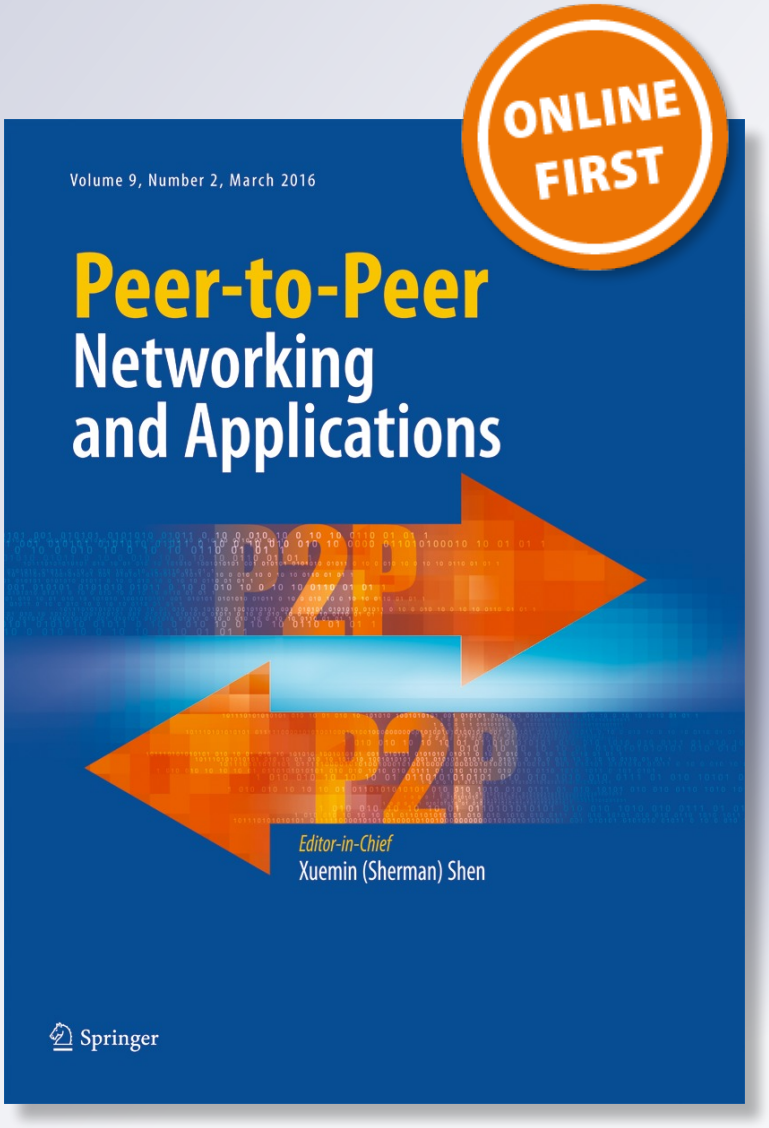
Characterizing user behaviors in location-based find-and-flirt services: Anonymity and demographics

Minhui Xue, Limin Yang, Keith W. Ross & Haifeng Qian

Peer-to-Peer Networking and Applications

ISSN 1936-6442

Peer-to-Peer Netw. Appl.
DOI 10.1007/s12083-016-0444-5



Springer

Your article is protected by copyright and all rights are held exclusively by Springer Science +Business Media New York. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

Author's personal copy

Peer-to-Peer Netw. Appl.
DOI 10.1007/s12083-016-0444-5

Characterizing user behaviors in location-based find-and-flirt services: Anonymity and demographics

A WeChat Case Study

Minhui Xue[1,2] · Limin Yang[1] · Keith W. Ross[2,3] · Haifeng Qian[1]

Received: 30 September 2015 / Accepted: 10 February 2016
© Springer Science+Business Media New York 2016

Abstract WeChat, both a location-based social network (LBSN) and an online social network (OSN), is an immensely popular application in China. In this paper we specifically focus on a popular WeChat sub-service, namely, the *People Nearby* service, which is exemplary of a *find-and-flirt* service, similar to those on Momo and Tinder. Specifically, the *People Nearby* service reads in the current geographic location of the device to locate a list of other people using WeChat who are in the same vicinity. The user can then request to establish a WeChat friendship relation with any of the users on the list. In this paper, we explore: (i) if one gender tends to use the *People Nearby* service more than another; (ii) if users of *People Nearby* are more anonymous than ordinary WeChat users; (iii) if ordinary WeChat users are more anonymous than Twitter users. We also take an in-depth examination of the user anonymity and demographics in a combined fashion and examine: (iv) if ordinary WeChat females are more anonymous than ordinary males; (v) if *People Nearby* females are more anonymous than *People Nearby* males. By answering these questions, we will gain significant insights into modern online dating and friendship creation, insights that should be able to inform sociologists as well as designers of future *find-and-flirt* services.

Keywords Location-based social networks · Anonymity · Demographics · Find-and-flirt services

1 Introduction

The wide proliferation of both smartphones and ubiquitous location-based services (LBSs) has driven the exponential growth of location-based social networks (LBSNs). LBSNs, a subcategory of LBSs, are designed to enable people to discover nearby users and establish on-the-spot communication. Currently, there is a plethora of popular LBSN applications – applications that help recommend nearby restaurants (e.g., Dianping, Yelp [12]); applications that create anonymous platforms for users (e.g., Yik Yak [14], Whisper [21]); and applications that help find potential nearby candidates for dating, often termed *find-and-flirt services* (e.g., Tinder [2], and the *People Nearby* service of WeChat).

In this paper we focus our study on WeChat. WeChat, which is both a LBSN and an OSN application, and which boasts more than 600 million users globally. Unlike Facebook and Google+, WeChat does not enforce a real-name policy, so that users may use pseudonyms during registration. WeChat's LBSN services – such as *"Shake", "Drift Bottles", "Circle of Friends"*, and *"People Nearby"* – facilitate users in creating friendships with nearby users. In this paper we specifically focus on the *People Nearby* service, which is an exemplary *find-and-flirt* service, similar

✉ Minhui Xue
minhuixue@nyu.edu

✉ Haifeng Qian
hfqian@cs.ecnu.edu.cn

Limin Yang
whyisyoung@ecnu.cn

Keith W. Ross
keithwross@nyu.edu

[1] East China Normal University, Shanghai, 200241, China

[2] NYU Shanghai, Shanghai, 200122, China

[3] New York University, New York, NY, 11201, USA

Published online: 26 February 2016

Author's personal copy

to the *find-and-flirt* services such as Tinder and Momo. Specifically, the *People Nearby* service reads in the current geographic location of the device to locate a list of other people using *People Nearby* who are in the same vicinity. The user can then request to establish a WeChat friendship relation with any of the users on the list. When reporting the nearby users, WeChat reports how far away each user is in bands of 100 meters.

