words, most of the Foursquare users are tip readers rather than publishers. Therefore, the average number of a user's published tips is only 0.89. We also rank the users according to the number of published tips. We find that the top 1% users published 47.54% of tips, and the top 10% users published 92.67% of tips.

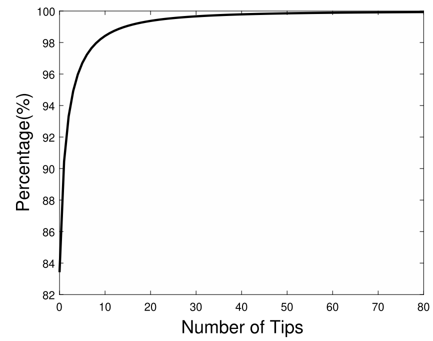


Fig. 2. CDF of the number of tips per user.

Besides considering all users as a whole, we also divide users into groups based on the availability of their profile photographs, gender information, and home country information, respectively. According to Fig. 3(a), among the users who have uploaded profile photographs, the average number of published tips per user is 1.28. In contrast, this number is only 0.09 for users without profile photographs. Therefore, whether the user has uploaded a profile photograph is an indicator of the number of published tips. For a user who has uploaded the profile photograph, the user has a higher chance to publish more tips on Foursquare. As shown in Fig. 3(b), the average number of tips published by male users is 0.85, while this number is 0.95 for female users. In other words, on average, female users published about 11.76% more than male users in terms of the number of tips. In Fig. 3(c), we look at the users' home countries and focus on the users from the ten countries with the highest Foursquare population. Among them, users in Russia and Mexico are more active in publishing tips, while users from Indonesia are the least active ones.

On Foursquare, about 0.071% of all users are known as "super users." These users are selected by Foursquare, and are allowed to edit the information of venues. Fig. 3(d) shows the difference between super users and ordinary users. Obviously, super users publish much more than ordinary users. As shown in Fig. 3(e), we group users according to their cross-site linking configurations [18]. Cross-site linking is a key function of Foursquare, allowing Foursquare users to link their accounts on leading OSNs, for example, Facebook and Twitter. We find that users who have enabled the cross-site linking function tend to publish more. We also classify users according to the configuration of the optional fields in their profiles [18]. On Foursquare, there are five optional fields in a user's profile, i.e., profile photograph, gender, residential location, last name, and biography. For users who have enabled all these five fields, we denote them as "open users," since they want to keep their profiles complete. For users who decline to provide anything to these five fields, we call them "cautious users," since they do not want to disclose any nonmandatory information. We regard the rest of users as "other users." According to Fig. 3(f), users who care more about their privacy publish less on Foursquare.

2) Venue Analysis: Each tip must be associated with a certain venue. Therefore, analyzing the venue data is also very important to understand the tip publishing patterns. In this section, we investigate the properties of venues on Foursquare.

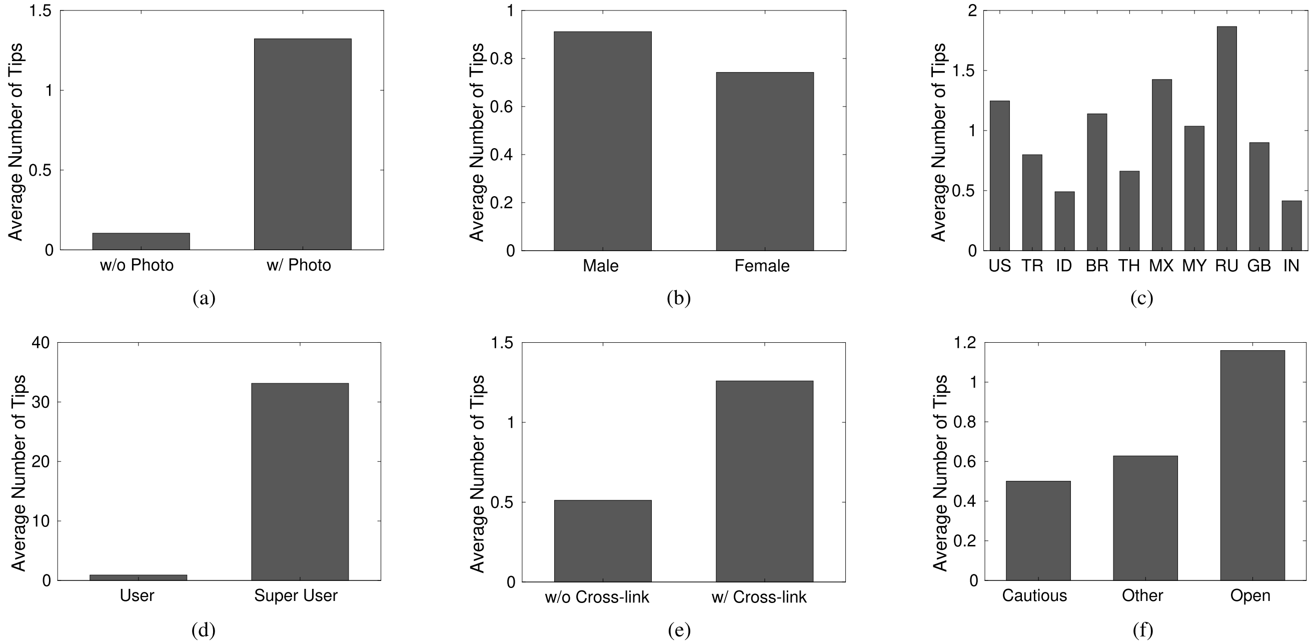


Fig. 3. Group-based analysis for the average number of tips per user. (a) Profile photograph. (b) Gender. (c) Home country. (d) User type. (e) Cross-site linking. (f) User privacy.

