공학박사학위논문

지리적 거리 정보를 활용한 가짜 팔로워 구매자 식별 방법

Fake Follower Market Customer Detection based on Geographical Distance Information

2019년 2월

서울대학교 대학원

컴퓨터공학부

장 보 연

지리적 거리 정보를 활용한
가짜 팔로워 구매자 식별 방법

Fake Follower Market Customer Detection based on Geographical Distance Information

지도교수 김종권

이 논문을 공학박사 학위논문으로 제출함

2018년 10월

서울대학교 대학원

컴퓨터공학부

장보연

장보연의 공학박사 학위논문을 인준함

2018년 12월

위원장 : 이상구 (인)

부위원장 : 김종권 (인)

위원 : 권태경 (인)

위원 : 강유 (인)

위원 : 양은호 (인)

Abstract

Fake Follower Market Customer Detection based on Follower Ratio using Geographical Distance

Bo Yeon Jang
Department of Computer Science & Engineering
The Graduate School
Seoul National University

The reputation of social media such as Twitter, Facebook, and Instagram now regard as one person's power in real-world. The person who has more friends or followers can influence more individuals. So the influence of users is associated with the number of friends or followers. On the demand of increasing social power, an underground market has emerged where a customer can buy fake followers. The one who purchase fake followers acts vigorously in online social network. Thus, it is hard to distinguish customer from celebrity or cyberstar. Nevertheless, there are unique characteristics of legitimate users that customers or fake followers cannot manipulate such as a small-world property. The small-world property is mainly qualified by the shortest-path and clustering coefficient. In the small-world network, most people are linked by short chains. Existing work has largely focused on extracting relationship features such as indegree, outdegree, status, hub, or authority. Even though these research explored the relationship features to classify abnormal users of fake follower markets, research that utilize the small-world property to detect abnormal users is not studied.

In this work, we propose a model that adapt the small-world property. Specifically, we study the geographical distance for 1hop-directional links using node′s geographical location to verify whether a social graph has the small-world property or not. Motivated by the difference of distance ratio for 1hop directional links, we propose a method which is designed to generate 1hop link distance ratio and classify a node as a customer or not. Experimental results on real-world Twitter dataset demonstrates that the proposed method achieves higher performance than existing models.

keywords : Online Social Network, Twitter, Fake Follower Market, Customer, Follow Spam, Geographical Distance

Student number : 2015-30272

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

We live in the small-world that two arbitrary people are likely connected by a short chain of intermediate friends. The Small-world phenomenon is a principle that most people in a society are linked by short chains of acquaintances. The Small-world phenomenon has been a matter of folklore in social networks. In 1967 Milgram [1] performed the experimental study about forwarding a document from a people in Nebraska to a person in Boston, and found that people in the United States are connected by ″six degrees of separation." A mathematical model to explain how the small-world phenomenon operates is proposed in [2]. The study showed that the small-world network is highly clustered and have shortest average paths. Their approach had been adopted to World Wide Web graph [3], and the results showed that distribution of connectivity decays with a power law tail. [4] conducted a small- world experiment on Facebook, a typical online social media service, which found that OSNs are networks of ″Four degrees of separation.″

In 2011, Kleinberg [5] suggested a decentralized algorithm that constructs short paths with non-negligible probability. The Kleinberg introduced a model that builds a grid graph with additional long-range random links with the probability. Though the Watts-Strogatz model [2] requires the whole link information to find the shortest path to a target, Kleinberg's model constructs an efficiently searchable network using decentralized search from the real-world networks. He found that the link probability between two nodes u, v follows the probability proportional to distance $(u,v)^{-r}$. In the geographical small-world model, the probability of linking to each node is proportion to d^{-2}. Many empirical studies are conducted to find the evidence of the exponent r=2. The authors in [6] measured the friendship link probability from LiveJournal which is one of the blogging sites in OSNs. They measured the friendship distance using the geographic information and found patterns of friendship link distribution across geography. The small-world property from other blog site has studied in [7]. [8] solved Pear-to-Pear routing problem by incorporating a small-world model. [9] found that the Boston subway had a small world property and presented the insight for the transportation networks.

Many existing researches showed that OSNs have the small-world property. However, there was no spam detection method using the small-world property that the link probability decreases as the distance of two nodes increases. Normally spam establishes a random link with no probability distribution so we can assume that a social relationship for spam does not follow this property. Therefore, in this paper, we focus the small-world property to detect the spam. Especially we target spam in the fake follower market which contains arbitrary relationships.

1.2 Fake Follower Markets

Twitter which is one of the top 20 most-viewed web sites [10] has a different form of relationship with other OSNs. Twitter's relationship is composed of follower and followee, which can be established even if the other party does not accept the link request. Therefore, a relationship can be easily created between users who do not know each other in the real-world. Also, users in Twitter have a various notification for others′ activities. Activities of friends or followers such as tweet, retweet or status update are forwarded through user′s pages. Due to its unique relationships and notifications, the influence or reputation of users are associated with the number of twitter followers or friends [11] [12]. The third-party service (Klout) [13], which measures social status, uses the number of followers and the frequency of re-sharing content as influence measures when distinguishing celebrities or normal users [14]. To gain reputation, some users accept friendship request that they receive from an unknown user or follow other users with the expectation of following back [15].

There are many ways to increase followers such as following others, sending a message to real friends, posting interesting tweets, assigning campaign, or buying followers from sellers. However, increasing followers through daily activities is a time-consuming job. The one who wants to increase their followers can be attempted to purchase fake followers from the spammer. According to the Twitter rules [16], the user who uses or promotes third-party services that claim to get more followers can be regarded as a spammer. However, it is easy to find promoting campaigns to increase followers. Even popular celebrities have fake followers from other researches [17] [18]. Also, malicious users may purchase fake followers to spread their malware or spam more rapidly [19] [20]. Many OSNs define buying some fake followers or others as a violation. However, it is easy to find fake follower market only by typing followers in google. Figure 1.1.1 shows one of the advertisements that can be found in Online. The advertisement shows that 1,000 followers only cost 17 dollars.

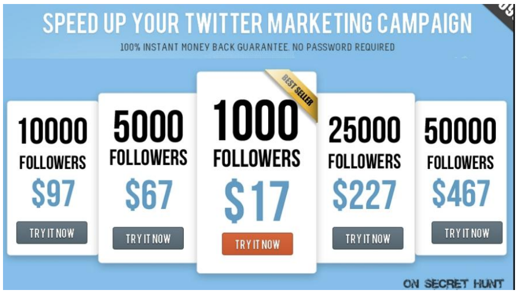

SPEED UP YOUR TWITTER MARKETING CAMPAIGN
100% INSTANT MONEY BACK GUARANTEE. NO PASSWORD REQUIRED
10000 FOLLOWERS $97 TRY IT NOW
5000 FOLLOWERS $67 TRY IT NOW
BEST SELLER
1000 FOLLOWERS $17 TRY IT NOW
25000 FOLLOWERS $227 TRY IT NOW
50000 FOLLOWERS $467 TRY IT NOW
ON SECRET HUNT

Figure 1.1.1: The Advertisement of Twitter Marketing Campaign

Fake follower market consists of three components that are merchantS, customers, and fake followers. Figure 1.1.2 shows the components of fake follower markets. Merchant posts advertisement and provides followers. And the customer who wants to increase followers gives a payment or other credential information to the merchant. The merchant provides fake followers through following activities to the customer. Fake followers could be compromised accounts, real-person accounts, or bot accounts based on its platform or service.

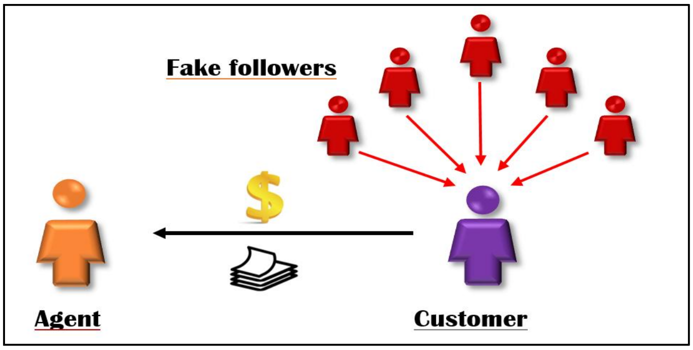


Figure 1.1.2: Component of Fake Follower Markets

1.3 Research Objectives

There are various previous researches to detect fake followers but only a few existing researches to detect customers. Detecting customers is more sensitive problem than detecting fake followers because some customers used to be legitimate users before being customers and they even don't realize that buying fake followers is a violation of OSNs site policy. However, if the customers who bought fake followers from markets can be caught then Twitter can suspend customer's account and fake followers' accounts. It will make the size of the fake follower markets smaller. In this research, we focus on the characteristics of fake followers and customers, then suggest the method to detect customers.

Given social relationship, our goal is to (i) verify whether a social network for fake follower market has a small-world property and (ii) classify customers using characteristic of follower distance distribution. The two problems are formally defined below:

Problem 1. Given an online social graph and node′s location, identify whether there exists a small-world property in a social network for fake follower market through measuring a correlation between online social relationship and geographical distance, then discover differences of customer′s follower link distance distribution from legitimate users.

Problem 2. Can we detect customer of fake followers? Given a set of users U, classify whether a user is a customer or legitimate user using its follower link distance distribution

Though we conclude to propose a detection method in this paper, the first problem is essential to solve the second problem. In order to solve these two problems, we first explain how to collect our dataset, then validate the first problem and solve the second problem based on answers to the first question.

1.4 Contributions

To investigate fake followers and customers, we collect twitter dataset containing fake followers and customers. By purchasing fake followers from sellers, we obtain 24,552 fake followers and identify other customers who purchase from the sellers using labeled fake followers. Also, we collected legitimate user dataset which contains profiles and relationship information.

As a result of analyzing the collected data, we discover that the link probability of the legitimate users in the Twitter dataset shows the small-world phenomenon that the link probability decreases as the distance between the two nodes increases. However, a customer of fake followers did not show this phenomenon. Based on this analysis we tried to find the differences between legitimate users and customers. We find that the legitimate user′s follower ratio between a number of followers in a specific distance range and the total number of followers also decreases when the distance between two nodes increases. But the follower ratio of customers does not show a decrease.

Considering that customer of fake followers is different in terms of geographical distance from legitimate, we conclude that this characteristic can be helpful to identify customers. Therefore, the classification was performed using the follower distance ratio. Our experiments on real Twitter dataset clearly show that our mechanisms perform better than other methods. Additionally, we test whether fake followers can be identified by following distance ratio and show that our method is effective than other methods.

In summary, we frame our contributions as followers;

- We show that the legitimate user′s friendship pattern in OSNs shows a similar pattern in the real-world by comparing link probability of each domain, but abnormal users do not show this phenomenon. In this research, we found that this small-world phenomenon can be applied only for legitimate users in OSNs and abnormal users do not follow small-world phenomenon.

- We identify characteristics that customers of fake followers have different follower distance ratio from legitimate users in OSNs. The follower ratio of legitimate users decreases when a distance from seed user increases. However, customer or fake followers did not follow this pattern and showed little change with distance.

- To the best of our knowledge, our approach is the first attempt with adapting a small-world property to detecting customers of fake followers in OSNs. Previous studies have only focused on finding the small-world property in legitimate social network and does not apply to spam domain. Also, other spam detection method has considered online network features or user features. However, physical distance is not applied to detect customers of fake follower market or other spam. We conduct customer detection experiment based on the small-world property and the previous finding that the follower ratio of the customer is different from others and shows suggested method is effective than other previous methods.

