5-2016

# Fake Likers Detection on Facebook

Prudhvi Ratna Badri Satya
*Utah State University*

# FAKE LIKERS DETECTION ON FACEBOOK

by

Prudhvi Ratna Badri Satya

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Science

Approved:

| | |
|---|---|
| Dr. Kyumin Lee<br>Major Professor | Dr. Curtis Dyreson<br>Committee Member |
| Dr. Tung Nguyen<br>Committee Member | Dr. Mark R. McLellan<br>Vice President for Research and<br>Dean of the School of Graduate Studies |

UTAH STATE UNIVERSITY
Logan, Utah

2016

## ABSTRACT

Fake Likers Detection on Facebook

by

Prudhvi Ratna Badri Satya, Master of Science

Utah State University, 2016

Major Professor: Dr. Kyumin Lee<br>
Department: Computer Science

In online social networking sites, gaining popularity has become important. The more popular a company is, the more profits it can make. A way to measure a company's popularity is to check how many likes it has (e.g., the company's number of likes in Facebook). To instantly and artificially increase the number of likes, some companies and business people began hiring crowd workers (aka fake likers) who send likes to a targeted page and earn money. Unfortunately, little is known about characteristics of the fake likers and how to identify them. To uncover fake likers in online social networks, in this work we (i) collect profiles of fake likers and legitimate likers by using linkage and honeypot approaches, (ii) analyze characteristics of fake likers and legitimate likers, (iii) propose and develop a fake liker detection approach, and (iv) thoroughly evaluate its performance against three baseline methods and under two attack models. Our experimental results show that our classification model significantly outperformed the baseline methods, achieving 87.1% accuracy and 0.1 false positive rate and 0.14 false negative rate.

(37 pages)

# PUBLIC ABSTRACT

Fake Likers Detection on Facebook

Prudhvi Ratna Badri Satya

As billions of users have used online social networks such as Facebook, Twitter, and Instagram, many online services have begun adapting signals of the popularity. For example, search engines rank popular users posts in the top results. A post of a popular user is automatically delivered to the followers or likers. Popular companies in online social networks also get good reputation from people. The popularity on Facebook is measured based on the number of likes in a page. Unfortunately, some crowdsourcing websites (e.g.Microworkers.com) or supply-driven marketplaces (e.g.Fiverr.com) began offering fake liking services in which these crowd workers sent fake likes to Facebook page owners and earned money from the page owners. This manipulation has been degrading information trust and threatening trustworthiness of the online social network. To stop these fake liking activities, in this thesis, we (i) collected profiles of fake likers and legitimate users from Facebook, (ii) analyzed characteristics of fake likers and legitimate users, and (iii) proposed and developed fake liker detection approaches. Our experimental results show that our approaches effectively identified fake likers, achieving 87.1% accuracy and 0.1 false positive rate and 0.14 false negative rate.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

Figure Page

# CHAPTER 1
# INTRODUCTION

Billions of people use online social networking sites (e.g., Facebook, Twitter, Instagram) to share information regarding their daily life, engage with friends and express their opinions (e.g., reviews, votes, likes) about products, posts and companies. On the other hand, companies create accounts and pages to engage with users and interact with them, expecting increasing credibility about the companies, getting more positive responses from the users and making more profits.

One of the popular activities in online social networks is “liking” an object such as a post, photo, video or company’s page (e.g., Nike page). Companies want to increase the number of likes associated with their pages to increase popularity. The liking function can be a good marketing tool since companies can recognize who are interested in the companies. In addition, newly created posts are automatically delivered to the likers similar with Twitter’s following function. Another benefit is if a company’s page has many likes, the page and associated posts will be considered important by search engines and will be located in top results.

Unfortunately, there are companies and business people which/who artificially increase the number of likes by hiring workers from crowdsourcing sites (e.g., Microworkers, Rapid-Workers) and seller-driven marketplaces (e.g., Fiverr). By paying $5, they get hundreds of fake likes. Similar problems like buying fake reviews, followers and manipulating keyword searches were reported by news media and researchers [1–8]. If we do not solve this manipulation, it will degrade information trust and threaten trustworthiness of online social networks and the Web. In general, detecting fake likers is a harder problem than detecting spammers/sybils, anomalous users and bots because some of the fake likers may use their personal accounts which make harder compared with spammers or anomalous users, who aggressively post advertisements or spam URLs and so on.