In this paper, we explore the following questions about *find-and-flirt* services: (i) does one gender tend to use the *People Nearby* service more than another; (ii) if users of *People Nearby* are more anonymous than ordinary WeChat users; (iii) if ordinary WeChat users are more anonymous than Twitter users. We also take an in-depth examination of the user anonymity and demographics in a combined fashion and examine: (iv) if ordinary WeChat females are more anonymous than ordinary males; (v) if *People Nearby* females are more anonymous than *People Nearby* males. By answering these questions, we will gain significant insights into modern online dating and friendship creation, insights that should be able to inform sociologists and psychologists, as well as designers of future find-and-flirt services.

This work aims to show the feasibility of exploring anonymity and demographics for characterizing user behavior in location-based find-and-flirt applications. Leveraging the methodology of [6] for data collection, for a period of 7 consecutive days, we collect in total 8,462 screenshots and 41,874 WeChat entries corresponding to 3,215 distinct Latin-Chinese character names. Following the methodology of [16], and with the help of 5 labmates, we classify these WeChat users into 4 categories: Anonymous, Identifiable, Partially Anonymous, and Unclassifiable.

Our findings are summarized in the following:

- For ordinary WeChat users, the fraction of male users and female users appears to be roughly the same. However, when it comes to using the *People Nearby* service, male users outnumber female users by roughly four to one.
- For ordinary WeChat users, anonymous users slightly outnumber identifiable users. However, when it comes to using the *People Nearby* service, anonymous users are twice as many as identifiable users.
- For ordinary WeChat users, the fraction of males who are identifiable is almost twice as many as the fraction of females who are identifiable, while the fraction of females who are anonymous greatly outnumbers the fraction of males who are anonymous. However, when it comes to using the *People Nearby* service, the fraction of identifiable males decreases significantly, and the fraction of anonymous males increases significantly.
- WeChat users are more anonymous than Twitter users in terms of the fraction of anonymous users.
- We finally build a machine learning classifier that can be used to detect anonymous and identifiable WeChat accounts by using the *People Nearby* query data sets and users' demographics information. We, therefore, are able to re-define the concept of anonymity in a novel fashion.

The rest of the paper is organized as follows. Section 2 briefly discusses the background of the WeChat application. Section 3 presents the framework of data collection and data sets. In Section 4, we preliminarily compare ordinary WeChat users to *People Nearby* users. In Section 5, we take an in-depth look at anonymity and gender, and persistence in a combined fashion. In Section 6, we revisit the anonymity of WeChat users by using machine learning. Section 7 surveys related work. Finally, Section 8 concludes the paper.

2 Background

We limit our study to one of the most prevalent LBSN applications – WeChat. Since its inception in 2011, it has become the de facto social network and messaging system in China, and has also become popular in many other countries. In this paper, we first describe the *People Nearby* service and then take a look at WeChat's real-name policy for usernames.

2.1 *People Nearby* service

The WeChat application provides a *People Nearby* service, which takes as input the current geo-location of a user's mobile device and returns a list of WeChat users in close proximity. Users can then send requests to people listed in hope of establishing WeChat friendship relationships with them. Once a friendship relationship is established between two users, the users can message each other and see each other's social media postings. For each user listed in *People Nearby*, WeChat gives an indication of how close the user is. Using bands of 100 meters, WeChat reports that a user is within 100 meters, within 200 meters, within 300 meters, and so on. For example, if Alice is 368 meters away from Bob, the WeChat server will only report that she is between 300 and 400 meters away from Bob. We emphasize that if a WeChat user does not use the WeChat *People Nearby* service, then the user is not traceable by the methods described in this paper. User mobility is only traceable if a user repeatedly queries WeChat's *People Nearby* service (Fig. 1).

As just described, WeChat lists the nearby people starting from closest to furthest within different distance ranges. Figure 2 shows a sample image of the *People Nearby* functionality of WeChat. The granularity of distances WeChat

Author's personal copy

Fig. 1 Screenshot of the WeChat's *People Nearby* functionality

Fig. 2 Fake GPS

reports is non-linear; distances up to 1,000 meters are reported in bands of 100 meters, but beyond 1,000 meters, the band size increases to 1,000 meters increments. Additionally, refreshing this list is susceptible to many factors such as synchronization issues. Therefore, we must take into account the fact that it is not always possible to discover a user even if he or she is near.

2.2 Classifying WeChat users

We rely on human knowledge to classify WeChat user accounts as Anonymous and Identifiable. We leverage 5 university labmates to help label these accounts (See Fig. 2). Specifically, following the methodology of [16], the labmates are asked to decide whether these usernames contain the following:

- just a first name;
- just a last name;
- both a first name and a last name;
- neither a first nor a last name;
- not sure or other.

We require the labmates to only choose the "neither a first or a last name" or "both a first name and a last name" options if they have full confidence, so as to avoid mislabeling particularly ambiguous user accounts. To take into account human error, each account is labelled by three labmates, and in the case of disagreement, a majority vote is used to decide. If, however, decision still cannot be reached, we (the authors) provide a definitive, final label for the account in question. By exploiting these labels, we define each WeChat user as one of the following:

- Identifiable: A WeChat account containing both a first name and a last name;
- Anonymous: A WeChat account containing neither a first nor a last name;
- Partially Anonymous: A WeChat account containing either a first name or a last name, but not both;
- Semi-Anonymous: A WeChat account that is either Partially Anonymous or Anonymous;
- Unclassifiable: A WeChat account that is neither Anonymous, Identifiable nor Partially Anonymous. For