1.5 Thesis Organization

The following chapters of this dissertation are organized as follows:

- Chapter 2 reviews the existing studies on the small-world phenomenon and online social abusing attack detection.

- Chapter 3 explains how to collect datasets, and introduces the characteristics of customers and followers.

- Chapter 4 shows the friendship pattern in OSNs by measuring node′s distance and shows that the friendship pattern in OSNs is similar to the real-world. Also, we analyze the difference of follower distance distribution between legitimate users and customers.

- Chapter 5 proposes a customer detection algorithm based on the previous findings and evaluates the performance.

- In Chapter 6, we present our conclusion on the research results of this dissertation.

Chapter 2

Related works

2.1 Small World Phenomenon

Stanley Milgram conducted a small-world experiment to measure the average path length of social networks through a letter-delivery experiment in the United States [1]. He found that letters averaged about six steps to reach their destination, which was later called ″Six degrees of separation.″ [2] defined the small-world network as neither completely regular nor a completely random network. They stated that this network is likely to be highly clustered. So, in the small-world network, nodes can be accessed with only a few hops. This phenomenon does not only occur in the offline social network. [4] conducted a similar small-world experiment on Facebook, a typical online social media service, which found that OSNs are networks of ″Four degrees of separation.″

In online society, people are connected through fewer steps. [5] found that the probability of being linked is reduced in inverse proportion to distance. This study found that the small-world network is searchable, and it is proven mathematically.

Previous researches showed that the small-world phenomenon exists in various networks such as the transportation system, citation network, or online network. For example, [9] found that the Boston subway had the small-world property and presented the insight on the general characteristics of the transportation networks. In addition, [3] stated that the development of the search engine caused the World Wide Web to be a small-world network with small path length. A representative blogging service called Blogosphere is also known as an example of a small world network [34, 21] analyzed the small- world property based on the continent (geographic information) of the users in Twitter site. The small-world phenomenon also occurs in the social collaboration networks, such as the film actor network [2], or the patent citation network [22].

Based on previous studies, it can be seen that human contact networks normally take the form of the small-world phenomenon. They commonly have a degree distribution, followed by a power law distribution [23]. In this study, they focus on the geographically limited movement of users of online social networking services (SNSs) [24] and observe the small-world phenomenon through the relationship between the location of Twitter users and the probability of being linked.

2.2 Online Social Abusing Attack Detection

As social media has begun to play a major role in people's information sharing and friendship, various social attack strategies aimed at social media are recently being conducted. However, excessive spam is damaging to the user's use of social media. For example, if a user wants to read real users' reviews before purchasing a particular product, he or she will see more advertisements than purely written reviews. He or she may also receive a lot of unwanted spam messages from SNSs that communicate with friends. Or he or she might see frequent spam content on his or her newsfeed. Such social attacks hinder the purity of the online social media and have a major impact on usability. Eventually, services that do not prevent social attack will deteriorate.

Even though inflating relationship like fake followers is a passive attack that gives less impact than other social abusing attack, it is one of the social abusing attacks that can hinder the purity of its domains. Also purchasing fake followers can be a means for malicious entities to spread malware and spam [25, 26, 27, 40]. Because of this, many previous studies have tried to detect social attacks including fake followers in the OSNs. We divide these into three ways to detect social abuse attacks: Contents-based, Social Network-based, and Behavior-based spam detection.

2.2.1 Contents-based Detection

Contents-based detection is the most general approach since the early email spam detection study. This method analyzes a sentence or vocabulary used by a social attacker, such as a spammer, when they send a message or upload spam content to social media. When a vocabulary or sentence with a spam characteristic appears in a specific document, it is judged whether or not it is spam, based on the contents learned in advance. In this case, sentence-based template matching is also used for this approach [29]. These are normally machine learning-based detection methods, such as Nave Bayes or SVM, which are frequently used for e-mail spam and SMS spam detection [30, 31].

Recently, previous researches have taken advantage of the characteristics of spam contents frequently seen in certain microblogging services, such as Twitter and Facebook. They noted that when using microblogging services, spammers frequently use URLs or hashtags with a limited number of characters. [32] proposed a method to detect a compromise attack through account hijacking, and found that the service usage pattern of compromised accounts is significantly different. In particular, they focused on the phenomenon that suddenly many URLs appeared in the content used by the attackers, or the language used changed [23]. [33] and [34] also use a spam detection method that simultaneously utilizes the use of hashtags and URLs.

2.2.2 Social Network-based Detection

The social network-based detection is a method of link analysis by expressing social media entities (i.e., users or pages) and their interactions in a social graph. The most commonly used methods are Pagerank [35] and HITS [36] based detection methods, which are useful for finding fake followers that increase the number of subscribers of a specific account or page during a social attack. This is because these approaches have been effectively used to detect spamdexing before the detection of social attacks [37]. Spamdexing was also performed by linking multiple fake pages to a target web page, in order to increase the rank of specific pages in the search engine. DetectVC [38] and CatchSync [39] both used the Hub and Authority scores computed by the HITS algorithm to detect fake followers at the microblogging sites. In addition to the Spamdexing type detection, there is a method to detect general spamming text that propagates spam contents to a large number of unspecified users. CollusionRank [40] is a typical Pagerank based detection method.

For Sybil attack detection, a social link analysis based approach is also used. SybilRank [41] is a method of ranking the Sybil nodes. This approach is a social network-based detection approach like SybilLimit [42] and SybilInfer [43], but shows better performance. FRAUDAR [44] is a dense subgraph detection method and a fraud detection approach using link analysis. Another approach, widely adopted in practice [45, 46, 47], is to find lockstep behavior from the OSNs. Lockstep behavior detection is one of the social network-based detection. Lockstep behavior occurs when groups of users act together. CopyCatch [46] is a method to identify suspicious lockstep behavior like Facebook page-like patterns by analyzing the social graph and finding bipartite cores. Authors in CROSSSPOT [47] suggest a solution to detect synchronized fake likes through building a suspiciousness metric and identifying suspicious blocks. LockInfer [45] uncovers lockstep behaviors by characterizing connectivity pattern from various kinds of network such as who-follows-whom network and Patent Citation network.

2.2.3 Behavior-based Detection

Behavior-based detection is a method of analyzing behaviors that can be defined as interactions made by users. The behaviors in OSNs include various activities such as Facebook like, Twitter reply / retweet, following / subscribing, sending e-mail/message, etc. So behavior-based detection is applied to different kinds of social abuse attacks. BPNN [48] presents a method for spam message filtering using rule-based processing and backpropagation neural networks. BPNN utilizes the spamming behaviors as features for describing emails. Authors in [49] proposed a method to detect malicious accounts by analyzing the aggregate clickstream behavioral pattern.

In regards to spambots, authors in [50] showed that legitimate accounts share lower behavior similarities between others and [51] have exploited a new approach to detect spambot groups using DNA-inspired behavior that is obtained from each account by encoding their behaviors as a digital DNA sequence. [52] is an unsupervised method to detect bots using a temporal pattern that is obtained from bot activities. Authors in [53] present a detailed study of Twitter follower markets and propose a method to detect customers of fake follower market using follower dynamics properties and static properties.

Chapter 3

Characteristic of Customers and Followers

We start by presenting our process for collecting a dataset of Twitter accounts that are grouped into legitimate user, fake followers and customers of fake follower markets. We then describe the analysis of a dataset to identify the difference between legitimate users and other groups.

3.1 Data Preparation

The dataset is collected using Twitter open API [54] between February 2017 and January 2018. The dataset contains the legitimate user set U_n, fake follower set U_{FF} and fake follower's customer set U_{FC}. We used Google Map Geocoding API [55] for obtaining coordination from user′s location. To obtain the Twitter fake followers we have to select merchants (or sellers) who sell fake followers first. The merchants can easily be found in the Twitter website, other OSNs site, or commercial sites. We can group the merchants into seller-driven markets and buyer-driven markets. Table 3.1.1 shows the fake follower markets that are categorized into two groups (seller-driven market and buyer-driven market). The seller-driven market is a traditional market where sellers advertise their service of follower delivery. The seller-driven market includes Fiverr, BigFollow, InterTwitter, GetmoreFollowers, BigFolo, NewFollow, SNSHelper, Devumi, and etc. The user who want to buy fake followers can easily purchase fake followers from one of these seller-driven market. However, the buyer-driven market is one of the crowdsourcing platforms where participants can select their work for profits. The fake followers from the seller-driven market tend to be fake accounts or compromised accounts. The buyer-driven market can be found in MicroWorkers, Amazon merchanical turk, CrowdFlower, or other crowdsourcing sites. However, any legitimate users who want some profit can be fake followers, because they can easily access and join the following tasks through the buyer-driven online website [56].

Table 3.1.1: Fake Follower Market List.

	Market list
Seller-driven market	Fiverr, BigFollow, InterTwitter, GetmoreFollowers, BigFolo, NewFollow, SNSHelper, Devumi
Buyer-driven market	MicroWorkers, Amazon merchanical turk, CrowdFlower

Our goal for this research is to find customers of fake follower markets through multiple following activities, so we cannot get enough information from the one who temporarily joins the fake following task. For this reason, we select the seller-driven market for collecting fake followers. The sellers from each market advertise their services and special additional conditions. The price for 1,000 followers varies between $1.7 and $14.0.

Table 3.1.2: Information about Fake Follower Market.

	Ordered Followers	Maximum Followers	Delivery start time after initiating an order	Duration time since the 1st follower
Seller A	2,500 / $35	3,377	29H	394H
Seller B	7,000 / $30	7,714	19H	19H
Seller C	4,800 / $16	5,048	31H	1H
Seller D	7,000 / $45	7,797	238H	12H

average (median)

We chose four sellers from the seller-driven market to collect the ground truth of fake followers. In particular, we bought 2,500 fake accounts from http://devumi.com (Seller A), 7,000 and 4,800 from http://www.fiverr.com (Seller B and C), and 7,000 fake accounts from http://intertwitter.com (Seller D). Purchasing fake followers is a general process that has performed in other previous research [53, 56, 57] to collect ground-truth followers. To ensure that our work done is supported by ethical principles, we followed the guidelines outlined by Hewson et al. [58]. In particular, we tried to follow social responsibility and minimizing harm. To do that, we created dummy Twitter accounts for the sole purpose of conducting experiments in this paper. We did not build any following relationship with other accounts and did not post any tweets or retweets.

As shown in Table 3.1.2, we purchase a different number of followers and get delivery of the fake followers to each seed account. Since the seed accounts were newly-created and had no followers or friends, we consider any account that started following our seed accounts a fake follower. The prices for fake followers are diverse, but all four sellers delivered extra followers than our payment. The start time of delivery for three sellers is between (19 and 31) hours, excepting seller D. We did not receive any response from seller D for eight days. We concluded that the seller must be a deceiver. But on the ninth day of the order, we received a message from seller D, and he completed our order with extra followers.

The active time for each delivery is within one day, except for seller A, as shown in Figure 3.1.1. Almost every seller shows burst time in the early stage. In particular, seller C gave us 4,800 followers within one hour. This phenomenon shows that the sellers can control multiple accounts in a short period. This means that seller might have multiple fake accounts, or seller could be employers of fake account users. Seller A shows a steady delivery for 17 days, in spite of their advertisement with 2-days-delivery. In total, we have 24,552 purchased fake followers, including makeup followers, which are given for the replacement of unfollowing users. Based on the account id from collected fake followers, we collected their profiles and relationships using Twitter API.