Recently, researchers [9–11] began studying the fake liker problem, but their approaches require temporal data (e.g., what pages each user liked when, and series of daily profile snapshots for each user) which is relatively expensive information compared with static data (i.e., a snapshot of user profile). Their approaches also require input parameters, finding optimal parameter values are hard, and their fake liker detection rate is low.

To complement the prior work, in this thesis, we are interested in answering following research questions: Did *fake likers*, who liked targeted pages because of money, use their own personal accounts or dummy accounts for fake liking activities? What social networks did they form? Did they behave differently from legitimate likers? Is it possible to automatically detect these fake likers? If online social networking service providers detect these fake likers, their accounts will be suspended, and information trust will increase.

To improve information quality by filtering fake likers, in this work, we used novel linkage and honeypot methods to collect Facebook fake likers' information from two sources – a popular crowdsourcing site and a seller-driven marketplace. Then we analyzed characteristics of these fake likers and compared them with legitimate likers. In addition, we proposed and evaluated a novel fake liker detection method based on less expensive data.

Concretely, we make following contributions in this research :

- First, we collected over 13,000 fake likers and legitimate likers' full profiles by using novel linkage and honeypot methods. Then, we analyzed how fake likers from two type of sources – a crowdsourcing site and a sell-driven marketplace – were different.
- Second, We analyzed characteristics of fake likers and legitimate likers in terms of page liking behaviors, categories of liked pages, posting activities and social interactions.
- Third, we proposed and extracted 5 types of feature sets toward building fake liker classifiers.
- Finally, our proposed approach significantly outperformed three baseline methods. Under individual and coordinated attack models, our approach consistently and robustly identified fake likers.

# CHAPTER 2

# RELATED WORK

In this chapter, we summarize research work related to crowdturfing, spammers, like farms and anomalous users in online social networks.

Researchers [12, 13] deployed honeypots to uncover spammers on social networks likes Twitter and Facebook, and developed machine learning based classification models to detect the spammers. Lee et al. [14] performed comprehensive analysis of Fiverr to reveal the existence of crowdturfing tasks in it, and developed a crowdturfing tasks detection method to remove such tasks in order to prevent their activity. Boshmaf et al. [15] used user-level activities and graph-level structures for effective fake account detection in online social networks. Song et al. [16] proposed a method to detect target objects of crowdturfing tasks (e.g., post, page, and URL) on Twitter.

De Cristofaro et al. [2] presented a comparative study of Facebook ads and like farms by analyzing the demographic, temporal and social characteristics of likers. They concluded that like farms followed two modi operandi in providing likes: one was instant burst like patterns and the other was a slow pattern which mimics legitimate user behavior on Facebook. However, they only performed the analysis among various like farms and did not investigate or detect fake likers' accounts. Compared with the work, we collected profiles of both legitimate and fake likers and analyzed characteristics of them.

Viswanath et al. [11] used the principal component analysis technique to identify the anomalous user behavior from normal user behavior in order to detect anomalous accounts on Facebook. They did not provide any analysis to show the differences between the users of various like platforms. We provided an in-depth analysis of user behavior on various like platforms with legitimate user behavior. Temporal features they used were very expensive, containing when a user liked which page and associated page category information.

CopyCatch [9], an anti-fraudulent tool deployed at Facebook, detected undesirable accounts generating fake likes by extracting near bipartite cores from users and Facebook pages based on likes and liked time. SynchroTrap [10] detected groups of malicious accounts by running hierarchical clustering based on their liking similarity during the same time interval. However, in our study, we found that existing Facebook security system using both methods did not filter fake likes and likers well, and did not remove them even two months later.