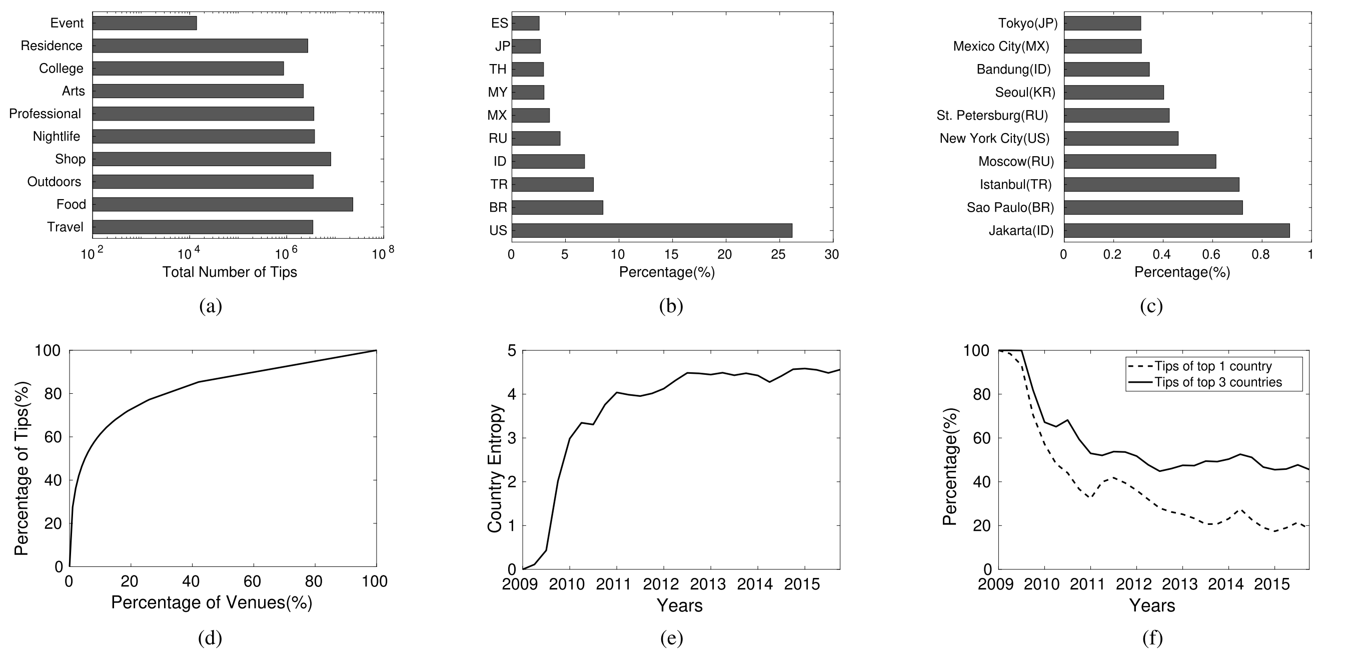


Fig. 4. Analysis of venues. (a) Venue category distribution. (b) Country distribution of the venues. (c) Top 10 cities of the venues. (d) Distribution of the number of tips per venue. (e) Evolution: Venue country entropy. (f) Evolution: Percentage of top countries.

For each tip, we also record the ID of the corresponding venue. By referring to all crawled tips, we have discovered 13.25 million venue IDs. We have further crawled the profiles of all these venues.

In Fig. 4(a), we show the venue category distribution of all published tips. The most popular venue category is "food," which has covered 45.06% of all tips. The second most popular venue category is "shop," covering 14.68% of all tips. For the rest of the categories, none of them has received more than 10% of tips. Therefore, restaurants are the most attractive venues for tip publishers on Foursquare. This is similar to the findings in [38], which are obtained based on the venue data covering 14 regions. Fig. 4(b) shows the percentages of the top ten countries in terms of the number of venues. We can see that these countries cover 68.28% of all venues. We also find that the top 20 countries cover 82.67% of all venues. In addition, the percentages of the top ten cities are shown in Fig. 4(c). These ten cities cover 5.21% of all venues,

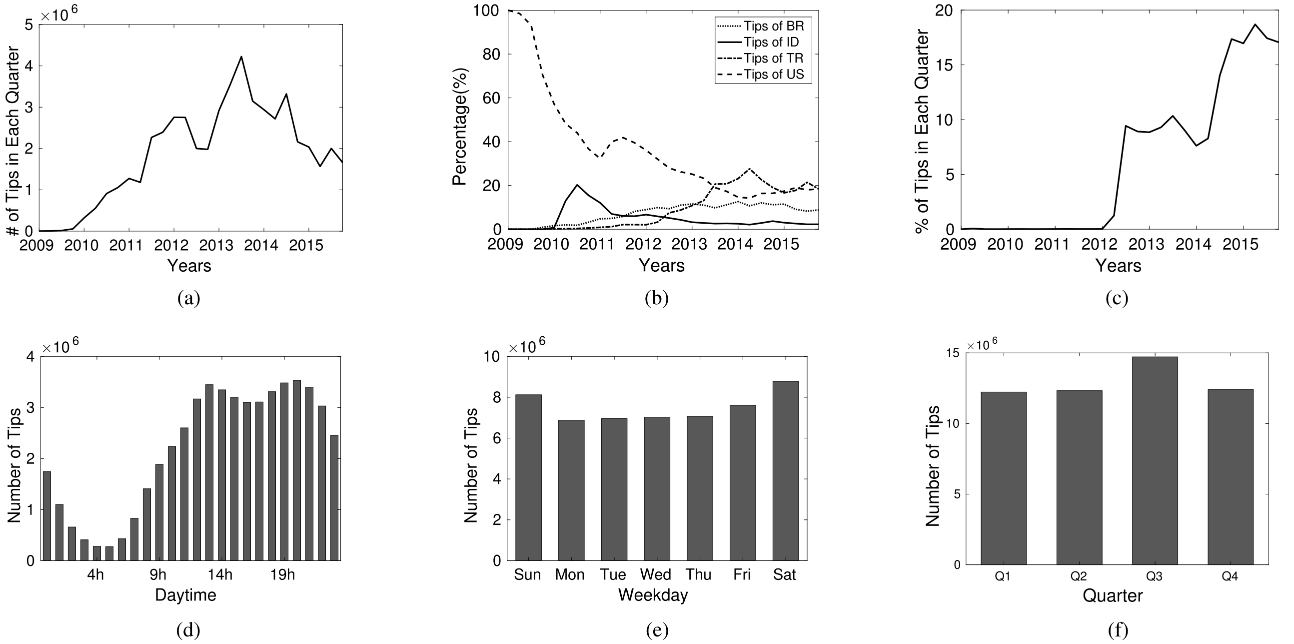


Fig. 5. Temporal analysis. (a) Number of tips. (b) % of tips in top four countries. (c) % of tips with photographs. (d) Daily temporal distribution. (e) Weekly temporal distribution. (f) Quarterly temporal distribution.

which is a noticeable percentage. We also rank the venues in terms of the number of received tips. According to Fig. 4(d), we find that the top 10% of venues have received 61.64% of all tips, and the top 20% of venues have received 73.14% of all tips. Therefore, a small portion of venues is attractive to the majority of users. Besides studying the evolution of the total number of tips, we are also interested in how the tips are distributed among different countries in each quarter. For a certain quarter, we use k to represent the number of countries which have been visited and use p_i to denote the probability of visiting the venues in the ith country. We introduce the concept of *venue country entropy* E, using the formula $E = -\sum_{i=1}^{k} p_i \log_2 p_i$. In Fig. 4(e), we show the evolution of the venue country entropy. It increased quickly in the first three years since 2009. This means the country distribution of visited venues was getting wider. Since the second quarter of 2012, the country distribution has reached a relatively stable status. In Fig. 4(f), we can see that the percentage of users coming from the top three countries was decreasing all the way until the second quarter of 2012. This confirms that in the first three years, Foursquare was popular in few countries, but its coverage has become wider since 2012.