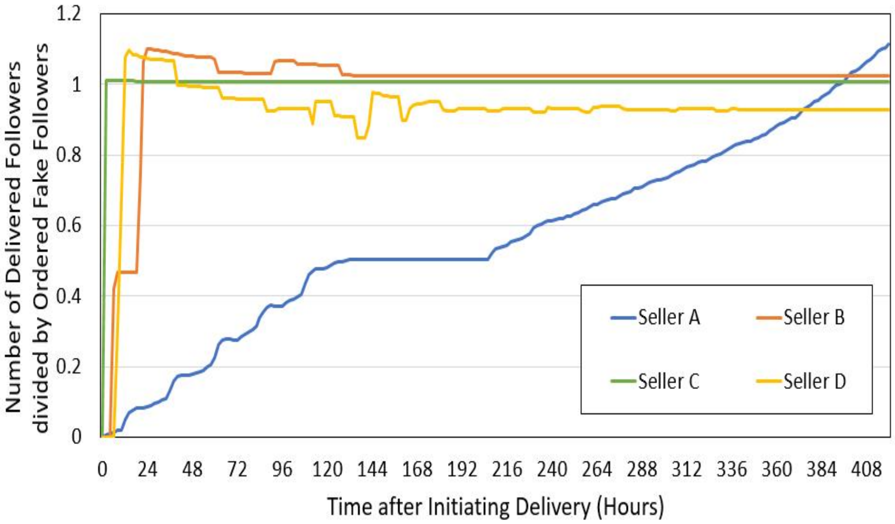


Figure 3.1.1: Follower Delivery Ratio.

After collecting the fake followers, the next step is to identify the customers of fake follower markets. Making fake follower accounts by joining Twitter doesn't require great effort. However, Twitter accounts are sellers' assets, so they usually reuse their accounts to sell fake followers. From this idea, we tried to extract the customers. If an account has the same fake followers that we identified by purchasing fake followers, then it could be a customer of common fake followers. We selected 3,000 candidate customers in descending order of the number of fake followers. After that, we crawled the profiles of candidates and excluded accounts that have no informative location, such as blank, emojis, meaningless sentence, or large area unit in their profiles. We also removed accounts that were suspended, or that had private link information. The final set of customers included 2,341 accounts.

Legitimate accounts of Twitter may have generic users, such as celebrities or local residents. In this paper, we use local news accounts, volunteer groups, local churches, and sports groups on Twitter to collect legitimate accounts. The reason for this is that our method uses local coordinates, and local news or volunteer groups originate from the assumption that local people will follow an official account to get information or interact with others. To collect account from the original group, we use different types of links; we collect followers for local news and volunteer groups and friends for local churches and sports groups. As in the same case of selecting customer candidates, the accounts that only have available link information and meaningful location information in the profile are added to our legitimate user dataset. We then have a total of 7,933 accounts as a legitimate dataset.

3.2 Fake Follower Properties

This section describes the characteristics of the collected fake follower dataset. We collect 24,552 profiles of purchased fake followers. We analyze the collected profiles of the fake followers and found their properties. Since we collect fake followers via OSNs, we could not verify whether each account is an artificial account or a real user's account who sell an account for a benefit. Thus both artificial accounts and real users' accounts could fall in the macro-category of fake followers in our dataset. However, the fake followers that were collected during our collection phase could fall in to the artificial accounts based on the reasons as follows: First, seller-driven markets provide fake accounts or compromised accounts and compromised accounts appear only in the pyramid markets that offer both premium and free subscription [53, 56]. However, we have selected the merchants who only provide premium service. Second, collected followers show the characteristics with fewer followers, big follower / following ratio, and low activities (tweet and favorite) compared with legitimate users. Third, the delivery ratio of fake followers shows a sharp increase, which is limited for real users to conduct in a short period. For these reasons, we can assume that the follower accounts that we collected are fake accounts.

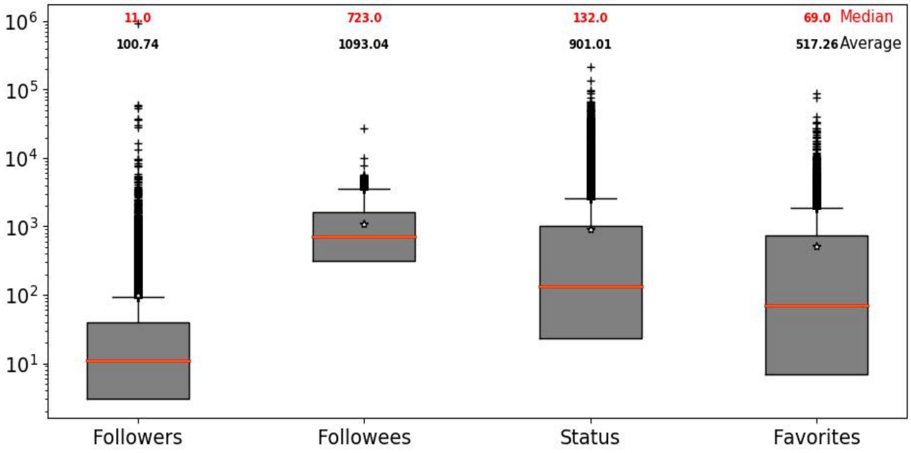


Figure 3.2.1: Properties of Fake Followers

Figure 3.2.1 shows the detailed information of analyzing the relationship and activities. The accounts that act as fake followers have 101 followers on average, with a median of 11 followers. We found that 7.3 % of fake follower accounts do not have any followers, and 45.9 % of them have less than 10 followers. Only 7.6 % of fake followers have more than 100 followers. The number of followees is relatively higher than the number of followers.

Average of followees is 1,093 with a median of 723.

Table 3.2.1: Comparison of Fake Followers from Various Sellers.

	Number. of Follower	Number of Followee	Number. of Tweet / Retweet	Number of Favorites
Seller 1	59.6 (56)	2,678.1 (2,182)	1,549.8 (1,509)	1,281.0 (1,226)
Seller 2	33.1 (8)	992.0 (540)	839.9 (117)	726.4 (86)
Seller 3	8.4 (4)	649.1 (319)	180.8 (137)	135.4 (107)
Seller 4	234.1 (13)	882.8 (649)	1,120.8 (26)	248.3 (5)

average (median)

Table 3.2.1 shows the fake follower′s activities of each seller. We can find that the characteristics of fake followers are different based on each seller. The number of followers for seller 4 is four times bigger than seller 1. The fake followers from seller A and seller B show relatively higher numbers of tweet/retweet or favorites. This can be explained by the sellers using fake follower accounts for different activities, such as fake favorite, or posting spam tweet/retweet. In the case of seller A, the website advertises retweets or favorites.

To investigate how long the fake followers can survive, we analyze each account creation date. We identify the number of active accounts at the time of the initial transaction by year of account creation and confirm whether the accounts are still active or suspended after 12 months. Figure 3.2.2 presents the results. In the first phase, fake follower accounts generated in 2016 account for the highest percentage of all accounts, with 6,139 accounts. In 2012, 4,198 accounts are ranked as second. However, after 12 months, many of the accounts created in 2016 are suspended, and the number of active accounts decreases by more than 80 % to 1,227. The second largest decrease is accounted for by 73 % of the accounts generated in 2014. However, the percentage of accounts suspended before 2013 is less than 30 %, indicating that recently created accounts are suspended at a relatively large rate. This confirms that there is an effort to identify the fake followers on Twitter itself, and to remove these accounts. Also, our results have a similar pattern to the previous research [57] that provides an analysis of Twitter spambot and proposes a detection method. The suspension rate of [57] is 73.5 % which is very close to our suspension rate fora fake account between 2014 and 2017.

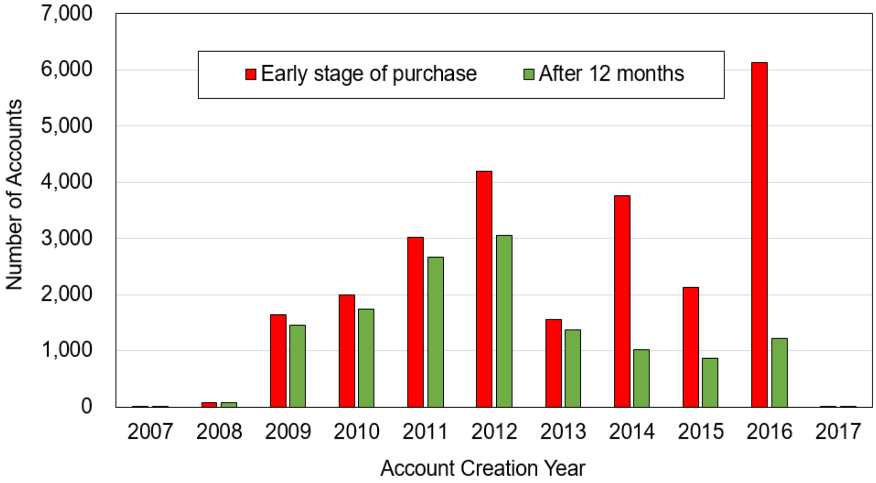


Figure 3.2.2: Number of Fake Followers Grouped by Year.

The key features in our research are the user's location and coordination. So, we collect all fake followers' location information from their profiles and convert the location to latitude and longitude using Google Map Geocoding API. Among the total fake followers, 8,212 accounts have no data for location field, and we can obtain only 15,339 meaningful coordinates (longitude and latitude) from the remaining 16,340 accounts that have location information in their profiles. This is because as the user can manually type the location information in the profile field at will, sometimes he or she can enter a meaningless place or an erroneous character. Although our experiment cannot be performed with the datasets that include the fake follower accounts without coordinate information, these datasets can be useful for other methods that do not utilize coordination information. Thus we exclude these accounts only when we measure the performance of our methods.

Figure 3.2.3 shows the Indegree/Outdegree of normal followers and fake followers. We use the fake follower set that we mentioned above. To obtain the normal followers, we utilize followers of legitimate users collected from the legitimate user dataset. As a result, we can find that there is a degree difference between normal followers and fake followers. In the ratio of outdegree and indegree, the ratio of the normal follower is 1.96, while that of the fake follower is 0.18 on average. The median ratio of normal followers is 0.33, while the median ratio of fake followers is 0.01.

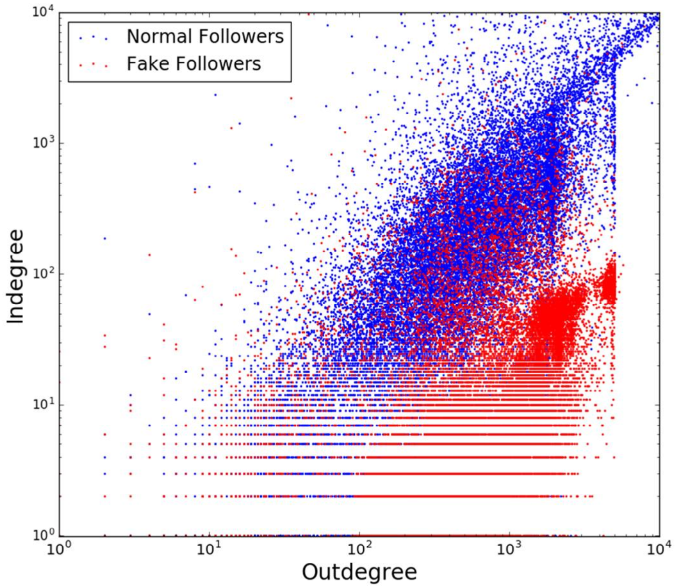


Figure 3.2.3: Indegree vs. Outdegree for Normal Followers and Fake Followers.