Author's personal copy

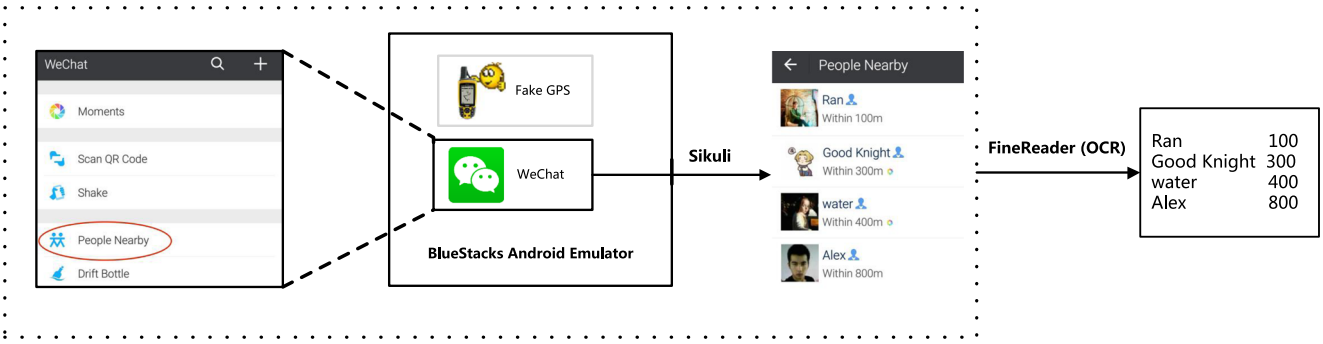


Fig. 3 Data Collection Framework: The Android emulator runs WeChat and Fake GPS; Sikuli interacts with the applications to set the fake GPS location, get lists of nearby people, and take screenshots of WeChat; FineReader then extracts the usernames and reported distances for each screenshot. Hence, tracking users can be done with all off-the-shelf softwares, rather than highly complex tools

example, a WeChat account that belongs to a name of a company or an organization.

To be sure, WeChat does not support complete anonymity (i.e., accounts that are never associated with any pseudonym). However, in this paper, we use the commonly-employed term *anonymous* in lieu of the term *pseudonymous* that appears frequently in cryptography. We regard the above two terms to be equivalent.

3 Methodology

3.1 Data collection

This paper aims to create a system that places two virtual probes in any location in the world and uses these probes to collect the names of nearby WeChat users and their distance with respect to the probe. Following the methodology of [6], we use an Android emulator, *BlueStacks Android Emulator*,[1] to simulate multiple Android devices. Within each emulator we run the WeChat app and another application, named *Fake GPS*,[2] to set the probe's GPS location to any place in the world by inputting the latitude and longitude of corresponding to the desired location (See Fig. 2). The WeChat application subsequently perceives this spoofed location to be the actual location of the user. We then use the commercial optical character recognition (OCR) software, *ABBYY FineReader*[3] to automate the data collection process and obtain the desktop images; we then process all the collected images through an optical character recognition (OCR) tool to extract usernames and reported distances for each WeChat screenshot corresponding to each probe (See Fig. 3). See [6] for a more detailed description of the data collection methodology.

We conducted our real-world experiment on Wall Street, downtown Manhattan of New York City. We then probed on Wall Street with 2 probes, each separated by 200 meters, and recorded the *People Nearby* query at both probe locations. Both probe locations took anywhere from 30 seconds to 2 minutes to scan for nearby people. Both probes are set to alternatively scan every 30 minutes in order to reduce traffic to the WeChat servers. We then used OCR to extract the usernames and reported distances for each WeChat screenshot. After obtaining WeChat usernames, we perform a large-scale analysis of these users. Spam users are filtered out, and relying on human knowledge all the users are subsequently classified as either Anonymous, Identifiable, Partially Anonymous, or Unclassifiable based on the criteria described in Section 2.2.

We collected data from 19 January 2015 to 25 January 2015. After the data collection, we used the OCR tool, *ABBYY FineReader*, to extract the textual data as a tuple in the form of <username, distance, timestamp, probe coordinate, gender>. For this Wall Street experiment, most of the characters in the names are latin characters, which can be automatically determined by OCR. For recognizing Chinese characters, we rely on human knowledge to intervene to improve the overall accuracy.

3.2 Data sets

We used one machine to collect data from two virtual probes set on Wall Street for 7 consecutive days. In total, 8,462 screenshots, 41,874 entries as tuples in the form of <username, distance, timestamp, probe coordinate,

[1] http://www.bluestacks.com

[2] https://play.google.com/store/apps/details?id=com.lexa.fakegps

[3] http://finereader.abbyy.com

Author's personal copy

gender> corresponding to 3,215 distinct Latin-Chinese character names were collected. With the help of 5 labmates, these users were split into 4 groups, corresponding to Anonymous, Identifiable, Partially Anonymous, and Unclassifiable users. The partitioning of users was done as described in Section 2.2. We also take into account the self-reported gender as displayed on the *People Nearby* lists with respect to each username. The number of distinct users that appear each day is displayed in Fig. 4. We see that the number of users that appear on any given day fluctuates around 800, which guarantees us to have enough distinct user accounts in our data sets, rather than very few users with high query frequencies.

In order to make a comparison with WeChat regular users who never trigger the *People Nearby* functionality of WeChat, we ask 5 female volunteers and 5 male volunteers to collect the usernames in their contact lists as our WeChat reference group. Specifically, in total, 1,136 distinct usernames were collected and classified into 4 categories by using the methodology described in Section 2.2.

We summarize our data sets as follows:

- WeChat Regular Users (WeChat Reference Group), who are the friends of our ten test volunteers: 1,136 users
- WeChat *People Nearby* Users: Users who appear at least once in any one day of 7 consecutive days during our experiment: 3,215 users