3) Temporal Analysis: We study the tip publishing behavior from a temporal view by counting the number of tips published during different time periods. We cover both evolutionary and periodical patterns. Note that we use the local time according to the venue location of each tip.

In Fig. 5(a), we calculate the number of published tips in each quarter and analyze the evolution of the total number of published tips on a quarterly base. From the very beginning, the total number of tips increased steadily. After reaching a significant peak in the third quarter of 2013, this number started to decrease. In the mid of 2014, Foursquare released the Swarm app with a focus on check-in and location sharing. In August 2014, the check-in interface was removed from Foursquare. This made Foursquare a dedicated tip-sharing app and could at least boost the tip posting for a while. As a result, we can observe another peak in the third quarter of 2014. Fig. 5(b) illustrates the evolution of the percentage of tips published in each of the top four countries. In the very beginning, most of the tips were published by users from the USA. The percentage dropped until early 2014. Meanwhile, the percentage of tips published by users from Turkey kept growing until the second quarter of 2014. In the last quarter of 2015, the percentages of tips published by users from the USA, Indonesia, Turkey, and Brazil were 18.41%, 2.27%, 18.39%, and 8.84%, respectively. Some users prefer to add photographs to a tip to make it more illustrative. In Fig. 5(c), we can see the percentage of tips with photographs is increasing. At the end of 2015, nearly 20% of tips were with photographs.

Regarding periodical patterns, by aggregating the published tips into 24 h, we can see the daily temporal distribution of tip publishing in Fig. 5(d). People are more active in tip publishing between 9:00 and 24:00. In particular, there are two peaks in a day. One is around 13:00, and the other is around 19:00–21:00. Fig. 5(e) shows the distribution of tips in a week. Obviously, during Saturdays and Sundays, users are more active in publishing tips. In Fig. 5(f), we find that in the third quarter of a year, more tips are published. By evaluating the temporal patterns of tip publishing, we believe Foursquare will be able to dynamically schedule its resources to better serve its users.

4) Sentiment Analysis: After analyzing the venue and temporal information, we dive into the main component of a tip, i.e., the tip text. This component records the tip publisher's detailed comment for a venue. To understand the publishers' opinion of the tips, a series of sentiment analysis are

TABLE III
SENTIMENT ANALYSIS—HOME COUNTRY

Country	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
BR	39.61	54.24	6.15	0.67
US	63.27	24.88	11.86	0.76
TR	51.85	42.06	6.09	0.73
ID	51.79	41.61	6.60	0.73

TABLE IV
SENTIMENT ANALYSIS—GENDER

Gender	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
Male	60.65	27.55	11.80	0.74
Female	66.14	22.59	11.27	0.77

introduced. In our study, we calculate a "sentiment score" for each tip. We use a representative and widely used natural language processing (NLP) library called NLTK[1] [2], to extract the publisher's attitude from the text. Using this tool, we can obtain a sentiment score for each tip by using the VADER algorithm [28], which is designed for analyzing social media texts. A sentiment score is within the range of $[-1, 1]$. A score of -1 means the tip is surely negative, and a score of 1 means the tip is certainly positive. If a score is within the range of $(0, 1]$, we regard the tip as a positive tip. If a score is within $[-1, 0)$, we classify the tip as a negative tip. The rest of the tips, with a sentiment score of zero for each of them, are defined as neutral tips. As VADER can only process tips written in English, we filter out all tips published in other languages. The sentiment analysis of tips written in other languages would be a potential future work. For example, Alrumayyan *et al.* [1] presented a sentiment analysis of Arabic tips on Foursquare.

Among all tips written in English, 63% of them are positive, 26% are neutral, and the rest 11% are negative. In other words, there are much more positive tips on Foursquare. We use F_{pos} to denote the fraction of positive tips, F_{neu} to represent the fraction of neutral tips and F_{neg} for the fraction of negative tips. We have $F_{\text{pos}}+F_{\text{neu}}+F_{\text{neg}} = 1$. To study the overall sentiment of a set of tips, we introduce a new metric, called happiness index (H_{idx}). Intuitively, a higher percentage of positive tips will lead to a higher H_{idx}. Similarly, having more negative tips indicates a lower H_{idx}. We define H_{idx} using the following equation:

$$H_{\text{idx}} = F_{\text{pos}} + F_{\text{neu}}/2. \tag{5}$$

The value of H_{idx} is within the range of [0, 1]. A higher value of H_{idx} indicates a higher level of satisfaction. For all tips in our study, the overall H_{idx} is 0.76.

We also undertake group-based analysis as follows. We first group the tips according to the tip publishers' home countries. According to Table III, tips published by users from the United States have the highest H_{idx}, while the tips published by users from Brazil have the lowest H_{idx}. Also, we classify the tips according to the gender of the publisher. From Table IV, we find that the male users have a slightly lower H_{idx} than the female users.

[1] **http://www.nltk.org/**

TABLE V
SENTIMENT ANALYSIS—YEARLY EVOLUTION

Category	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
2009	58.29	33.44	8.27	0.75
2010	60.38	29.48	10.13	0.75
2011	58.94	28.72	12.34	0.73
2012	59.91	27.24	12.86	0.74
2013	64.92	23.08	12.00	0.77
2014	66.67	23.08	10.25	0.78
2015	69.63	21.37	9.00	0.80

TABLE VI
SENTIMENT ANALYSIS—HOUR OF THE DAY

Time slot	From-To	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
morning	6:00-10:00	62.62	26.17	11.21	0.76
noon	10:00-1400	63.26	25.76	10.98	0.76
afternoon	14:00-18:00	63.21	25.42	11.37	0.76
evening	18:00-22:00	63.27	25.06	11.67	0.76
night	22:00-2:00	63.66	24.75	11.59	0.76
late night	2:00-6:00	62.12	26.00	11.88	0.75