3.3 Customer Properties

In order to distinguish the customers from the legitimate users, we first analyze the properties of customers that are identified from fake followers. We analyze 2,341 customer accounts. The customer has an average of 1,347 fake followers, and the proportion of fake followers among all followers is 26 % on average. However, since we are limited to identifying the entire fake followers during collection of the fake follower dataset, we assume that the ratio of the fake followers is higher than our measured ratio.

The data collected from the user profiles is shown in Table 3.3.1. In the table, 'Days since Registration' is the number of days between account creation date and January 31, 2018. Average of 'Days since Registration' is 1,778 days for customers, and 2,373 days for legitimate users. In the follower link, the number of customers is ten times more than the number of legitimate users. In the case of followee numbers, the customers have about three times more followees than legitimate users. While customers and legitimate users show a big difference with respect to the number of links, there is a relatively small difference between activities such as tweet and favorite. In particular, the `Number of Tweet/Retweet' of customers is 1.6 times greater on average than that of legitimate users, but the median of the customer is almost the same as that of the legitimate user. `Number of Favorites' shows differences, which means that customers use their favorites more often than legitimate users.

Table 3.3.1: Profile Attributes of Customers and Legitimate Users.

	Customer	Legitimate user
Days since registration	1,778 (1,726)	2,373 (2,480)
Number of Follower	46,518 (12,442)	4,650 (537)
Number of Followee	5,060 (648)	1,580 (859)
Number of Tweet / Retweet	7,947 (1,386)	5,123 (1,359)
Number of Favorites	4,869 (495)	1,898 (295)

average (median)

To measure follower distance, we calculate a distance between the seed node and each follower node, then average all distances for the seed node. Figure 3.3.1 shows the follower distance for customers and legitimate users. The average follower distance of customers is 2,568 kilometers, while the average follower distance of the legitimate user is 1,180 kilometers. This means that the legitimate user′s followers are located closer than the customer′s followers.

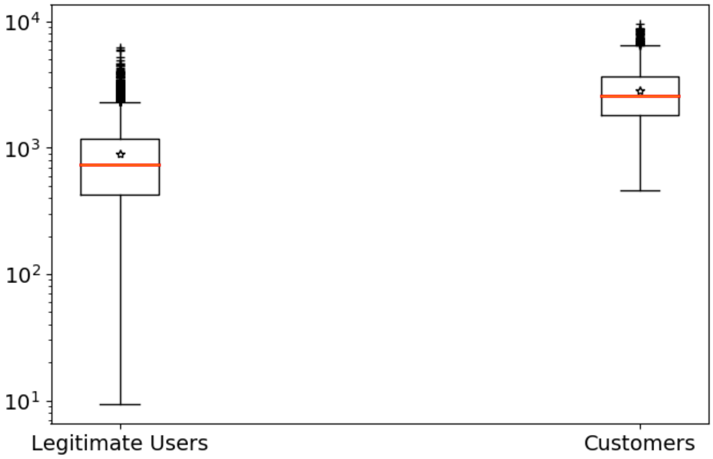


Figure 3.3.1: Comparison of Follower Distance

Chapter 4

SOCIAL RELATIONSHIP AND GEOGRAPHICAL DISTANCE

The first question we set out to answer is the question for the existence of a small-world property. This can be implied to whether the geographical distance has any correlation with the online social relationship. Also, we try to find the differences of follower distance distribution between legitimate users and customers. To answer these questions, we measure the link distances of each relationship, calculate the link probability, and obtain each distribution of relationships. Then we compare the follower distance distribution of customers and normal users.

4.1 Geographical Distance in OSNs

As the first step in addressing RQ1, we take a look at how geographical distance affects the online social relationship. To assess the correlation, we calculate the link probability that two individual online accounts are connected as a function of their geographical distance. Link probability can be calculated in many ways. However, we compute the link probability as a ratio between a real relationship and all pairs of individuals. Given a set of $G = (V, E)$, we calculate the geographical distance d_{uv} for every edge $e_{vu} \in E$ where d_{uv} is the surface distance between two geodesic points on the earth. Also, we calculate the distance between all pairs for nodes $v_i, v_j \in V,$ and i $\neq$ j. Then we bucket by intervals of 10 kilometers to compute the total number of all pairs and the number of real edge pairs. We use two different datasets to calculate link probability. One is from our dataset that we previously explained in chapter 3. In the Twitter dataset, we aggregate all Twitter accounts from our dataset and we also collect other Twitter accounts using the Twitter streaming API. The number of nodes for the Twitter dataset is 1.4 million and we use reciprocal relationship as links. The other dataset that we used is a Foursquare dataset that has 0.72 million nodes. We obtain this dataset from the open dataset, which is introduced in [59].

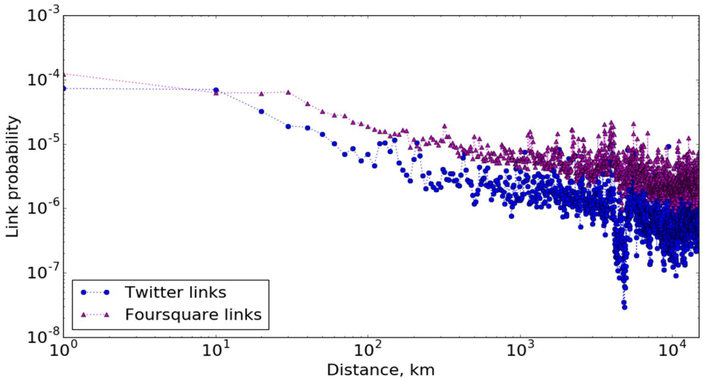


Figure 4.1.1: Link Probability for Twitter (Blue) and Foursquare (Magenta) Links.

Figure 4.1.1 shows the link probability by distance. The link probability that two nodes which are separated by distance d are friends can be modeled as $d^{-\alpha} + \varepsilon$. The link probability in Foursquare data decrease as $d^{-\alpha} + \varepsilon$, with $\alpha = 0.56$ and $\varepsilon = 9.3\mathrm{E} - 08$ and the result from a Twitter dataset has values with $\alpha = 0.55$ and $\varepsilon = 7.8\mathrm{E} - 05$. The ε value can be regarded as the background link probability of a non-geographic friendship between two users. Though two datasets show slightly different exponents and constant value due to its platform properties, we can find a similar pattern with exponential distribution. Also, other research for OSNs shows a similar pattern in [24]. The results in Figure 4.1.1 can imply that two datasets show the small-world property.

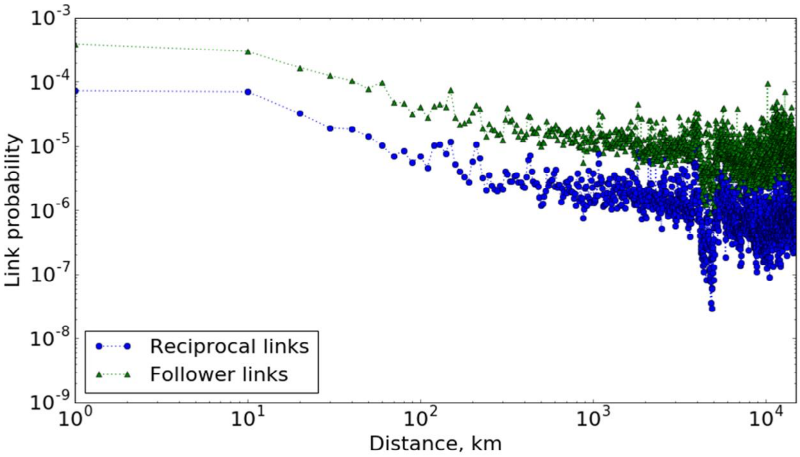


Figure 4.1.2: Link Probability of Reciprocal Links (Blue) and Follower Links (Green).

Now we compare the link probability for reciprocal links and follower links because our main focus is to find the characteristics from the user's followers. Figure 4.1.2 shows the results with the value of $\alpha = 0.34$ for follower links. The graph of the follower links shows a similar pattern to the reciprocal links graph. Therefore, it can be seen that the link probability tends to decrease as the follower links of nodes increase with similar distance.

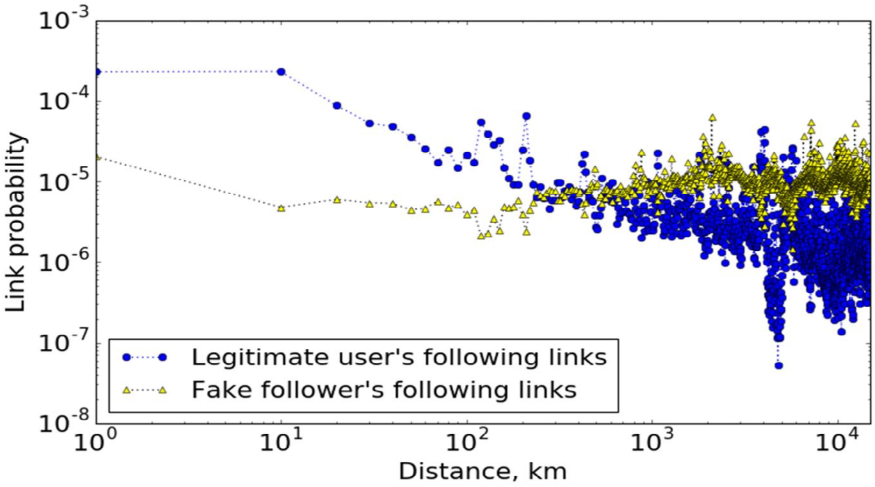
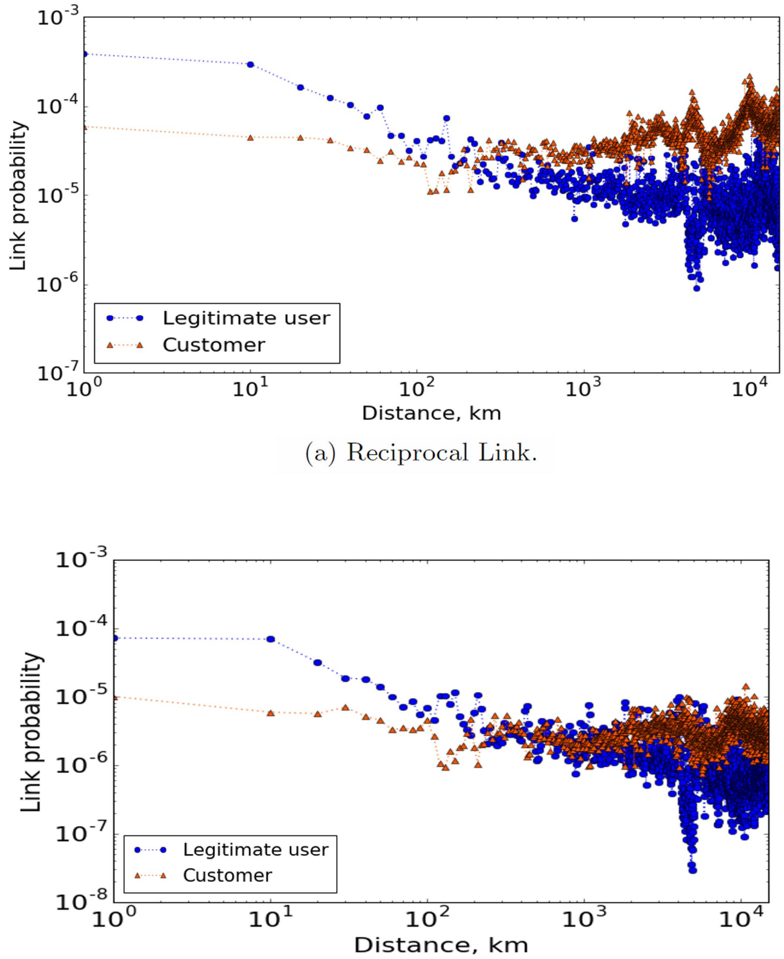


Figure 4.1.3: Following Link Probability of Legitimate User (Blue) and Fake Follower (Yellow).