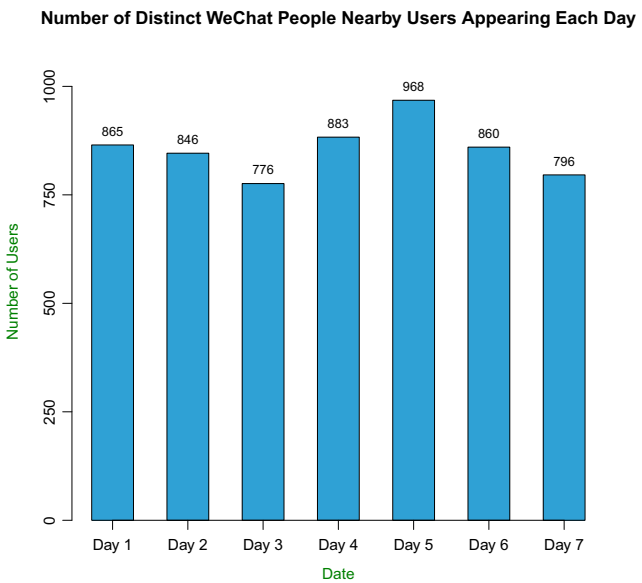


Fig. 4 Number of distinct WeChat *People Nearby* users appearing each day

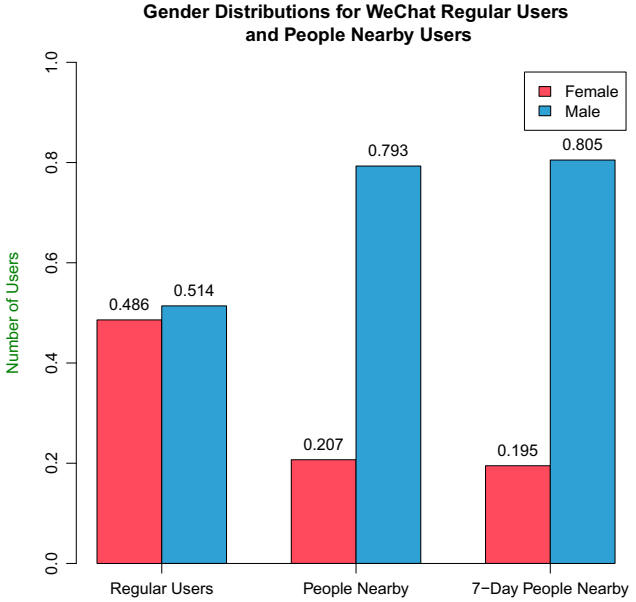


Fig. 5 Gender distributions for WeChat regular users and *People Nearby* users

- WeChat 7-Day *People Nearby* Users: Users who appear in all 7 consecutive days during the experiment: 82 users
- Twitter Users: Data sets adapted from [16] (50,173 accounts)

4 Comparing ordinary users to *People Nearby* users

In this section we compare the the gender and anonymity of ordinary WeChat users to *People Nearby* users.

4.1 Gender analysis

We first begin by comparing the gender distributions for ordinary WeChat users and *People Nearby*. As we see from Fig. 5, for regular WeChat users, the fraction of male users and female users appears to be roughly the same. However, when it comes to using the *People Nearby* service, male users outnumber female users by roughly four to one. This bias towards male users appears for users who appear at least once in the 7 consecutive days, and for users who appear in each of the 7 consecutive days.

From this analysis, we can see that males are more active than females in *People Nearby*. This could possibly be because men are less privacy concerned; or it could simply be that men are more aggressive when trying to establish new friends and mates. Further study is needed by psychologists and sociologists to understand on a deeper

Author's personal copy

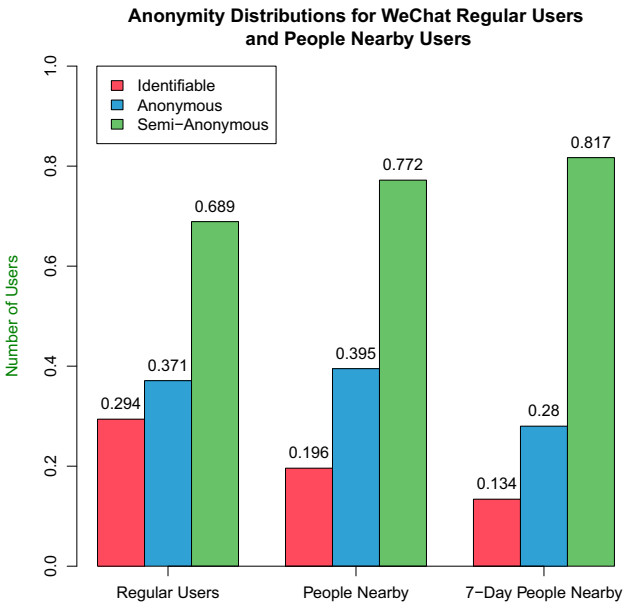


Fig. 6 Anonymity distributions for WeChat regular users and *People Nearby* users

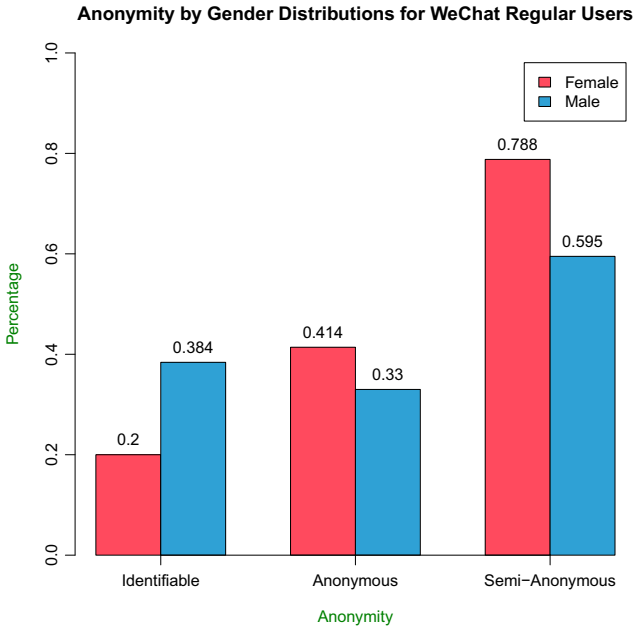


Fig. 7 Anonymity by gender distributions for WeChat regular users

level the incentives and dis-incentives for using the *People Nearby* service, and why males are more prevalent in *People Nearby*.

4.2 Anonymity analysis

We now compare the anonymity distributions for ordinary WeChat users with those for *People Nearby*. As we observe from Fig. 6, for regular WeChat users, anonymous users slightly outnumber identifiable users. However, when it comes to using the *People Nearby* service, anonymous users are twice as many as identifiable users, and semi-anonymous users outnumber identifiable users by roughly five to two. This bias towards anonymous users in *People Nearby* appears for users who appeared at least once, and for users who appeared in each of the 7 consecutive days.