Compared with the previous research work, we analyzed (i) how fake likers from two different sources (i.e.,Microworkers and Fiverr) were different, and (ii) characteristics of fake likers and legitimate likers on Facebook. Based on the analysis, we proposed and extracted less expensive features toward building fake liker classifiers. Under two attack simulations, our approach robustly and consistently identified fake likers. Our research will complement the existing research base.

# CHAPTER 3
# PROBLEM DEFINITION AND DATASET

In this chapter, we discuss the problem definition and the data collection process involved in the collection of fake and legitimate likers ground truth data.

## 3.1 Problem Definition

In social networking sites like Facebook, each user $u_i$ has a profile $p_i$ consisting of bio, status messages, a list of friends and so on. A **fake liker detection problem** is to predict whether $u_i$ is a fake liker or a legitimate liker through a classifier $c$ when $p_i$ is given. A classifier

$$c : u_i \rightarrow \{fake\, liker, legitimate\, liker\}$$

approximates whether $u_i$ is a fake liker.

To build a classifier $c$, we extract features from each profile $p_i$. In this work, a fake liker is a user who performed fake liking activity at least twice on Facebook with receiving compensation/money from requesters. Fake/paid liking activities are prohibited on Facebook [17, 18].

## 3.2 Data Collection

In the fake liker detection study, the first step is to collect a dataset consisting of fake liker profiles and legitimate liker profiles. First, we collected fake liker profiles by linking Fiverr and Microworkers users to Facebook users. In the meantime, we selected and collected legitimate liker profiles from a randomly selected Facebook user pool and members of conference and research groups on Facebook. Figure 3.1 shows a high-level overview of our data collection process. From each of the selected users, we crawled only publicly available profile information such as posts, liked pages, friends, longevity, gender,