We have previously discussed the legitimate user's reciprocal link and follower link characteristics. Based on this analysis we now try to investigate the fake follower's link probability and the customer's link probability to get an insight into characterizing them. To characterize the link property of fake followers, we focused on the following links of fake followers because it is a general proposition for fake followers to follow customers to get an incentive.

The result in Figure 4.1.3 shows the difference of following link probability for legitimate users and fake followers. The α values of legitimate users$'$ links and fake followers$'$ links are (0.65 and -0.06), respectively. The following link probability of legitimate users shows the similar pattern with the reciprocal link probability of legitimate users. However, the link probability for fake followers does not decrease when the distance increases. This means that the following link probability for fake followers has less correlation with distance compared with the legitimate users$'$ one and the fake follower's network does not have the small-world property.

Figure 4.1.4 shows the link probability of the legitimate user and the customer. As can be seen in the graphs, the link probability of the legitimate user and the link probability of the customer show a clear difference. The reciprocal link probability of the legitimate users decreases when the distance increases. However, the link probability of the customer tends not to decrease with distance between nodes, but rather to increase slightly as the distance increases. The link property for the follower links shows a similar pattern to the reciprocal links that means the follower link probability of the customers does not decrease when the distance increases. We can find that tendency from the α values of the customer$'$s reciprocal links $(\alpha = -0.11)$ and follower links $(\alpha = -0.24)$.



(a) Reciprocal Link.

(b) Follower Link.

Figure 4.1.4: Comparison of Link Probability.

4.2 Follower Ratio

Although in the previous section we have confirmed that the customer link property is different from the legitimate user, it is limited to classifying each user directly. Therefore, we further explore the characteristics of each user′s follower link according to distance. First, we divide the distance range into n sections then we count the number of followers in each distance section. After that, we calculate the follower ratio for each distance section. If u is a selected node, the follower ratio of u node is the ratio between the number of followers and the total number of followers, as follows:

$$R_j = \frac{N_j}{N_{Total}} \qquad (1)$$

$$\textit{where} \quad N_{Total} \ \textit{is the total number of follower for node u,}$$

$$\textit{and} \ N_j \ \textit{is the number of followers for node s in the j th distance section}$$

Figure 4.2.1 shows the average of the follower ratio according to the distance interval for 7,933 legitimate users and 2,341 customers. The follower ratio of the customers increases when the distance increases. But the follower ratio of the legitimate users does not increase sharply. For the legitimate users, the follower ratio is the highest at 0.24 in the $[10^{0.5}{\sim}10^{1.0})$ section. However, for the customers, the highest follower ratio is $[10^{3.0}{\sim}10^{3.5})$. The results reveal that the legitimate user has closer followers in a short physical distance. However, the customer′s followers are located far from the customers. This means that some of the follower relationships for customers are artificially generated, rather than in a natural human continuum.

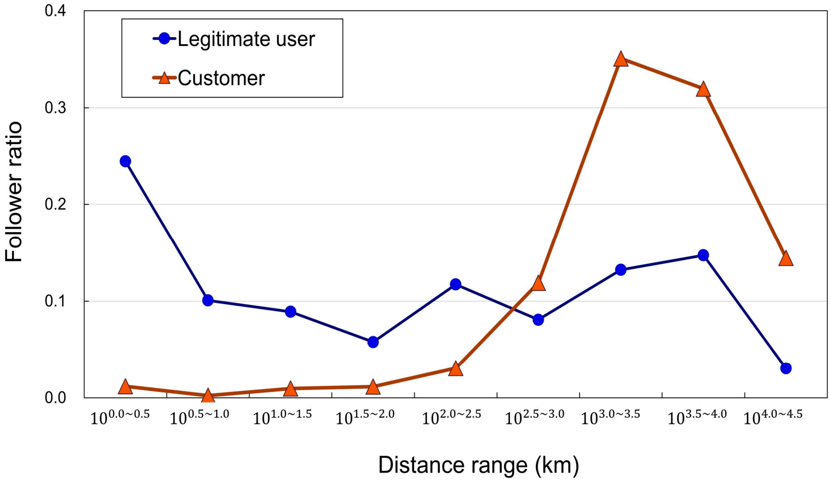
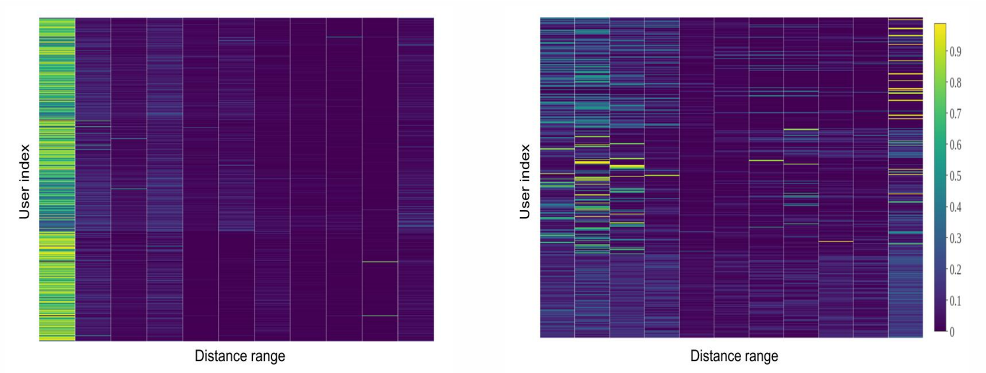


Figure 4.2.1: Follower Ratio of Legitimate Users (Blue) and Customers (Orange).

To visualize the follower ratio pattern of each user, we display the follower ratio of each user as a heat map, as shown in Figure 4.2.2. In the graph, the y-axis is divided into 15 sections in the range of (0 to 15,000) km, and the last section shows the follower ratio exceeding 15,000 km. In the case of the y-axis, it represents each user and includes all of the legitimate users and the customers in the dataset. The legitimate user graph is relatively darker than the customer graph, except for the first distance range. This means that most of the followers of each legitimate user are located within the [(0 - 1,000) km] distance range. However, in the case of the fake follower graph, it is difficult to identify the intensified region.



(a) Legitimate user

(b) Customer

Figure 4.2.2: Heatmap of Follower Ratio.

Chapter 5

Detecting Customers

Our previous study showed that the network of fake follower market does not have a small-world property and the ratio of follower′s distance for the fake customers is different from the legitimate users. The follower ratio for the legitimate users normally decreases when the distance from the centered node increase while the follower ratio of the customers increases. Thus, it would be useful to build a method that can detect the customers. In this section, we conduct a customer detection experiment based on the key features that we found in the previous analysis, then we compare our method with the other previous methods.

5.1 Key Features for Customer Detection

The key features of our method are the geographical distance between the two users and the follower ratio for each user. First, we define the geographical distance between user u and user v as follows:

$d_{uv} =$

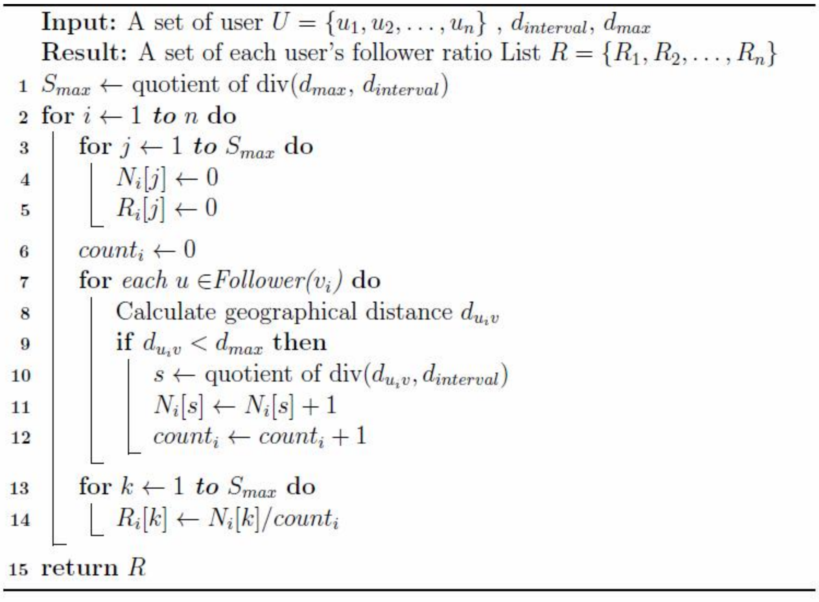
$$\text{Radious} \times \quad \arctan \frac{\sqrt{(\cos\phi_v \cdot \sin(\Delta\lambda)^2 + (\cos\phi_u \cdot \sin\phi_v - \sin\phi_u \cdot \cos\phi_v \cdot \cos(\Delta\lambda))^2}}{\sin\phi_u \cdot \sin\phi_v + \cos\phi_u \cdot \cos\phi_v \cdot \cos(\Delta\lambda)}$$

(2)

where ϕ_u, λ_u *and* ϕ_v, λ_v *is the geographical latitude and longitude of user u and v, and* $\Delta\lambda$ *is the absolute differences of* λ*. We set the* $\text{Radius} = 6372.795km$

After calculating each geographical distance, we compute the follower distance ratio for each node. Algorithm 1. is a pseudo-code to compute a follower distance distribution. Algorithm 1. computes the list of follower ratios for each distance section. We calculate the geographical distance using Equation. (2).

Algorithm 1. Generate Follower Distance Distribution



```
Input: A set of user $U = \{u_1, u_2, \ldots, u_n\}$ , $d_{interval}$, $d_{max}$
Result: A set of each user's follower ratio List $R = \{R_1, R_2, \ldots, R_n\}$
$S_{max} \leftarrow$ quotient of div($d_{max}$, $d_{interval}$)
for $i \leftarrow 1$ to $n$ do
    for $j \leftarrow 1$ to $S_{max}$ do
        $N_i[j] \leftarrow 0$
        $R_i[j] \leftarrow 0$
    $count_i \leftarrow 0$
    for each $u \in Follower(v_i)$ do
        Calculate geographical distance $d_{u_i v}$
        if $d_{u_i v} < d_{max}$ then
            $s \leftarrow$ quotient of div($d_{u_i v}, d_{interval}$)
            $N_i[s] \leftarrow N_i[s] + 1$
            $count_i \leftarrow count_i + 1$
    for $k \leftarrow 1$ to $S_{max}$ do
        $R_i[k] \leftarrow N_i[k]/count_i$
return $R$
```

5.2 Performance matrices

To evaluate the performance of our method and other classification algorithms, we build a confusion matrix as in Table 5.2.1, and use the following evaluation matrices: accuracy, precision, recall (sensitivity), false positive rate (FPR), and false negative rate (FNR), F1-score, and Matthews Correlation Coefficient (MCC) [60].

Table 5.2.1: Confusion Matrix.