From this analysis, we can see that anonymous WeChat users are more active in *People Nearby*. This could possibly be because users maybe prefer not to expose their identities in *find-and-flirt* services such as *People Nearby* service. Since the identity of anonymous users is not clear, anonymous users may be more comfortable in using the service. Furthermore, some users may choose to be anonymous for the explicit purpose of using the *People Nearby* service. Further study is needed by psychologists and sociologists to understand on a deeper level the incentives and dis-incentives for using the *People Nearby* service, and why anonymous users are more prevalent in *People Nearby*.

5 Anonymity by gender analysis

In this section we take a deeper dive into our data sets, examining anonymity and gender in a combined fashion.

We first begin by comparing the anonymity by gender distributions for ordinary WeChat users and *People Nearby*. As we see from Fig. 7, out of all the 1,136 accounts, the fraction of males who are identifiable is almost twice as many as the fraction of females who are identifiable, while the fraction of females who are semi-anonymous greatly outnumbers the fraction of males who are semi-anonymous. Thus we see that, for regular WeChat users, females choose anonymous names more frequently than males. This is perhaps because females are more privacy concerned, or because females enjoy using unusual, humorous names. More research, perhaps using traditional surveys, is needed to understand the motivations choosing anonymous names.

However, when it comes to using the *People Nearby* service (see Fig. 8), out of all the 3,215 distinct user accounts, we see that the fraction of identifiable males decreases significantly, and the fraction of anonymous males increases significantly. This shows that males are much more comfortable using the *People Nearby* service with anonymous names. The fraction of female users also decreases, but not as significantly. This is likely because a relatively small fraction of regular females are identifiable, so there is less room to further decrease this percentage. By exploiting anonymity as camouflage, male users are more likely to be proactive when seeking potential nearby candidates.

Author's personal copy

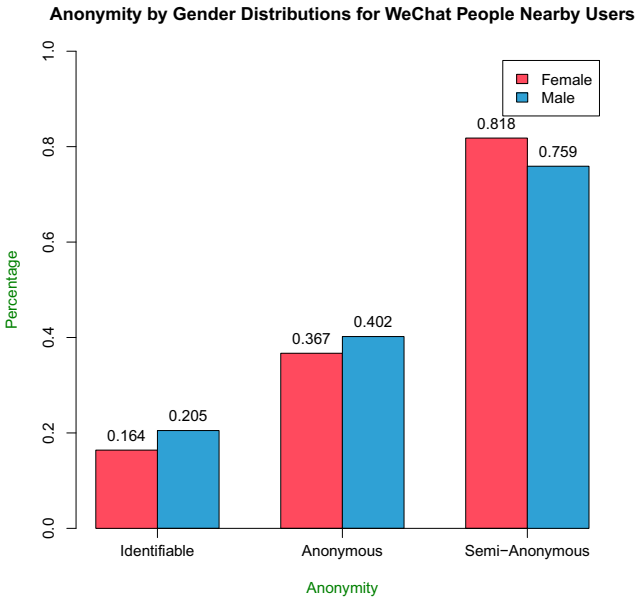


Fig. 8 Anonymity by gender distributions for WeChat *People Nearby* users

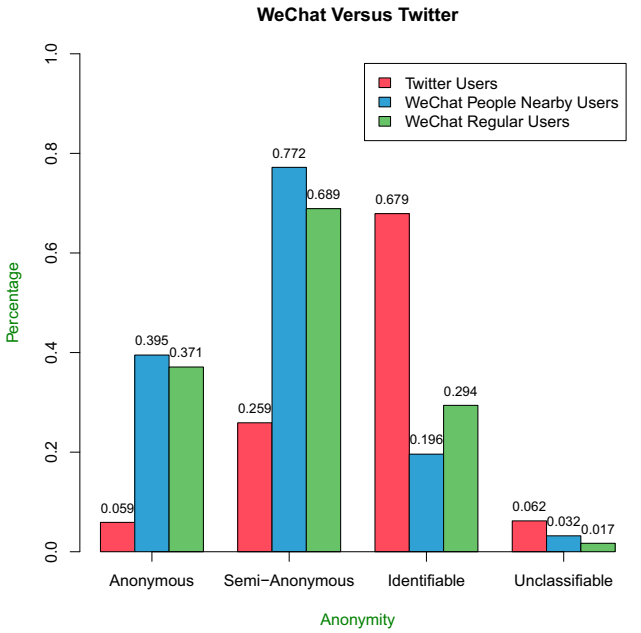


Fig. 9 Anonymous and identifiable accounts of WeChat versus Twitter

5.1 Anonymity of WeChat regular users versus Twitter users

From the data set of [16], all the 50,173 accounts are labelled by using Amazon Mechanical Turk. Figure 9 plots each anonymity percentages for WeChat and Twitter. As anonymity is a crucial feature for OSNs, we see that with at least 37.1 % (resp. 5.9 %) of WeChat regular users (resp. Twitter users) using non-identifiable pseudonyms. Thus, we can deduce that WeChat is more anonymous than Twitter in terms of the fraction of anonymous users. Furthermore, the identifiable user group has 67.9 % of Twitter accounts, while it only takes up approximately 30 % for WeChat regular accounts. As the fraction of each anonymity category shown in Fig. 9, the composition of WeChat regular users are quite different from that of Twitter.

This is partially because WeChat users can send and receive messages in a closed circle with familiar friends. People in a closed circle are able to recognize each other even with anonymous usernames since the WeChat system binds the user mobile phone or the ordinary telephone that can be uniquely identified by friends. Furthermore, WeChat users often accept friendship (such as with the *People Nearby* functionality) without really knowing the others. They attempt to exchange messages with the discovered nearby users, thereby attempting to make new friends and possibly meeting up with them in an anonymous manner. This indicates that user anonymity plays a significant role in WeChat. Giving users a sense of security when querying *People Nearby* by not having a strict real-name policy is a selling point for WeChat. In addition, WeChat has only developed a set of APIs to build custom features for verified businesses or public organizations. (These accounts are classified into Unclassifiable.). WeChat has not yet developed private verified accounts.

In comparison, Twitter has been proactive in creating a more manageable experience for its high profile members through private verified accounts. They are highly sought members in music, journalism, and sports, who are absolutely classified into identifiable Twitter users. As is to be expected, these Twitter users are more willing to advertise themselves to win new followers.