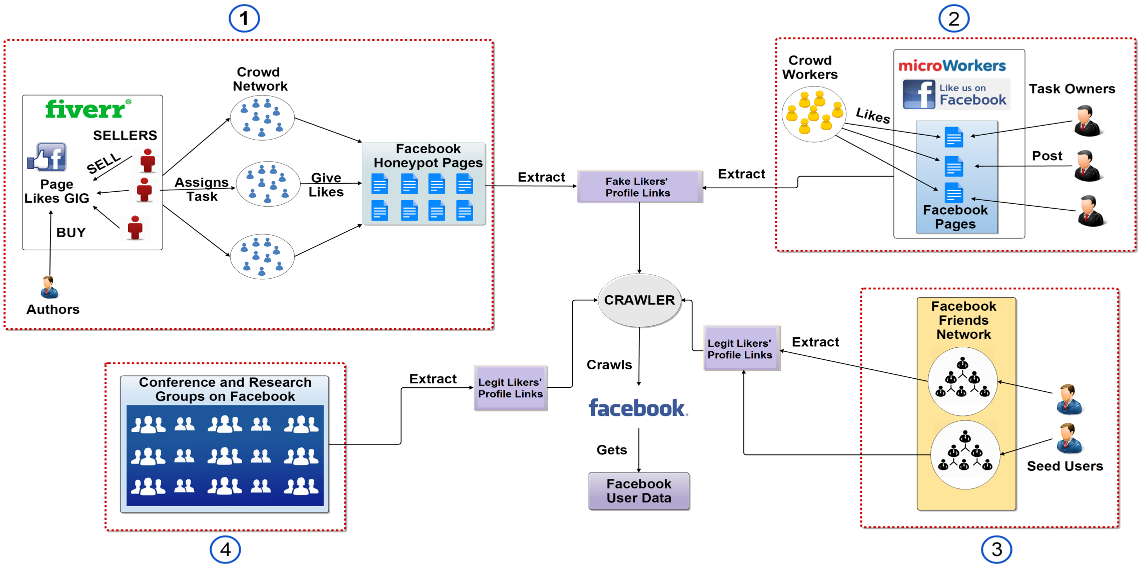


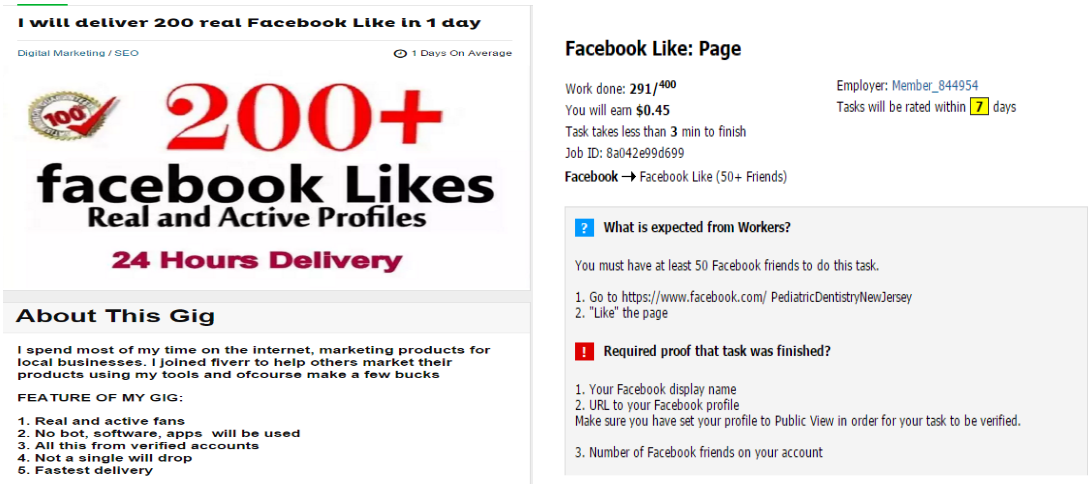
Fig. 3.1: Collecting fake likers from Fiverr and Microworkers, and legitimate likers from a random pool and conference groups.

social interaction received by posts such as comments, likes, and shares, and bio such as affiliation, education, location and etc. Next, we describe our data collection approaches in detail.

### 3.2.1 Fake Liker Profiles

Detecting fake likers is a harder problem than detecting spammers because most fake liking activities are hard to recognize without direct evidence of requested fake liking tasks. To identify the evidence, we chose Fiverr, a seller-driven marketplace, and Microworkers, a requester-driven crowdsourcing platform, since other researchers [14, 19] reported that unethical sellers and requesters on these marketplaces and crowdsourcing platforms created malicious tasks (e.g., fake likes, fake followers) targeting social networking sites.

Sellers in Fiverr create tasks/services as shown in Figure 3.2(a). We selected 10 sellers who promised passing fake likes to any Facebook page. Then, we created 10 honeypot Facebook pages, and intentionally left them blank without posting any update, image and etc. To make sure not receiving any false like from legitimate Facebook users, we clearly displayed "This is a fake page please don't like this" in the about section. Each honeypot



I will deliver 200 real Facebook Like in 1 day

Digital Marketing / SEO

1 Days On Average

About This Gig

I spend most of my time on the internet, marketing products for local businesses. I joined fiverr to help others market their products using my tools and ofcourse make a few bucks

FEATURE OF MY GIG:

1. Real and active fans
2. No bot, software, apps  will be used
3. All this from verified accounts
4. Not a single will drop
5. Fastest delivery

(a) A Facebook like task in Fiverr

Facebook Like: Page

Work done: **291**/$^{400}$
You will earn **$0.45**
Task takes less than **3** min to finish
Job ID: 8a042e99d699
**Facebook** → Facebook Like (50+ Friends)

Employer: Member_844954
Tasks will be rated within 7 days

? **What is expected from Workers?**

You must have at least 50 Facebook friends to do this task.

1. Go to https://www.facebook.com/ PediatricDentistryNewJersey
2. "Like" the page

! **Required proof that task was finished?**

1. Your Facebook display name
2. URL to your Facebook profile
Make sure you have set your profile to Public View in order for your task to be verified.