		Predicted	
		Customer	Legitimate user
Classified	Customer	True Positive (TP)	False Positive (FP)
	Legitimate user	False Negative (FN)	True Negative (TN)

The accuracy is the ratio of correctly classified users and all of the users. The accuracy is calculated as Accuracy = $\frac{True\ Positive+True\ Negative}{True\ Positive+False\ Positive+False\ Negative+True\ Negative}$. Precision is defined as the truly classified customers. It is expressed as follows : Precision = $\frac{True\ Positive}{True\ Positive+False\ Positive}$. Recall is defined as the ratio of correctly classfied customers and total real customers, and it is expressed as Recall = $\frac{True\ Positive}{True\ Positive+False\ Negative}$. The false positive rate

(FPR) and false negative rate (FNR) are calculated as $\mathrm{FPR} = \frac{False\ Positive}{True\ Negative+False\ Positive}$, $\mathrm{FNR} = \frac{False\ Negative}{False\ Negative+True\ Positive}$. A false positive (FP) means that a customer was misclassified as a legitimate users, while a false negative (FN) is a legitimate users misclassified as a customer. F1-score is the harmonic mean of precision and recall, and it is expressed as $\mathrm{F1-score} = \frac{2\times Precision\ \times Recall}{Precision+Recall}$. Matthews Correlation Coefficient (MCC) is the estimator of the correlation between the predicted class and the real class of the users and it is expressed as $\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$

5.3 Experiments

As we mentioned in Chapter 3, we collected a real-world dataset for 12 months between February 2017 and January 2018. The dataset consists of 7,933 legitimate users, 24,552 fake followers, and 2,341 customers. We set the maximum distance range to be 15,000 km and split the distance range into 150 sections. To find the best classifier, we built and tested statistical models implemented in WEKA [61] with 10-fold cross-validation. We compared the performance of four classification algorithms: Logistic Regression, J48, Support Vector Machine (SVM), and RandomForest.

Figure 5.3.1 shows the classification performance evaluation using follower ratio by distance. To evaluate the performance of the classifier, we use precision, recall, and accuracy as evaluation metrics. We can find that RandomForest shows the best performance in precision (0.941), accuracy (0.980) and recall (0.974). Overall, RandomForest shows outstanding performance. From these results, we select RandomForest as our classifier and use only RandomForest for further evaluation.

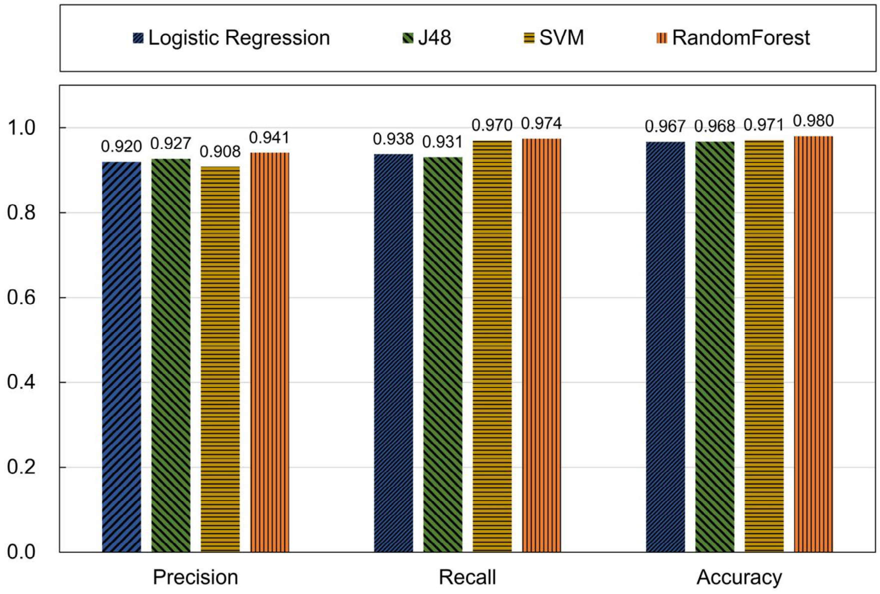


Figure 5.3.1: Performance of Logistic Regression, J47, SVM, and Random Forest.

To gain the best performance, it is important to set up the best distance interval and maximum distance to calculate the follower ratio. For this reason, we have experimented extensively with the maximum distance and distance interval. Figure 5.3.2 shows the performance with various maximum distance.

We set the distance interval as 100 km. Accuracy increases gradually as the maximum distance increases, but decreases slowly after 10,000 km. The best performance for accuracy is 0.981. The precision and recall tend to follow a similar pattern even though recall changes a little bit more sharply. The reason for showing the best performance when the maximum distance is 10,000 km is that the follower ratio at this distance section is higher than at any other distance section $[10^{3.0} \sim 10^{3.5})$, in the previous Figure 4.2.1. To achieve the best performance, we set the maximum distance at around 10,000 km for our other experiments.

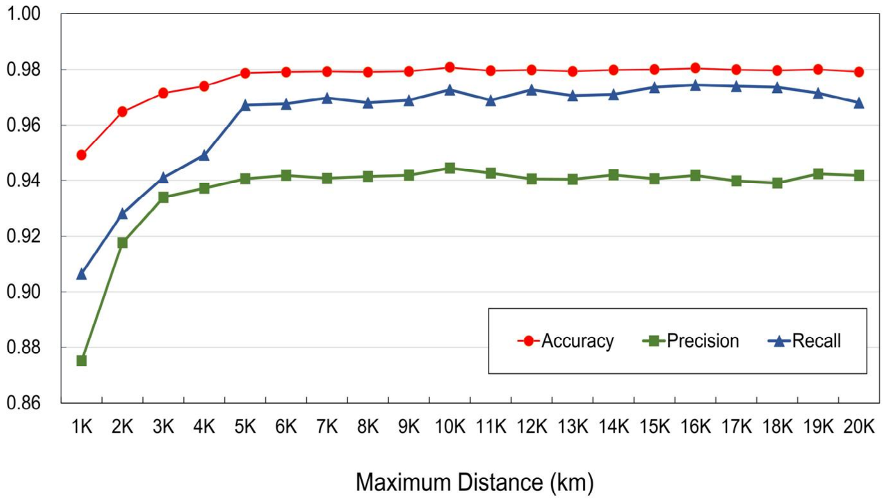


Figure 5.3.2: Experiment Results with Various Maximum Distance.

After fixing the maximum distance, we conducted the experiments with various distance intervals in [50 km, 1,000 km]. Figure 5.3.3 is the result of the performance evaluation considering various distance intervals. The performance of the algorithm deteriorates when the distance interval increases. However, the performance of the 50 km-distance interval is slightly lower than the performance of the 100 km-distance interval. This can be explained that the 100 km-distance interval can show enough difference of follower ratio between legitimate users and customers.

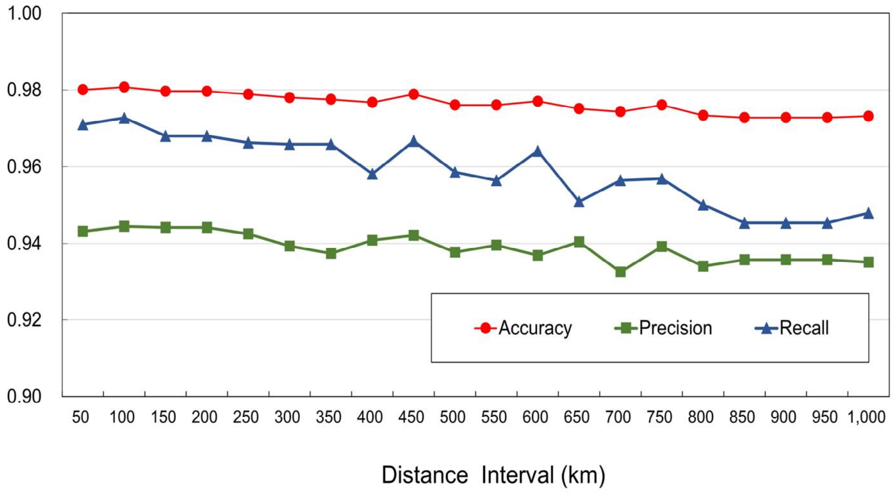


Figure 5.3.3: Experiment Results with Various Distance Intervals.

5.4 Comparison with Baseline Methods

The previous experiment showed that the RandomForest classifier achieved the best prediction results for detecting customers with a maximum distance range of 10,000 km, and 100 km distance intervals. In this section, we compare our classification method using the follower ratio with other baseline methods. The baseline methods include DetectVC [38] and CatchSync [39]. We used the output of this algorithm as a feature. DetectVC is a graph structure-based algorithm to detect voluntary followers and customers. This algorithm returns spam probability in a range between (0 and 1). The probability is propagated from the seed nodes labeled as the fake follower along with the following links, and a higher score indicates a fake follower or a customer. DetectVC can be applied when the node relationships and a priori knowledge are given. We measured the performance of DetectVC with a priori knowledge of identifying 2,000 fake followers. CatchSync introduces synchronicity and normality to capture the suspicious graph pattern. It is a HITS-based detection algorithm, so it can be applied when at least 2-hop link information for a network is given. DetectVC and CatchSync are similar to our method in terms of utilizing the graph. However, our proposed method can be applied with only 1-hop link information and its location information. Also, it does not need any prior knowledge of label information. This is the most powerful advantage for our method.

Table 5.4.1: Comparison of Experiment Results in the Entire Dataset.

	Accuracy	FPR	FNR	F1-score	MCC
Our Method	0.981	0.017	0.027	0.958	0.946
CatchSync	0.935	0.037	0.157	0.856	0.814
DetectVC	0.952	0.031	0.104	0.895	0.865

Table 5.4.1 presents the experimental results of two baseline methods and our method. Our method achieved 0.981 accuracy, 0.017 FPR, 0.027 FNR, 0.958 F1-score, and 0.946 MCC, improving up to 0.029 (= 0.981 - 0.952) accuracy, compared with the baseline methods. CatchSync and DetectVC achieved (0.935 and 0.952) accuracy, respectively. To sum up, our method using follower distance distribution outperforms better than other baseline methods.

The reason for achieving the best performance for our method is that it considers geographical distance, which is a spatial characteristic. CatchSync and DetectVC use link characteristics and prior knowledge to identify node features, but our algorithm measures the geographical distance between links using additional information in the profile. Our method requires additional node's location information, and it can be a limitation for our method. However, using the node characteristics in the network rather than using prior knowledge can be an advantage of the proposed algorithm in the real-world, because we cannot obtain all information of a network to answer whether a node is a legitimate user or not. Furthermore, even if we know some fake follower labels, we cannot identify customers in other campaigns. Also, unlike hits-based CatchSync, which requires knowing all links between nodes, this algorithm requires less effort, because it only needs the 1-hop link information of the node to be identified and the location information of the corresponding node.

Figure 5.4.1 shows the Receiver Operating Characteristic (ROC) for our method and others. Compared to other methods, our method has the highest

Area Under the Curve (AUC) at 0.9968. CatchSync and DetectVC achieved (0.9679 and 0.9811) AUC, respectively.

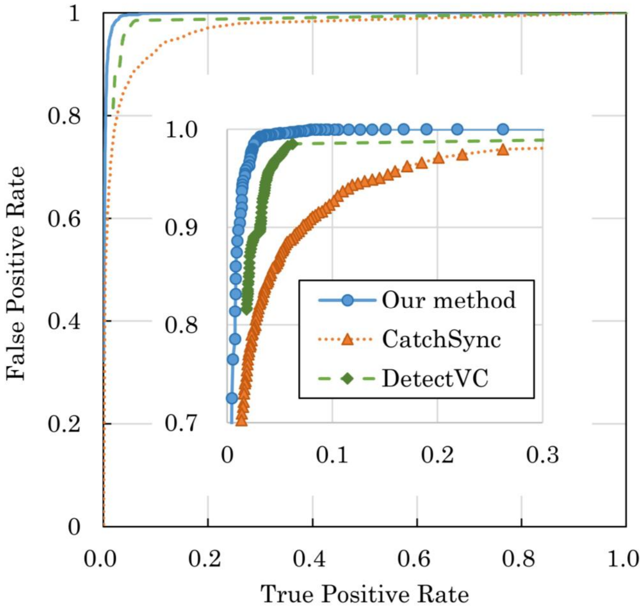


Figure 5.4.1: ROC Curve.