6 Automatically discovering anonymous WeChat nearby users

In this section, we take an in-depth dive into our *People Nearby* data sets, arguing that the total number of queries has something to do with demographics. By using machine learning, we attempt to automatically discover anonymous WeChat users by combining the *People Nearby* service and gender information, and to re-define the concept of anonymity in a novel fashion.

Author's personal copy

Table 1 Labelled data for anonymity versus number of queries of WeChat 7-Day *People Nearby* Users

Label	# of Queries < 50		50 ⩽ # of Queries < 100		100 ⩽ # of Queries < 150		150 ⩽ # of Queries < 200		200 ⩽ # of Queries	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Anonymous	234	984	7	22	3	8	2	6	1	3
Semi-Anonymous	517	1,822	23	67	5	18	3	14	2	10
Identifiable	98	497	4	20	1	5	1	3	0	2
Unclassifiable	11	88	0	0	0	2	0	1	0	1
Total	3,033		114		31		22		15	

6.1 Feature selection for classification

In order to select the contributed features to support the classification results with sufficient accuracy, we further explore the relationship among the data sets over each category we obtained in Table 1.

Figures 10 and 11 show that there exits a strong relationship among user anonymity, gender, and the total number of queries in a period of 7 consecutive days. Since a lot of data over-plotting can sometimes obscure important patterns, in this situation, it can be useful at the exploratory data analysis phase to "jitter" the data so that underlying data can be viewed making it easier to see patterns. Jittering means adding additional noise to your data as a way of "anonymizing" spatial data while maintaining, for example, neighborhood patterns. When we take a look at the data set with jittering as illustrated in Figs. 10 and 11, we observe that most of the users with a large number of queries belong to male users who are semi-anonymous and female users query much less than male users. This validates that analyzing users' demographics and query data sets can help identify anonymous and identifiable users. We then use a machine learning classifier as it is able to accommodate these two features, which are 1 boolean (i.e., Gender) and 1 numeric (i.e., Total Number of Queries). We use the classifier implementations in Weka toolkit [7].

6.2 Customized classifier for user anonymity classification

There are four class instances as shown in Table 1. We cannot train the machine learning classifier only using these class instances because our training and testing data sets contain instances belonging to classes that are neither anonymous nor identifiable. In our case, when using all four classes, we noticed low precision and recall values being reported for unclassifiable accounts. Then we considered using multiple-class (in our case, three-class) classification. In three-class classification, multiple classifiers are built, one for each pair of classes, and the final classification label for an instance is determined on the basis of a voting mechanism. The final results of the three classifiers are then predicted as "Anonymous", "Partially Anonymous", and "Identifiable". When using above three classes, we also noticed low precision being reported for each class. Therefore, in order to optimize the overall accuracy with sufficient recall with not similar size distributions across each category, we simply dropped the Unclassifiable class and instead utilize a customized two-class classification. The final results of the two classifiers are then predicted as "Semi-Anonymous" and "Identifiable". The training data set (3,112 accounts in total) containing two classes, labelled as "Semi-Anonymous" and "Identifiable", is passed through a two-class Support Vector Machine (SVM) classifier with a (Gaussian) radial basis function kernel optimized for detecting above two accounts. By tuning the cost parameters for the two-class classifiers, we can trade off the precision and recall values. In this paper, we intend to choose the cost

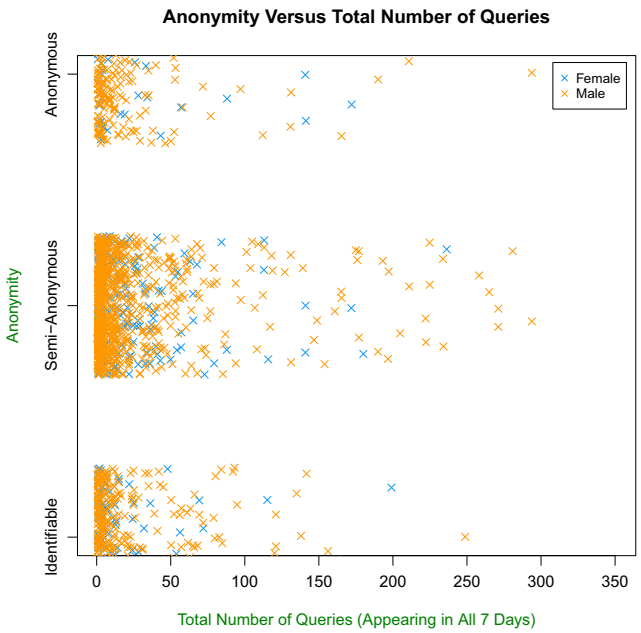


Fig. 10 Anonymity versus total number of queries

Author's personal copy

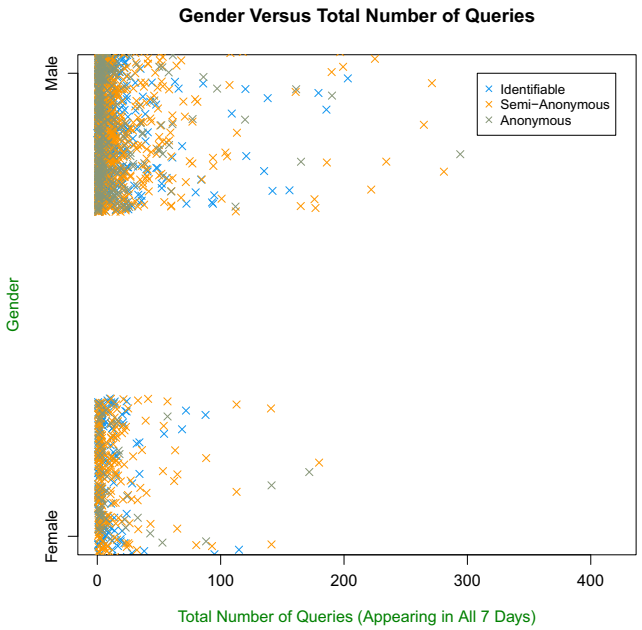


Fig. 11 Gender distribution versus total number of queries

parameters such that two classes achieve sufficient accuracy. Table 2 tabulates the accuracy tested on the given data set.