TABLE VII
SENTIMENT ANALYSIS—DAY OF THE WEEK

Category	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
Sunday	63.12	25.53	11.35	0.76
Monday	62.94	25.99	11.07	0.76
Tuesday	63.05	25.98	10.97	0.76
Wednesday	63.03	25.96	11.01	0.76
Thursday	63.10	25.56	11.34	0.76
Friday	63.59	24.67	11.74	0.76
Saturday	63.49	24.40	12.12	0.76

TABLE VIII
SENTIMENT ANALYSIS—VENUE CATEGORY

Category	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
Travel	60.05	24.77	15.19	0.72
Food	67.25	21.91	10.83	0.78
College	49.64	37.52	12.84	0.68
Nightlife	65.53	24.06	10.41	0.78
Event	59.30	33.47	7.23	0.76
Shop	61.38	27.00	11.63	0.75
Residence	50.17	36.96	12.86	0.69
Professional	55.10	33.63	11.27	0.72
Outdoors	62.66	28.03	9.31	0.77
Arts	61.35	27.57	11.08	0.75

Looking at the happiness index from a temporal perspective, Table V shows the yearly evolution of H_{idx}. We can see that the H_{idx} value dropped in 2011–2012, but increased again since 2013. From Tables VI and VII, we can see that H_{idx} varies very little among different slots of the day,[2] or among different days of the week.

To figure out which types of venues have a higher chance to receive positive tips, we undertake sentiment analysis for each venue category. From Table VIII, we can see that the tips published in "food," "nightlife," and "outdoors" categories have the highest values of H_{idx}, while the tips published in "college" and "residence" categories have the lowest H_{idx} values. The difference between Table VI/VII and Table VIII shows that, with different grouping criteria, we may observe different variations in H_{idx} values.

5) Summary and Discussion: In this subsection, we analyze the tip publishing behavior by referring to all published tips on Foursquare. By studying the numbers of tips published by different users, we can see that most of the Foursquare users

[2] As in [8], we divide the 24 h of a day into six slots.

are tip readers rather than publishers, and the top 10% users published 92.67% of tips. We also look into the contents of all published tips from the aspects of tip venues, temporal patterns, and sentiment. Based on our venue analysis, we find that "food" is the most popular venue category, and the coverage of Foursquare venues has expanded from a few countries to around the world. According to our temporal analysis, we explore both evolutionary and periodical patterns of tip publishing. Also, for sentiment analysis, we propose the "happiness index" metric, indicating the sentiment difference among different tip publishers, tip publishing time, and venue categories.

C. Prediction of Influentials

In Sections III-A and III-B, we study the global social graph and tip publishing behavior of Foursquare users. In this subsection, we aim to explore the relationship between a user's social influence and the user's profile and UGC. In a social network, different users have different levels of "social influence," which is a well-known concept in sociology and viral marketing [9]. Discovering influentials is useful for determining which users are more important within the social network.

In previous works, researchers have proposed different definitions of social influence, and have often chosen the Twitter platform for the case study. For example, Cha *et al.* [9] explored three social influence metrics, i.e., in-degree (number of followers), the number of retweets, and the number of mentions. Similarly, Kwak *et al.* [34] proposed three social influence metrics, i.e., PageRank, the number of followers, and the number of retweets. Since PageRank [34], [39], [52] and its extensions [37], [48], [50], [60] have been widely used in quantifying the social influence, in our work, we select the users within the top 0.1% PageRank values as *influentials*. We regard the rest of the users as *ordinary users*. We use a set P to represent the influentials. The average number of published tips and check-ins of these users are 82.14 and 2005.69, respectively. These values are much larger than those of ordinary users. Considering the gender composition, we can see 60.82% of users in P are male, 28.33% are female, and the rest 10.85% choose to hide their gender information. Therefore, the percentage of male users in P is significantly larger than that of the entire Foursquare. Regarding the percentage of enabling the cross-site linking function [18], the values of P and the whole Foursquare are 87.17% and 57.06%, respectively. This shows that the overwhelming majority of users in P have linked their profiles to Facebook and/or Twitter.

Although the aforementioned graph-based metrics are widely used to quantify the social influence in OSNs, they are more helpful for OSN service providers. From the perspective of third-party application providers, obtaining the social connectivity information of a selected OSN user might not be feasible. A number of mainstream OSNs, such as Facebook, allow a user to hide her list of friends/followings/followers. Therefore, a third-party application provider might not be able to quantify a user's social influence if social connectivity information is not fully available. Our goal is to find an approach to quickly determine whether a user is an influential within the OSN by referring to her publicly visible information, i.e., her profile and UGC. Our approach does not need to refer to the social connectivity information of the partial or entire graph.

A number of key features are selected to distinguish between the two groups of users. We classify them into two categories, i.e., content generation features and demographic features.

1) Content generation features (three features) are related to a user's content publishing behavior. We consider the number of tips and the number of visited countries. Also, we involve the number of check-ins on the Swarm app.
2) Demographic features (seven features) are related to the information fields of the user profile. There are some optional fields, including gender, lastname, profile photograph, home location, and biography. Therefore, we have five corresponding features, i.e., "has_gender," "$has\_lastname$," "$has\_profile\_photo$," "$has\_home\_location$," and "$has\_biography$." On the other hand, a Foursquare user can choose to link her profile to her Twitter and Facebook accounts. Accordingly, there are two more features, i.e., "$has\_Twitter$" and "$has\_Facebook$." If an optional field is enabled, the corresponding feature value is set as 1. Otherwise, the feature value is set as 0.

According to Table IX and Fig. 6, we can see the difference between influentials and ordinary users in terms of these features. To judge whether a user is an influential or an ordinary user based on her profile and content generation information, we introduce a binary classifier based on supervised machine learning technologies. In other words, we can simply refer to the profile page of a Foursquare user to accurately determine whether she is an influential, instead of relying on the structural information of the social network. Our approach can be adopted by third-party application providers to uncover influentials with a low measurement cost.

In our study, we randomly select 10 000 influentials and 10 000 ordinary users as the training and validation set. We compare the prediction performance between a number of supervised machine learning algorithms, including the emerging XGBoost [10] algorithm, which has been widely used in recent machine learning contests like Kaggle. We also study classic algorithms such as support vector machine (SVM) [22], CART decision tree (DT) [45], Random Forest (RF) [4], and Naive Bayers (NB) [31].