5.5 Comparison with Feature-based Methods

Our proposed method is a network-based method. However, each user in OSNs has additional information that can be retrieved from the profile. In this section, we examine feature-based methods in [62] to compare with our proposed method. The authors of [62] combined various features into three categories that are Class A (profile), Class B (timeline), and Class C (relationship) to detect fake followers. Since we did not collect users tweets, we evaluate Class A classifier and Class C classifier. Also, we exclude 'bot in biography', 'duplicate profile pictures', and 'default image after two months' from Class A due to the absence of additional information. We employ RandomForest classifier.

Table 5.5.1 reports the results of the classifiers using three feature set: Class A features, Class C features, and a combination of Class A and Class C features. Among the three categories, the combination of Class A and C obtains the best results with 0.935 accuracy. This performance is very similar to the results of CatchSync [39] which is one of the other baseline methods. Originally the target of [62] is fake follower accounts, but the target of our method is to detecting customers of fake followers. For this reason, the results of Feature-based methods in [62] show lower results compared to our method.

Table 5.5.1: Comparison of Experiment Results with Feature-based Methods.

	Accuracy	FPR	FNR	F1-score	MCC
Our Method	0.981	0.017	0.027	0.958	0.946
Class A	0.918	0.041	0.223	0.811	0.760
Class C	0.875	0.056	0.361	0.669	0.625
Class A and Class C	0.935	0.032	0.174	0.854	0.813

5.6 Impact of Balanced Dataset

In this research, we used an unbalanced dataset that has a different number of customers and legitimate users. If the dataset is unbalanced, it can influence the result of the classifier. For this reason, we performed additional experiments with a balanced dataset. We build a balanced dataset that includes 2,000 legitimate users and 2,000 customers. We used random under-sampling so that each user in the balanced dataset is selected randomly from the entire dataset. Table 5.6.1 shows the experimental results with a balanced dataset. This shows that the performance of our proposed method is slightly changed. Our method outperformed CatchSync and DetectVC, achieving 0.981 accuracy, 0.032 FPR, 0.008 FNR, 0.981 F1-score, and 0.961 MCC. Compared with a result using an entire dataset, the accuracy is not changed as 0.981. The FPR increased while the FNR decreased. In customer or spam detection, it is important to prevent the misclassified legitimate user. For this reason, we used the unbalanced dataset in our major experiments.

Table 5.6.1: Comparison of Experiment Results in the Balanced Dataset.

	Accuracy	FPR	FNR	F1-score	MCC
Our Method	0.981	0.032	0.008	0.981	0.961
CatchSync	0.920	0.079	0.082	0.920	0.840

DetectVC	0.954	0.046	0.047	0.951	0.908

5.7 Fake Follower Detection

This research was conducted focusing on customer detection using the follower link. However, as confirmed in Chapter 4, the following links of fake followers showed a different tendency from the following links of legitimate users. In this section, we perform fake follower detection in this regard. First, unlike customer detection, we focused on the following links in this experiment, so we used the following link ratio as a feature. Figure 5.7.1 confirms the change of the following ratio according to distance.

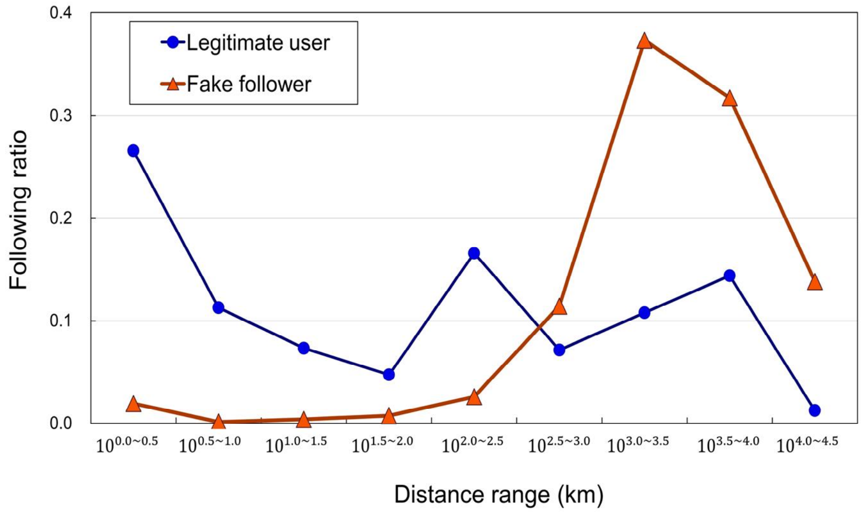


Figure 5.7.1: Following Ratio for Legitimate Users and Fake Followers.

We perform the experiment with a balanced dataset that has 2,000 legitimate users and 2,000 fake followers. Table 5.7.1 shows the result. Our method achieved 0.987 accuracy, 0.018 FPR, 0.009 FNR, and 0.974 MCC, improving to 0.096 (= 0.987 − 0.891) accuracy, compared with the baseline methods. Surprisingly, the improvement for fake follower detection is greater than the improvement for customer detection that improves accuracy to 0.069 more than the other methods. The improvement of the proposed algorithm in the fake follower detection is significant, because the fake followers make the following relationship only for the delivery of orders, rather than making meaningless following links to improve personal relationships. Therefore, unlike legitimate users, the following ratio varies greatly depending on the distance. Thus, fake follower detection is an easier problem to solve than customer detection, because some legitimate users follow customers due to customer's popularity or postings.

Table 5.7.1: Experiment Results for Fake Follower Detection.

	Accuracy	FPR	FNR	F1-score	MCC
Our Method	0.987	0.018	0.009	0.987	0.974
CatchSync	0.891	0.115	0.103	0.892	0.783
DetectVC	0.879	0.126	0.116	0.879	0.787

Chapter 6

Future Work

6.1 The Absence of Location Information

Many OSNs websites such as Twitter and Facebook collect user's home location or activity location. Also, mobile devices and location-based service initiate a new type of OSNs which is called location-based social networks(LBSN). Our proposed method presented in this paper can identify users who have arbitrary relations by using relationship and location information in this environment. However, the location information of the user or activity may not be mandatory based on each platform. Thus we exclude the users who have no location information from our experiments.

Even though our method did not deal with the absence of location information, many other researches have suggested methods to solve this problem. [63] proposed an unsupervised framework to predict named entity city-level location using the tweet content, user profile information, and the tweet GPS coordinate information. [64] proposed an end-to-end neural network to predict the location of a tweet using the tweet content, user profile information such as creation time, UTC offset, time zone, and account creation time. [65] extracted Twitter user's location from a set of location-related word that is contained in user's tweet messages. [66, 67] identify location indicative words by adapting a feature selection method. While there are prior works [63, 64, 65, 66, 67] exploit user profile, tweet context, or tweet metadata, [68, 69] have utilized the 1-hop network relationship. The authors in [68] proposed the method to predict the user profile location based on the geographic distribution of the user's friends. [69] presented an unsupervised approach to improve the accuracy of predicting location by exploiting check-in data, spatial-temporal feature, and social relationship.

If we adapt location inferring method to measuring 1-hop link distance between users who have no location information, our proposed method will achieve higher performance. But these inferring methods are fundamentally only available when OSNs provide location information. Therefore, OSNs which do not have location information is limited to utilize our proposed method. However, if users are grouped through graph clustering or community detection method [70, 71, 72], we can measure the weights between 1-hop link using the similarity of groups or relation between other groups and the weights can replace the distance attribute.

6.2 Hybrid Detection Method with Link Ratio and Profile Information

The proposed method utilizes a small-world property of OSNs to detect customers and fake followers that do not have the small-world property. We used only the user's location information for detection. However, there is other information such as the user's profile and activities. Therefore, the hybrid method with the combination of these additional information and our method can achieve better performance than other methods. In this section, we have performed a simple experiment to show the effectiveness of the hybrid method. All additional profile features used in this experiment can be extracted from the user's profile metadata. Experiments were performed using RadomForest Classifier. The results are shown in Table 6.2.1. It showed a slight increase in performance when using the profile features. In other words, we can verify that the hybrid method can achieve better performance through combination of profile features and relationship information. We believe that if we design more sophisticated algorithms, we can greatly improve performance.

Table 6.2.1: Experiment Results for Hybrid Methods.

Link Ratio	Additional Profile Features	Accuracy	MCC
○	×	0.981	0.946
○	Follower Count, Followee Count	0.982	0.948
○	Follower Count, Followee Count, Status Count, Favorite Count	0.981	0.946
○	Has name, Has image, Has address, Has biography , Belongs to a list	0.980	0.945
○	Followers Count, Followee Count, Has name, Has Image, Has Address, Has Biography, Belongs to a List	0.982	0.949
×	Followers Count, Followee Count, Has name, Has Image, Has Address, Has Biography, Belongs to a List	0.967	0.905

Chapter 7

Conclusion

On Twitter, one of the most popular social networking services (SNSs), a new kind of spamming strategy has emerged known as fake followers. The goal of this paper is to verify whether a fake follower market network has a small-world property and to classify customer of fake followers by utilizing social network properties and user profile properties. To solve this problem, we proposed geographical distance based fake customer detection method.

These approaches analyze and exploit social network properties such as follower link and distance property. We conducted large-scale experiments on real Twitter datasets. The results from analyzing distance based follower ratio support our assumption that a fake customer′s follower distance ratio is different from a legitimate user′s follower distance ratio. We compared our approaches to DetectVC and CatchSync, the HITs-based representative algorithm of the fake follower and customer detection. Also, we compared our approaches to feature-based algorithm. Our proposed method was found to be the most competitive and superior approach.

In conclusion, with a high proportion of accuracy (98.1%), our approaches are very secure and practical mechanisms that can be applied as fake customer detection systems.

Bibliography

[1] Travers, Jeffrey, and Stanley Milgram. "The small world problem." Phychology Today 1.1 (1967): 61-67.

[2] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of ‘small-world’networks." nature 393.6684 (1998): 440.

[3] Adamic, Lada A. "The small world web." International Conference on Theory and Practice of Digital Libraries. Springer, Berlin, Heidelberg, 1999.

[4] Backstrom, Lars, et al. "Four degrees of separation." Proceedings of the 4th Annual ACM Web Science Conference. ACM, 2012.

[5] Kleinberg, Jon. The small-world phenomenon: An algorithmic perspective. Cornell University, 1999.

[6] Liben-Nowell, David, et al. "Geographic routing in social networks." Proceedings of the National Academy of Sciences 102.33 (2005): 11623-11628.

[7] Java, Akshay, et al. "Why we twitter: understanding microblogging usage and communities." Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007.

[8] Zhang, Hui, Ashish Goel, and Ramesh Govindan. "Using the small-world model to improve freenet performance." INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. Vol. 3. IEEE, 2002.

[9] Latora, Vito, and Massimo Marchiori. "Is the Boston subway a small-world network?." Physica A: Statistical Mechanics and its Applications 314.1-4 (2002): 109-113.

[10] Mislove, Alan, et al. "Measurement and analysis of online social networks." Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007.

[11] Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. AcM, 2010.

[12] Hong, Sounman, and Daniel Nadler. "Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience." Government Information Quarterly 29.4 (2012): 455-461.

[13] Edwards, Chad, et al. "How much Klout do you have… A test of system generated cues on source credibility." Computers in Human Behavior 29.5 (2013): A12-A16.