7 Related work

Location privacy concerns on LBSNs has attracted much attention in recent years. Recent work has developed theoretical and practical methodologies to locate any target *People Nearby* user within a narrow area. Using number theory, Xue et al. [24] theoretically prove that any location-based social discovery user can be located within a circle of radius no greater than one meter, but the model does not consider the possible location errors reported by the LBSN systems. In addition, [6, 11] demonstrate a possibility for either an unsophisticated or sophisticated adversary (a user of the LBSNs) to identify individuals' locations and mobility patterns in any targeted area.

Many psychology and social scientists have also shown that anonymity is the key feature in a social group [1, 4, 10]. Studies focusing on online social networks have shown that a user's username and anonymity may indicate potential user behavior and thus reveal their privacy in a certain degree [9, 13, 15, 17]. Some researchers believe that in LBSNs, anonymity is still an key feature involving privacy and user behavior [18]. Therefore, user anonymity is another focus on LBSNs [8]. By analyzing privacy and anonymity in social networks, Peddinti et al. [16] perform a large scale analysis of Twitter in order to study the prevalence and behavior of Anonymous and Identifiable users and finds a correlation between content sensitivity and a user's anonymity. In contrast, our study is the first work to quantify the relationship between user anonymity and demographics of WeChat users.

Recently, both the popular press and academic scholars suggest that LBSN applications would fundamentally change the nature of urban sociability and alter the way people coordinate gatherings, decide which places to visit or even meet new people [19]. Sex or the possibility of sex is a significant motivator in using the apps such as Grindr and Tinder, and motivates user concerns around self-presentation, identifiability and potential stigmatization (especially around casual sex [5]), and privacy and safety. Although it is true that many users have concerns about privacy [20], it is also true that there are motivators, such as sex, that outweigh these concerns. (e.g., being willing to use the *People Nearby* service in exchange for warranting of others willing to do the same). Another issue raised by the use of current dating apps (i.e., Grindr, Tinder, and Momo) is the way in which people exchange information and flirt with each other in their interactions that lead to sex [3]. Image exchange within conversations, for example, makes it possible to share sensitive images of one's body with a partner before meeting them [23]. In the meantime, however, users of an app like Tinder that provides more information about mutual friends and users of an app like WeChat that can recognize friends when using the *People Nearby* service might be hesitant to act in a manner that this person could tell their friends about that probably leads to a huge embarrassment.

Taking these preliminary analyses of location-based *find-and-flirt* services all together, we know little, however, about how these interactions play out, what the consequences of them might be, or how they play out differently for people across various demographics with different goals or interests. Hence, our study is the first work as a pilot study to attempt to quantify the relationship between user anonymity and demographics of WeChat users.

Table 2 Classifier performance based on 10-fold cross validation

Type of Error	Value
Training Error	0.20
10-Cross Validation Error	0.20
Testing Error	0.19

8 Conclusion

This paper notices that WeChat users do not necessarily use real names as their usernames. Rather, users are given the option to be either anonymous or identifiable. Through

Author's personal copy

close examination of a popular WeChat sub-service, namely, the *People Nearby* service, we conclude: (i) For ordinary WeChat users, the fraction of male users and female users appears to be roughly the same. However, when it comes to using the *People Nearby* service, male users outnumber female users by roughly four to one; (ii) For ordinary WeChat users, anonymous users slightly outnumber identifiable users. However, when it comes to using the *People Nearby* service, anonymous users are twice as many as identifiable users, and semi-anonymous users outnumber identifiable users by roughly five to two; (iii) WeChat users are more anonymous than Twitter users in terms of the fraction of anonymous users. We also take an in-depth look at the user anonymity and demographics in a combined fashion and we find: (iv) For ordinary WeChat users, the fraction of males who are identifiable is almost twice as many as the fraction of females who are identifiable, while the fraction of females who are semi-anonymous greatly outnumbers the fraction of males who are semi-anonymous. However, when it comes to using the *People Nearby* service, the fraction of identifiable males decreases significantly, and the fraction of anonymous males increases significantly; (v) We finally build a machine learning classifier that can be used to detect anonymous and identifiable WeChat accounts by using the *People Nearby* query data sets and users' demographics information. We, therefore, are able to re-define the concept of anonymity in a novel fashion.

To the best of our knowledge, this is the first detailed case study focused on WeChat that quantifies the relationship between user anonymity and demographics of location-based *find-and-flirt* services. By completely answering all these questions we proposed in this paper, we expect our study to gain more significant insights into modern online dating and friendship creation, insights that should be able to not only inform sociologists and psychologists but also inform designers of future *find-and-flirt* services.

Acknowledgments This paper is an extended version of [22]. We would like to thank our 5 labmates for helping classify numerous accounts. This work was supported in part by the NSF under Grant CNS-1318659. This work was also supported in part by the National Natural Science Foundation of China, under Grant 61571191, in part by the Science and Technology Commission of Shanghai Municipality under Grant 13JC1403502.