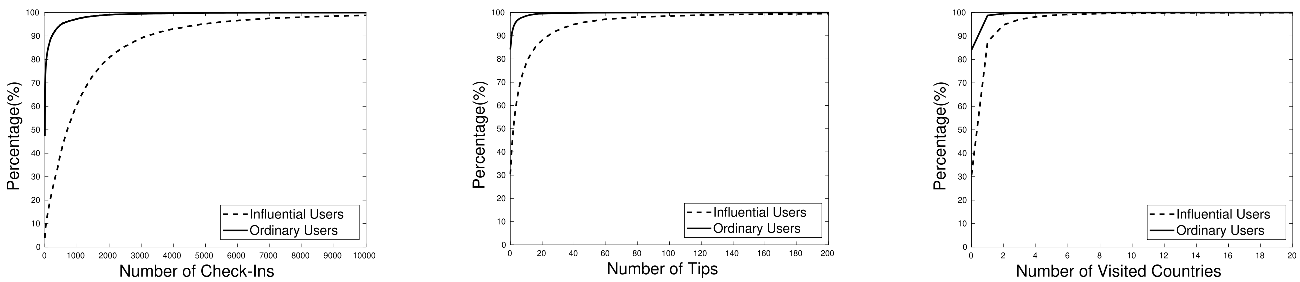
To evaluate the prediction performance of the classifier, we use four classic metrics, i.e., precision, recall, F1-score, and AUC [15]. For each algorithm, we have to select a set of parameters. Once the parameters are chosen, we apply tenfold cross-validation[3] for the evaluation. For each algorithm, we use a grid search to go through the parameter space and

[3]In tenfold cross-validation, we randomly divide the training and validation data set into ten subsets with equal size. Among these ten subsets, a single subset is retained as the validation data to evaluate the model, and the remaining nine subsets are applied for training. The cross-validation procedure is repeated ten times, with each of the ten subsets selected once as the validation data.

TABLE IX

COMPARISON BETWEEN INFLUENTIALS AND ORDINARY USERS ACCORDING TO OPTIONAL PROFILE FIELDS

	has_gender	*has_biography*	*has_profile_photo*	*has_Facebook*	*has_Twitter*	*has_home_location*
Influentials	96.15%	31.95%	98.75%	78.09%	70.39%	96.08%
Ordinary Users	92.55%	4.17%	67.75%	52.58%	14.91%	89.73%



(a) (b) (c)

Fig. 6. Comparison between Influentials and Ordinary users according to content posting. (a) Number of check-ins. (b) Number of tips. (c) Number of visited countries.

TABLE X

PREDICTION OF INFLUENTIALS

Model	Parameter	Precision	Recall	F1-Score	AUC
XGBoost	n_estimators=100, learning_rate=0.2, min_child_weight=7, max_depth=3, gamma=0.99, subsample=0.3, colsample_bytree=0.5, lambda=1, alpha =10^{-5}, objective=binary:logistic	0.85	0.89	0.87	0.88
RF	63 trees,depth=3,K(# of features)=5	0.83	0.88	0.85	0.87
SVMr	Kernel γ =1000, C =1.0	0.84	0.88	0.86	0.86
SVMp	Kernel degree d =3, C =1000	0.84	0.86	0.85	0.85
DT	Min samples leaf=20, Max depth=5	0.85	0.88	0.86	0.87
NB	-	0.81	0.84	0.82	0.83

record the set of parameters which can lead to the highest F1-score.

After the optimal parameter set of each model is obtained, we randomly select 5000 influentials and 5000 ordinary users as the test set. We evaluate the prediction performance of each algorithm using this test set. The parameters obtained during the training process are used, and the classifier predicts whether a user is an influential according to the selected features. Our results are shown in Table X. We can see that XGBoost performs the best, and we can achieve an F1-score of 0.87 and an AUC value of 0.88. The selected features can be used to accurately uncover the influentials for third-party application providers, without referring to the social connectivity information.

Among these selected features, we use χ^2 (Chi-square) statistic [66] to evaluate the discriminative power of each of them. According to Table XI, the three most discriminative features are "the number of checkins," "number of tips," and "has_Twitter," respectively.

TABLE XI

χ^2 STATISTIC

Rank	Feature	χ^2
1	Number of checkins	16046604.48
2	Number of tips	232337.46
3	Number of visited countries	8069.35
4	has_Twitter	5523.22
5	has_biography	3226.53
6	has_profile_photo	844.67
7	has_Facebook	750.57
8	has_location	34.13
9	has_lastname	18.69
10	has_gender	10.46

IV. RELATED WORK

Social graph analysis has been used for analyzing social networking services, such as Twitter, which also has a directed social graph. Kwak *et al.* [34] studied the complete Twitter social graph in 2009, including 41.7 million users. They analyzed the follower–following topology and found several features that differentiate Twitter from ordinary social networks. Watanabe *et al.* [59] collected a Twitter social graph with 469.9 million users and 28.7 billion relationships in 2012, and studied the graph from the aspects of degree distribution, reciprocity, degree of separation, and diameter. Gabielkov *et al.* [16] crawled about 93.77% of the complete Twitter social graph and studied its macrostructure. Leskovec *et al.* [36] studied the geospatial structure of a planetary-scale social network of 240 million users, i.e., the communication network of Microsoft Instant Messenger. They explored the interplay among topological, geographical, and algorithmically generated paths between the users. None of these works have studied the global social graph of Foursquare, the representative LBSN around the world.

There are also some works on tips in LBSNs. Li *et al.* [38] explored the venue's popularity on Foursquare by collecting data from 24 million venues in 14 different regions. For each venue, they collected the venue profile and statistical information, including the number of tips. They summarized three key factors related to venue popularity, i.e., the completeness of venue profile information, the category of the venue, and the age of the venue. Vasconcelos *et al.* [54] crawled 1.6 million Foursquare venues and extracted 527K user IDs from the obtained venue data. Based on the profiles of these users, they also studied the distribution of the number of tips per user. However, the set of studied user IDs were obtained in a biased way. Both studies only focused on the numbers of tips, instead of the content of each tip. Alrumayyan *et al.* [1] collected 12 000 tips from over 1000 venues in Riyadh, Saudi Arabia. They applied Lexicon-based sentiment analysis on Arabic tips and used Latent Dirichlet Allocation (LDA) algorithms to detect communities. Their investigation focused on a single city, paying particular attention to Arabic tips. Kwon *et al.* [35] used Foursquare tips and Yelp reviews to study user engagement in LBSNs. In particular, they focused on the long-term producers, who wrote more than 50 reviews, and examined the behavioral characteristics of these users. However, these users only covered 1.27% of Foursquare users, who are the most active ones in terms of tip publishing. Capdevila *et al.* [7] made use of the 309,640 Foursquare tips from the Manhattan region to verify the usefulness of their venue recommendation system, called GeoSRS, which was able to combine the advantages of text analysis and collaborative filtering. They did not provide discussions about tip publishing patterns. Costa *et al.* [12] studied the spam tips in Apontador, a popular Brazilian LBSN system. They proposed a spam detection mechanism, which was able to identify most of the spam tips. Vasconcelos *et al.* [53] explored the prediction of the future popularity of tips. In our work, we study all published tips on Foursquare from different aspects, including the numbers of tips of different users, as well as the venues, temporal patterns, and sentiment information of published tips.