[14] Hutto, Clayton J., Sarita Yardi, and Eric Gilbert. "A longitudinal study of follow predictors on twitter." Proceedings of the sigchi conference on human factors in computing systems. ACM, 2013.

[15] Krombholz, Katharina, Dieter Merkl, and Edgar Weippl. "Fake identities in social media: A case study on the sustainability of the facebook business model." Journal of Service Science Research 4.2 (2012): 175-212.

[16] Luca, Michael, and Georgios Zervas. "Fake it till you make it: Reputation, competition, and Yelp review fraud." Management Science 62.12 (2016): 3412-3427.

[17] Stringhini, Gianluca, et al. "Poultry markets: on the underground economy of twitter followers." ACM SIGCOMM Computer Communication Review 42.4 (2012): 527-532.

[18] De Veirman, Marijke, Veroline Cauberghe, and Liselot Hudders. "Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude." International Journal of Advertising 36.5 (2017): 798-828.

[19] Boyd, Danah M., and Nicole B. Ellison. "Social network sites: Definition, history, and scholarship." Journal of computer□mediated Communication 13.1 (2007): 210-230.

[20] Bhat, Sajid Yousuf, and Muhammad Abulaish. "Community-based features for identifying spammers in online social networks." Advances in

Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013.

[21] Shi, Xiaolin, Belle Tseng, and L. Adamic. "Looking at the blogosphere topology through different lenses." Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007). Vol. 1001. 2007.

[22] Hung, Shiu-Wan, and An-Pang Wang. "Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network." Scientometrics 82.1 (2010): 121-134.

[23] Thomas, Kurt, et al. "Suspended accounts in retrospect: an analysis of twitter spam." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011.

[24] Cho, Eunjoon, Seth A. Myers, and Jure Leskovec. "Friendship and mobility: user movement in location-based social networks." Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011.

[25] Gao, Hongyu, et al. "Detecting and characterizing social spam campaigns." Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, 2010.

[26] Thomas, Kurt, et al. "Design and evaluation of a real-time url spam filtering service." Security and Privacy (SP), 2011 IEEE Symposium on. IEEE, 2011.

[27] Grier, Chris, et al. "@ spam: the underground on 140 characters or less." Proceedings of the 17th ACM conference on Computer and communications security. ACM, 2010.

[28] Stringhini, Gianluca, Christopher Kruegel, and Giovanni Vigna. "Detecting spammers on social networks." Proceedings of the 26th annual computer security applications conference. ACM, 2010.

[29] Gao, Hongyu, et al. "Spam ain't as diverse as it seems: throttling OSN spam with templates underneath." Proceedings of the 30th Annual Computer Security Applications Conference. ACM, 2014.

[30] Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. "Spam filtering with naive bayes-which naive bayes?." CEAS. Vol. 17. 2006.

[31] Sculley, David, and Gabriel M. Wachman. "Relaxed online SVMs for spam filtering." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

[32] Egele, Manuel, et al. "Compa: Detecting compromised accounts on social networks." NDSS. 2013.

[33] Benevenuto, Fabricio, et al. "Detecting spammers on twitter." Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). Vol. 6. No. 2010. 2010.

[34] Martinez-Romo, Juan, and Lourdes Araujo. "Detecting malicious tweets in trending topics using a statistical analysis of language." Expert Systems with Applications 40.8 (2013): 2992-3000.

[35] Page, Lawrence, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

[36] Kleinberg, Jon M., et al. "The web as a graph: measurements, models, and methods." International Computing and Combinatorics Conference. Springer, Berlin, Heidelberg, 1999.

[37] Gyöngyi, Zoltán, Hector Garcia-Molina, and Jan Pedersen. "Combating web spam with trustrank." Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. VLDB Endowment, 2004.

[38] Liu, Yuli, et al. "Pay Me and I'll Follow You: Detection of Crowdturfing Following Activities in Microblog Environment." IJCAI. 2016.

[39] Jiang, Meng, et al. "Catchsync: catching synchronized behavior in large directed graphs." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

[40] Ghosh, Saptarshi, et al. "Understanding and combating link farming in the twitter social network." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[41] Cao, Qiang, et al. "Aiding the detection of fake accounts in large scale social online services." Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012.

[42] Yu, Haifeng, et al. "Sybillimit: A near-optimal social network defense against sybil attacks." IEEE/ACM Transactions on Networking (ToN) 18.3 (2010): 885-898.

[43] Danezis, George, and Prateek Mittal. "SybilInfer: Detecting Sybil Nodes using Social Networks." NDSS. 2009.

[44] Hooi, Bryan, et al. "Fraudar: Bounding graph fraud in the face of camouflage." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[45] Jiang, Meng, et al. "Inferring lockstep behavior from connectivity pattern in large graphs." Knowledge and Information Systems 48.2 (2016): 399-428.

[46] Beutel, Alex, et al. "Copycatch: stopping group attacks by spotting lockstep behavior in social networks." Proceedings of the 22nd international conference on World Wide Web. ACM, 2013.

[47] Jiang, Meng, et al. "Spotting suspicious behaviors in multimodal data: A general metric and algorithms." IEEE Transactions on Knowledge and Data Engineering 28.8 (2016): 2187-2200.

[48] Wu, Chih-Hung. "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks." Expert Systems with Applications 36.3 (2009): 4321-4330.

[49] Wang, Gang, et al. "You Are How You Click: Clickstream Analysis for Sybil Detection." USENIX Security Symposium. Vol. 9. 2013.

[50] Cresci, Stefano, et al. "DNA-inspired online behavioral modeling and its application to spambot detection." IEEE Intelligent Systems 31.5 (2016): 58-64.

[51] Cresci, Stefano, et al. "Social Fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling." IEEE Transactions on Dependable and Secure Computing (2017).

[52] Chavoshi, Nikan, Hossein Hamooni, and Abdullah Mueen. "DeBot: Twitter Bot Detection via Warped Correlation." ICDM. 2016.

[53] Stringhini, Gianluca, et al. "Follow the green: growth and dynamics in twitter follower markets." Proceedings of the 2013 conference on Internet measurement conference. ACM, 2013.

[54] T. Inc., Api reference index, accessed: 2017-10-01. URL https://developer.twitter.com/en/docs/api-reference-index

[55] G. Developers, Developer guide, accessed: 2017-10-01. URL https://developers.google.com/maps/documentation/geocoding/intro

[56] Thomas, Kurt, et al. "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse." USENIX Security Symposium. 2013.

[57] Cresci, Stefano, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017.

[58] Hewson, Claire, and Tom Buchanan. "Ethics guidelines for internet-mediated research." The British Psychological Society, 2017.

[59] Bao, Jie, Yu Zheng, and Mohamed F. Mokbel. "Location-based and preference-aware recommendation using sparse geo-social networking data." Proceedings of the 20th international conference on advances in geographic information systems. ACM, 2012.

[60] Baldi, Pierre, et al. "Assessing the accuracy of prediction algorithms for classification: an overview." Bioinformatics 16.5 (2000): 412-424.

[61] Holmes, Geoffrey, Andrew Donkin, and Ian H. Witten. "Weka: A machine learning workbench." Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on. IEEE, 1994.

[62] Cresci, Stefano, et al. "Fame for sale: efficient detection of fake Twitter followers." Decision Support Systems 80 (2015): 56-71.

[63] Shen, Wei, Yinan Liu, and Jianyong Wang. "Predicting Named Entity Location Using Twitter." 2018 IEEE 34th International Conference on Data Engineering (ICDE). IEEE, 2018.

[64] Lau, et al. “End-to-end Network for Twitter Geolocation Prediction and Hashing.” In Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)

[65] Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[66] Han, Bo, Paul Cook, and Timothy Baldwin. "Geolocation prediction in social media data by finding location indicative words." Proceedings of COLING 2012 (2012): 1045-1062.

[67] Han, Bo, Paul Cook, and Timothy Baldwin. "A stacking-based approach to twitter user geolocation prediction." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2013.

[68] Liu, Zhi, and Yan Huang. "Closeness and structure of friends help to estimate user locations." International Conference on Database Systems for Advanced Applications. Springer, Cham, 2016.

[69] Huang, Chao, and Dong Wang. "Exploiting spatial-temporal-social constraints for localness inference using online social media." Proceedings of

the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Press, 2016.

[70] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.

[71] Shun, Julian, et al. "Parallel local graph clustering." Proceedings of the VLDB Endowment 9.12 (2016): 1041-1052.

[72] Ruan, Yiye, David Fuhry, and Srinivasan Parthasarathy. "Efficient community detection in large networks using content and links." Proceedings of the 22nd international conference on World Wide Web. ACM, 2013.

국문초록

장보연
컴퓨터공학부
서울대학교 대학원

온라인 소셜 네트워크가 발달함에 따라 트위터, 페이스북, 인스타그램과 같은 소셜 미디어에서의 평판은 이제 현실 세계에서도 그 사람의 파워로 간주되는 경향이 있다. 다수의 친구나 팔로워를 갖고 있는 사용자는 자신의 활동들을 게시하고 상태를 업데이트하며 여러 내용을 포스팅 하거나 트윗 혹은 리트윗하면서 자신의 팔로워에게 더 많은 영향을 줄 수 있다. 일반적으로 온라인 소셜 네트워크 사용자의 영향력은 친구 또는 팔로워의 수와 밀접한 관계가 있다. 즉 친구나 팔로워가 많은 사용자가 더 큰 영향력을 갖고있다 할 수 있다. 영향력을 높이기 위해 팔로워 수를 높이는 방법에는 여러가지가 있지만 단기간에 보다 빠르게 팔로워를 증가시키고 싶은 사용자들은 불법적인 시장의 유혹에 쉽게 빠져들 수 있다. 이러한 시장을 통해 팔로워를 구매한 구매자들은 건전한 온라인 소셜 네트워크의 생태계를 저해하지만 기존의 방법으로는 이 구매자들을 식별하기 제한된다. 가짜 팔로워를 구매하는 고객들은 일반적으로 온라인 소셜 네트워크 상에서 활발하게 활동을 하는 사용자들이며, 따라서 이들을 유명인이나 활동적인 일반 사용자와 구별하기에 어렵기 때문이다. 하지만 소셜 네트워크 특성상 가짜 팔로워나 구매자가 인위적으로 쉽게 조작할 수 없는 Small-world property 와 같은 고유의 특징들이 나타난다.

Small-world property 는 최단경로특성과 집단화 수 특성을 나타내며, 이 네트워크는 사람들간의 짧은 체인으로 연결되는 현상을 보인다. 기존 연구에서는 네트워크내 피참조도나 참조도, 중심성 등을 활용하여 악의적인 사용자를 식별하거나, 사용자의 고유 특징을 나타내는 프로파일을 활용하기도 하였다. 하지만 가짜 팔로워나 구매자를 식별하는데 small-world property 를 활용한 연구는 아직 제시된 바가 없다.

따라서 본 연구에서는 이 특성을 활용하여 가짜 팔로워를 구매하는 사용자를 식별하는 방법을 제안하고 있다. 본 연구에서는 일반 사용자와 가짜 팔로워 구매자의 1Hop 네트워크내 링크의 지리적 거리가 분포가 small-world property 를 따르는지 여부를 확인하였으며, 가짜 팔로워 구매자들은 이를 따르지 않음에 착안하여 이들을 식별하는 방법을 제안하였다. 또한 수집한 트위터 데이터셋에서의 실험 결과를 통해 기존에 제시된 방법들보다 높은 성능을 보여주고 있다.

주요어 : 온라인 소셜 네트워크, 보안, 트위터, 가짜 팔로워, 스팸 탐지, 지리적 거리

학 번 : 2015-30272