References

1. Barreto M, Ellemers N (2002) The impact of anonymity and group identification on progroup behavior in computer-mediated groups. Small Group Res 33(5):590–610
2. Bilton N (2014) Tinder, the fast-growing dating app, taps an age-old truth. NY Times 29
3. Blackwell C, Birnholtz J, Abbott C (2014) Seeing and being seen: co-situation and impression formation using grindr, a location-aware gay dating app. New Media & Society, p 1461444814521595
4. Chesney T, Su DK (2010) The impact of anonymity on weblog credibility. Int J Human-Comput Stud 68(10):710–718
5. Conley TD (2011) Perceived proposer personality characteristics and gender differences in acceptance of casual sex offers. J Person Soc Psychol 100(2):309
6. Ding Y, Peddinti ST, Ross KW (2014) Stalking beijing from timbuktu: a generic measurement approach for exploiting location-based social discovery. In: Proceedings of the 4th ACM workshop on security and privacy in smartphones & mobile devices. ACM, pp 75–80
7. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Explor Newslett 11(1):10–18
8. Huang J, Qi J, Xu Y, Chen J (2015) A privacy-enhancing model for location-based personalized recommendations. Distrib Parallel Databases 33(2):253–276
9. Jain P, Kumaraguru P, Joshi A (2013) @ i seek'fb. me': identifying users across multiple online social networks. In: Proceedings of the 22nd international conference on world wide web companion. International World Wide Web Conferences Steering Committee, pp 1259–1268
10. Lelkes Y, Krosnick JA, Marx DM, Judd CM, Park B (2012) Complete anonymity compromises the accuracy of self-reports. J Exper Soc Psychol 48(6):1291–1299
11. Li M, Zhu H, Gao Z, Chen S, Yu L, Hu S, Ren K (2014) All your location are belong to us: breaking mobile social networks for automated user location tracking. In: Proceedings of the 15th ACM international symposium on mobile ad hoc networking and computing. ACM, pp 43–52
12. Lim Ys, Van Der Heide B (2015) Evaluating the wisdom of strangers: the perceived credibility of online consumer reviews on yelp. J Comput-Mediated Commun 20(1):67–82
13. Liu J, Zhang F, Song X, Song YI, Lin CY, Hon HW (2013) What's in a name?: an unsupervised approach to link users across communities. In: Proceedings of the sixth ACM international conference on web search and data mining. ACM, pp 495–504
14. Nemelka CL, Ballard CL, Liu K, Xue M, Ross KW (2015) You can yak but you can't hide. In: Proceedings of the third edition of the ACM conference on online social networks. ACM
15. Peddinti ST, Korolova A, Bursztein E, Sampemane G (2014) Cloak and swagger: understanding data sensitivity through the lens of user anonymity. In: 2014 IEEE symposium on security and privacy (SP). IEEE, pp 493–508
16. Peddinti ST, Ross KW, Cappos J (2014) On the internet, nobody knows you're a dog: a twitter case study of anonymity in social networks. In: Proceedings of the second edition of the ACM conference on online social networks. ACM, pp 83–94
17. Perito D, Castelluccia C, Kaafar MA, Manils P (2011) How unique and traceable are usernames? In: Privacy enhancing technologies. Springer, pp 1–17
18. Preoţiuc-Pietro D, Cohn T (2013) Mining user behaviours: a study of check-in patterns in location based social networks. In: Proceedings of the 5th annual ACM web science conference. ACM, pp 306–315
19. De Souza e Silva A, Frith J (2010) Locative mobile social networks: mapping communication and location in urban spaces. Mobilities 5(4):485–505

Author's personal copy

20. Toch E, Levi I (2013) Locality and privacy in people-nearby applications. In: Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing. ACM, pp 539–548
21. Wang G, Wang B, Wang T, Nika A, Zheng H, Zhao BY (2014) Whispers in the dark: analysis of an anonymous social network. In: Proceedings of the 2014 conference on internet measurement conference. ACM, pp 137–150
22. Wang R, Xue M, Liu K, Qian H (2015) Data-driven privacy analytics: a wechat case study in location-based social networks. In: Wireless algorithms, systems, and applications. Springer, pp 561–570
23. Weisskirch RS, Delevi R (2011) Sexting and adult romantic attachment. Comput Human Behav 27(5):1697–1701
24. Xue M, Liu Y, Ross KW, Qian H (2015) I know where you are: Thwarting privacy protection in location-based social discovery services. In: 2015 IEEE conference on computer communications workshops (INFOCOM WKSHPS). IEEE, pp 179–184

Minhui Xue is pursing his Ph.D. degree at the Department of Computer Science of East China Normal University, focusing primarily on computer science and mathematics, including data-driven privacy analytics, machine learning applications, online social networks, and cryptography. He received a Bachelor of Science degree in the field of fundamental mathematics from East China Normal University in July 2013, recipient of the Elite Student Scholarship from Fundamental Mathematics Honors Program (National Science Base Class). He is currently serving as a researcher in New York University Shanghai, co-advised by Professor Keith W. Ross (NYU) and Professor Haifeng Qian (ECNU).

Limin Yang is pursing his master's degree at the Department of Computer Science of East China Normal University with a focus on information security, machine learning applications, and cryptography. He received a Bachelor of Science degree from the Computer Science Department at East China Normal University in July 2015. He is currently serving as a research assistant at Institute of Computer Science & Technology of Peking University, focusing primarily on vulnerability discovery and automated exploit detection.

Keith W. Ross is the Dean of Engineering and Computer Science at NYU Shanghai and the Leonard J. Shustek Chair Professor in the Computer Science and Engineering Dept. at NYU. Before joining NYU-Poly in 2003, he was a professor at University of Pennsylvania (13 years) and a professor at Eurecom Institute (5 years). Professor Ross was the Department Head of the CSE Department at NYU-Poly from 2008 to 2013, and he joined NYU Shanghai in 2013. He received a B.S.E.E from Tufts University, a M.S.E.E. from Columbia University, and a Ph.D. in Computer and Control Engineering from The University of Michigan. Professor Ross's current research interests are in data-driven analysis of online social networks and privacy. He has also worked on peer-to-peer networking, Internet measurement, video streaming, applied probability and Markov decision processes. He is an ACM Fellow, IEEE Fellow, recipient of the Infocom 2009 Best Paper Award (1,435 papers submitted), and recipient of 2008 and the 2011 Best Paper Awards for Multimedia Communications (awarded by IEEE Communications Society). His work has been featured in the New York Times, NPR, Bloomberg Television, Huffington Post, Fast Company, Ars Technia, and the New Scientist. Professor Ross is co-author (with James F. Kurose) of the popular textbook, Computer Networking: A Top-Down Approach Featuring the Internet, published by Addison-Wesley (first edition in 2000, sixth edition 2012). It is the most popular textbook on computer networking, both nationally and internationally, and has been translated into fourteen languages. Excluding introductory programming textbooks, it is the fifth most popular CS textbook overall. Professor Ross is also the author of the research monograph, Multiservice Loss Models for Broadband Communication Networks, published by Springer in 1995.

Professor Ross has served on numerous journal editorial boards and conference program committees. He was PC co-chair for ACM Multimedia 2002, ACM CoNext 2008, and IPTPS 2009. From July 1999 to July 2001, Professor Ross took a leave of absence to found and lead Wimba, which develops voice and video applications for online learning. He was the Wimba CEO and CTO during this period. Wimba was acquired by Blackboard in 2010.

Haifeng Qian is a professor at the Department of Computer Science of East China Normal University, Shanghai, China. He received a BS degree and a master degree in algebraic geometry from the Department of Mathematics at East China Normal University, in 2000 and 2003, respectively, and the PhD degree from the Department of Computer Science and Engineering at Shanghai Jiao Tong University in 2006. His main research interests include network security, cryptography, and algebraic geometry.