V. Conclusion

In this article, we present a comprehensive analysis of Foursquare user behavior based on the crawled data of all 61.43 million Foursquare users. Our study covers two key building blocks of Foursquare, i.e., social connections and tips. We study the global Foursquare social graph and present a set of unique and undiscovered characteristics of this large graph, including a moderate level of reciprocity (0.42), a small average clustering coefficient (0.065), a giant strongly connected components (covering nearly 60% of users), and a significant community structure (Q value ≈ 0.6). Besides the singletons, almost all Foursquare users are weakly connected with each other. In addition, we conduct a detailed study on all published tips on Foursquare. On one hand, we analyze the numbers of tips published by different groups of users. On the other hand, we investigate the tips from the perspectives of tip venues, temporal patterns, and sentiment. Our analytical findings provide the first comprehensive view of Foursquare tips. Last but not least, as a practical scenario to help third-party application providers, we propose a supervised machine learning-based approach to predict influentials in LBSNs without referring to the social connectivity information. Our data-driven evaluation shows that our approach can reach a good prediction performance with an F1-score of 0.87 and an AUC value of 0.88. Our findings will be helpful for LBSN service providers, ISPs, and third-party application providers. For the next step, we wish to explore the following topics.

First, the massive Foursquare data tell us a lot about the user interactions and human mobility. Therefore, we aim to leverage the rich spatial and temporal information to improve urban planning. For example, for a selected city, we aim to collect user location data from Foursquare and involve data from other sources, such as mobile cellular data [13], [63] and human population data [51]. We will uncover the periodic phenomena and the long-term tendency of users according to the collected data. We believe that the user mobility data on Foursquare would be very useful for city computing-related applications [49], [62]. For example, our findings could help the governments decide where to build a new metro station, and could predict a possible traffic jam.

Second, there are some malicious accounts on Foursquare, and they might publish some incorrect tips to mislead legitimate users. As reported by Gong *et al.* [17], nearly 30% of all users on Dianping, another leading LBSN, are malicious accounts. Although Foursquare has applied some spam reporting and detection methodologies, still, spam tips keep appearing. We will further investigate malicious account detection problem using machine learning technologies [17], [20] and social graph-based technologies [6], [55]. We will consider both social interactions and spatiotemporal information to detect malicious users in an accurate way. In particular, since Foursquare records rich spatial and temporal information of users, we aim to introduce long short-term memory (LSTM) neural networks [24] for the detection.

Last but not least, we plan to consider other definitions of influential users, for example, considering other types of social interactions, instead of relying on the social graph only. In [19], we used the number of received "likes" to evaluate the social influence. Similarly, on Foursquare, users can upvote/downvote a tip to express their opinions. Therefore, a user who receives more upvotes could be an influential. We will further study how to uncover influential users according to the upvote/downvote information of the published tips.

References

[1] N. Alrumayyan, S. Bawazeer, R. AlJurayyad, and M. Al-Razgan, "Analyzing user behaviors: A study of tips in Foursquare," in *Proc. 5th Int. Symp. Data Mining Appl.*, 2018, pp. 153–168.

[2] S. Bird, "NLTK: The Natural Language Toolkit," in *Proc. COLING/ACL*, 2006, pp. 1–8.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.

[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[5] A. Broder *et al.*, "Graph structure in the Web," *Comput. Netw.*, vol. 33, nos. 1–6, pp. 309–320, Jun. 2000.

[6] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proc. NSDI*, 2012, pp. 197–210.

[7] J. Capdevila, M. Arias, and A. Arratia, "GeoSRS: A hybrid social recommender system for geolocated data," *Inf. Syst.*, vol. 57, pp. 111–128, Apr. 2016.

[8] E. Celikten, G. Le Falher, and M. Mathioudakis, "Modeling urban behavior by mining geotagged social data," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 220–233, Jun. 2017.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. AAAI ICWSM*, 2010, pp. 1–8.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[11] Y. Chen, J. Hu, H. Zhao, Y. Xiao, and P. Hui, "Measurement and analysis of the swarm social network with tens of millions of nodes," *IEEE Access*, vol. 6, pp. 4547–4559, 2018.

[12] H. Costa, F. Benevenuto, and L. H. C. Merschmann, "Detecting tip spam in location-based social networks," in *Proc. 28th Annu. ACM Symp. Appl. Comput. (SAC)*, 2013, pp. 724–729.

[13] P. Deville *et al.*, "Dynamic population mapping using mobile phone data," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 45, pp. 15888–15893, Nov. 2014.

[14] C. Ding, Y. Chen, and X. Fu, "Crowd crawling: Towards collaborative data collection for large-scale online social networks," in *Proc. 1st ACM Conf. Online Social Netw. (COSN)*, 2013, pp. 183–188.

[15] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[16] M. Gabielkov, A. Rao, and A. Legout, "Studying social networks at scale: Macroscopic anatomy of the Twitter social graph," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst. (SIGMETRICS)*, 2014, pp. 277–288.

[17] Q. Gong *et al.*, "DeepScan: Exploiting deep learning for malicious account detection in location-based social networks," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 21–27, Nov. 2018.

[18] Q. Gong, Y. Chen, J. Hu, Q. Cao, P. Hui, and X. Wang, "Understanding cross-site linking in online social networks," *ACM Trans. Web*, vol. 12, no. 4, pp. 25:1–25:29, Nov. 2018.

[19] Q. Gong *et al.*, "Exploring the power of social hub services," *World Wide Web*, vol. 22, no. 6, pp. 2825–2852, Nov. 2019.

[20] Q. Gong *et al.*, "Detecting malicious accounts in online developer communities using deep learning," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1251–1260.

[21] J. Han, D. Choi, B.-G. Chun, T. Kwon, H.-C. Kim, and Y. Choi, "Collecting, organizing, and sharing pins in pinterest: Interest-driven or social-driven?" in *Proc. ACM SIGMETRICS*, 2014, pp. 15–27.

[22] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[23] A. Hicks, S. Comp, J. Horovitz, M. Hovarter, M. Miki, and J. L. Bevan, "Why people use Yelp.com: An exploration of uses and gratifications," *Comput. Hum. Behav.*, vol. 28, no. 6, pp. 2274–2279, Nov. 2012.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[25] D. Hric, R. K. Darst, and S. Fortunato, "Community detection in networks: Structural communities versus ground truth," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, no. 6, Dec. 2014, Art. no. 062805.

[26] T. Hu, E. Bigelow, J. Luo, and H. Kautz, "Tales of two cities: Using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 55–66, Mar. 2017.

[27] Y. Huang, Y. Chen, Q. Zhou, J. Zhao, and X. Wang, "Where are we visiting? Measurement and analysis of venues in Dianping," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[28] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. AAAI ICWSM*, 2014, pp. 1–10.

[29] L. Jiao, J. Li, T. Xu, W. Du, and X. Fu, "Optimizing cost for online social networks on geo-distributed clouds," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 99–112, Feb. 2016.

[30] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos, "Understanding user behavior in online social networks: A survey," *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 144–150, Sep. 2013.

[31] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. UAI*, 1995, pp. 338–345.

[32] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 179–188.

[33] H. Kwak, Y. Choi, Y.-H. Eom, H. Jeong, and S. Moon, "Mining communities in networks: A solution for consistency and its evaluation," in *Proc. 9th ACM SIGCOMM Internet Meas. Conf. (IMC)*, 2009, pp. 301–314.

[34] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 591–600.

[35] Y. D. Kwon, D. Chatzopoulos, E. ul Haq, R. C.-W. Wong, and P. Hui, "GeoLifecycle: User engagement of geographical exploration and churn prediction in LBSNs," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 92:1–92:29, Sep. 2019.

[36] J. Leskovec and E. Horvitz, "Geospatial structure of a planetary-scale social network," *IEEE Trans. Comput. Soc. Syst.*, vol. 1, no. 3, pp. 156–163, Sep. 2014.

[37] J. Li, W. Peng, T. Li, T. Sun, Q. Li, and J. Xu, "Social network user influence sense-making and dynamics prediction," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5115–5124, Sep. 2014.

[38] Y. Li, M. Steiner, L. Wang, Z.-L. Zhang, and J. Bao, "Exploring venue popularity in Foursquare," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 3357–3362.

[39] Q. Liu *et al.*, "An influence propagation view of PageRank," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 3, pp. 30:1–30:30, Apr. 2017.

[40] M. Moricz, Y. Dosbayev, and M. Berlyant, "PYMK: Friend recommendation at myspace," in *Proc. 2010 Int. Conf. Manage. Data (SIGMOD)*, 2010, pp. 999–1002.

[41] M. E. J. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 66, no. 3, Sep. 2002, Art. no. 035101.

[42] A. Noulas, B. Shaw, R. Lambiotte, and C. Mascolo, "Topological properties and temporal dynamics of place networks in urban environments," in *Proc. 24th Int. Conf. World Wide Web (WWW Companion)*, 2015, pp. 431–441.

[43] D. Preoţiuc-Pietro and T. Cohn, "Mining user behaviours: A study of check-in patterns in location based social networks," in *Proc. 5th Annu. ACM Web Sci. Conf. (WebSci)*, 2013, pp. 306–315.

[44] J. M. Pujol, V. Erramilli, and P. Rodriguez, "Divide and conquer: Partitioning online social networks," *CoRR*, vol. abs/0905.4918, 2009. [Online]. Available: https://arxiv.org/abs/0905.4918

[45] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993, pp. 306–315.

[46] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora, "Distance matters: Geo-social metrics for online social networks," in *Proc. 3rd Workshop Online Social Netw. (WOSN)*, 2010, pp. 1–9.

[47] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *Proc. AAAI ICWSM*, 2011, pp. 1–8.

[48] A. Silva, S. Guimarães, W. Meira, and M. Zaki, "ProfileRank: Finding relevant content and influential users based on information diffusion," in *Proc. 7th Workshop Social Netw. Mining Anal. (SNAKDD)*, 2013, pp. 1–9.

[49] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro, "Revealing the city that we cannot see," *ACM Trans. Internet Technol.*, vol. 14, no. 4, pp. 26:1–26:23, Dec. 2014.

[50] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere," in *Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2007, pp. 971–974.

[51] F. R. Stevens, A. E. Gaughan, C. Linard, and A. J. Tatem, "Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data," *PLoS ONE*, vol. 10, no. 2, Feb. 2015, Art. no. e0107042.

[52] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2012, pp. 743–752.

[53] M. Vasconcelos, J. M. Almeida, and M. A. Gonçalves, "Predicting the popularity of micro-reviews: A Foursquare case study," *Inf. Sci.*, vol. 325, pp. 355–374, Dec. 2015.

[54] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida, "Tips, dones and todos: Uncovering user profiles in Foursquare," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2012, pp. 653–662.

[55] B. Wang, J. Jia, L. Zhang, and N. Z. Gong, "Structure-based Sybil detection in social networks via local rule-based propagation," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 3, pp. 523–537, Jul. 2019.

[56] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: An analysis of quora," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, 2013, pp. 1341–1352.

[57] G. Wang, S. Y. Schoenebeck, H. Zheng, and B. Y. Zhao, "'Will check-in for badges': Understanding bias and misbehavior on location-based social networks," in *Proc. AAAI ICWSM*, 2016, pp. 1–10.

[58] T. Wang *et al.*, "The power of comments: Fostering social interactions in microblog networks," *Frontiers Comput. Sci.*, vol. 10, no. 5, pp. 889–907, Oct. 2016.

[59] M. Watanabe and T. Suzumura, "How social network is evolving?: A preliminary study on billion-scale Twitter network," in *Proc. 22nd Int. Conf. World Wide Web (WWW Companion)*, 2013, pp. 531–534.

[60] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential twitterers," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining (WSDM)*, 2010, pp. 261–270.

[61] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. 4th ACM Eur. Conf. Comput. Syst. (EuroSys)*, 2009, pp. 205–218.

[62] R. Xie, Y. Chen, Q. Xie, Y. Xiao, and X. Wang, "We know your preferences in new cities: Mining and modeling the behavior of travelers," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 28–35, Nov. 2018.

[63] F. Xu, P. Zhang, and Y. Li, "Context-aware real-time population estimation for metropolis," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput.*, Sep. 2016, pp. 1064–1075.

[64] T. Xu, Y. Chen, L. Jiao, B. Y. Zhao, P. Hui, and X. Fu, "Scaling microblogging services with divergent traffic demands," in *Proc. ACM/IFIP/USENIX Middleware*, 2011, pp. 20–40.

[65] C. Yang *et al.*, "Understanding the behavioral differences between American and German users: A data-driven study," *Big Data Mining Analytics*, vol. 1, no. 4, pp. 284–296, 2018.

[66] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. ICML*, 1997, p. 35.

[67] Z. Zhang *et al.*, "On the validity of geosocial mobility traces," in *Proc. 12th ACM Workshop Hot Topics Netw. (HotNets-XII)*, 2013, pp. 1–7.

[68] X. Zhao *et al.*, "Multi-scale dynamics in a massive online social network," in *Proc. ACM Conf. Internet Meas. Conf. (IMC)*, 2012, pp. 171–184.

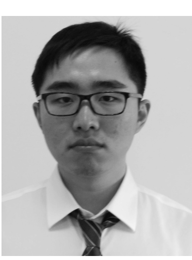
[69] J. Zhou and J. Fan, "JPR: Exploring joint partitioning and replication for traffic minimization in online social networks," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 1147–1156.

[70] Q. Zhou *et al.*, "Measurement and analysis of the reviews in Airbnb," in *Proc. IFIP Netw. Conf. (IFIP Netw.)*, May 2018, pp. 1–9.

Yang Chen (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2004 and 2009, respectively.

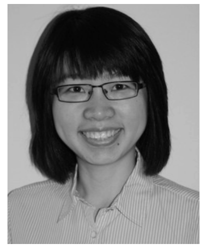
He is currently an Associate Professor with the School of Computer Science, Fudan University, Shanghai, China, where he also leads the Mobile Systems and Networking (MSN) Group. From April 2011 to September 2014, he was a Post-Doctoral Associate with the Department of Computer Science, Duke University, Durham, NC, USA, where he served as a Senior Personnel with the NSF MobilityFirst Project. From September 2009 to April 2011, he has been a Research Associate and the Deputy Head of the Computer Networks Group, Institute of Computer Science, University of Göttingen, Göttingen, Germany. He visited Stanford University, Stanford, CA, USA, in 2007, and Microsoft Research Asia, Beijing, from 2006 to 2008 as a Visiting Student. He was a Nokia Visiting Professor with Aalto University, Espoo, Finland, in 2019. His research interests include online social networks, Internet architecture, and mobile computing.

Dr. Chen served as an OC/TPC Member for many international conferences, including SOSP, WWW, IJCAI, AAAI, ECAI, DASFAA, IWQoS, ICCCN, GLOBECOM, and ICC. He serves as an Associate Editor-in-Chief of the *Journal of Social Computing*, an Associate Editor of IEEE ACCESS, and an Editorial Board Member of the Transactions on Emerging Telecommunications Technologies (ETT).



Jiyao Hu received the B.S. degree from the School of Computer Science, Fudan University, Shanghai, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Duke University, Durham, NC, USA.

He was a Student Research Assistant with the Mobile Systems and Networking (MSN) Group, Fudan University, from 2014 to 2017. He visited Aalto University, Espoo, Finland, as a Research Intern in 2017. His research interests include online social networks and data mining. He published refereed articles in USENIX NSDI, ACM Transactions on the Web, and the IEEE ICDCS.



Yu Xiao (Member, IEEE) received the bachelor's and master's degrees in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, and the Ph.D. degree (Hons.) in computer science from Aalto University, Espoo, Finland, in January 2012.

She is currently an Assistant Professor with the Department of Communications and Networking, Aalto University, where she leads the Mobile Cloud Computing Group. Her research interests include edge computing, mobile crowdsensing, and energy-efficient wireless networking.

Dr. Xiao has received three best paper awards from IEEE/ACM Conferences. She is also a recipient of the three-year postdoc grant from the Academy of Finland.

Xiang Li (Senior Member, IEEE) received the B.S. and Ph.D. degrees in control theory and control engineering from Nankai University, Tianjin, China, in 1997 and 2002, respectively.

In 2008, he joined Fudan University, Shanghai, China, as a Professor of the Electronic Engineering Department. He was with the City University of Hong Kong, Hong Kong, Int. University Bremen, Bremen, Germany, and Shanghai Jiao Tong University, Shanghai, China, as a Post-Doctoral Research Fellow, Humboldt Research Fellow, and an Associate Professor from 2002 to 2004, 2005 to 2006, and 2004 to 2007, respectively. He served as the Head of the Electronic Engineering Department, Fudan University, from 2010 to 2015. He is currently a Distinguished Professor of Fudan University and chairs the Adaptive Networks and Control (CAN) Group and the Research Center of Smart Networks and Systems, School of Information Science and Engineering, Fudan University. He has (co-)authored 4 research monographs, 6 academic chapters, and more than 200 peer-refereed publications in journals and conferences. His main research interests cover network science and system control in both theory and applications.

Dr. Li received the IEEE Guillemin-Cauer Best Transactions Paper Award from the IEEE Circuits and Systems Society in 2005, Shanghai Natural Science Award (1st class) in 2008, the Shanghai Science and Technology Young Talents Award in 2010, the National Science Foundation for Distinguished Young Scholar of China in 2014, the National Natural Science Award of China (2nd class) in 2015, Ten Thousand Talent Program of China in 2017, among other awards and honors. He served as an Associate Editor for several journals including the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: Regular Papers from 2010 to 2015, and serves as the Associate Editor of *Journal of Complex Networks* and the IEEE Circuits and Systems Society Newsletter, and the Guest Editor of the IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING.

Pan Hui (Fellow, IEEE) received the B.Eng. and M.Phil. degrees from the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, and the Ph.D. degree from Computer Laboratory, University of Cambridge, Cambridge, U.K.

He has been an Adjunct Professor of social computing and networking with Aalto University, Espoo, Finland, since 2012. He has been a Faculty Member of the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, since 2013. He has been the Nokia Chair in data science and a Full Professor in computer science with the University of Helsinki, Helsinki, Finland, since September 2017. He has authored or coauthored more than 200 research articles with over 18 000 citations. He has around 30 granted/filed European patents.

Dr. Hui is an Associate Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING and the IEEE TRANSACTIONS ON CLOUD COMPUTING, a Guest Editor for *IEEE Communication Magazine*, an ACM Distinguished Scientist, and a member of Academia Europaea (Academy of Europe).