3. Number of Facebook friends on your account

(b) A Facebook like task in Microworkers

Fig. 3.2: Examples of Facebook like tasks in Fiverr and Microworkers.

page's URL was sent to each Fiverr seller, requesting passing some fake likes to the page. All the sellers completed the tasks and delivered fake likes within the promised time period. Our crawler collected a snapshot of each of our honeypot pages (including the number of current likes) once every one hour for 10 days. Overall, 10 sellers sent 3,916 fake likes to the 10 honeypot pages. Another crawler collected each fake liker's profile information. Out of 3,916 fake likers, 3,207 users had no privacy setting, so finally we collected 3,207 fake liker profiles including a total of 1.86 million pages they liked, 0.21 million posts and 0.82 million friends information from Facebook.

In contrast to Fiverr, Microworkers is a requester-driven crowdsourcing platform where a requester creates a task as shown in Figure 3.2(b), and workers perform the task. A task contains a title and task description which consisting of task instruction with a target URL. First we selected 353 tasks, titles of which contained "Facebook Like" and extracted targeted Facebook page URLs. Then, we linked the URLs to targeted pages, and extracted a list of likers of the targeted pages. We counted how many pages among 353 targeted pages each unique liker liked, and randomly selected 4,482 fake likers who liked at least two targeted pages. Out of 4,482 fake likers, 3,688 users did not have any privacy setting, so we collected their profiles including a total of 1.63 million pages they liked, 0.33 million posts, and 1.54

million friends information. We marked likers from Fiverr and Microworkers as fake likers.

### 3.2.2 Legitimate Liker Profiles

To analyze characteristics of fake likers and legitimate likers in the following chapters, we also collected legitimate liker profiles by using two data collection approaches: (i) random user collection; and (ii) user collection from conference and research communities.

In the random user selection approach, first we randomly selected 20 seed users who lived in various countries. Then from each seed user, we collected his/her friends network and friends' friends network by reaching to two hops based on breadth first search (BFS). We had 28,200 user list, and randomly selected 5,700 users in order to further randomize collected user profiles. Out of 5,700 random users, 3,779 users did not have any privacy setting, so we collected 3,779 user profiles. Then, two labelers conducted manual labeling process for the 3,779 user profiles, since almost 2% accounts on Facebook are undesirable accounts according to a Facebook SEC filing report [20]. The labelers carefully investigated each user's posts and timeline information. They achieved 99.7% agreement and 0.78 kappa coefficient [21]. We removed a user, whose profile was labeled as a suspicious user by at least one labeler. Finally, 3,701 legitimate user profiles remained. These legitimate users liked 0.6 million pages and posted 0.28 million posts, and had 2.24 million friends.

Next, we collected 2,552 user profiles from 13 conference groups on Facebook[1]. As other researchers did [11], we also assumed that these technically savvy users were less likely to be infected by malware or other attacks, and did not perform fake liking activity since they were associated with the premier institutions or well-known IT companies. In addition, we sampled 1,000 out of the 2,552 user profiles, and checked whether they were suspicious accounts. There was no suspicious account in the sample set. Based on the prior research and our verification process, we considered 2,552 conference users as legitimate users. These legitimate users (i.e., users in conference groups) liked 0.57 million pages, posted 0.17 million posts, and had 0.87 million friends. We marked the random users

[1] SIGCOMM, COSN, CIKM, SIGMOD, RecSys, SIGMOBILE, MobiSys2014, HCI Korea, ACM SIGCHI, ISWC, ICDE, CIKM 2011, VLDB 2013.

and the users collected from conference groups as legitimate likers after the above manual labeling and careful selection.

Table 3.1: Dataset.

| Types | \|Likers\| | \|pages\| | \|posts\| | \|friends\| |
|---|---|---|---|---|
| F. likers – Fiverr | 3,207 | 1.86M | 0.21M | 0.82M |
| F. likers – Microwor. | 3,688 | 1.63M | 0.33M | 1.54M |
| L. likers – random | 3,701 | 0.60M | 0.28M | 2.24M |
| L. likers – conference | 2,552 | 0.57M | 0.17M | 0.87M |
| Total | 13,148 | 4.66M | 0.99M | 5.47M |

Overall, we collected 6,895 faker liker profiles and 6,253 legitimate liker profiles with a list of 4.66 million pages they liked, 0.99 million posts and 5.47 million friends information as shown in Table 3.1, and used the dataset in the following chapters.

### 3.2.3 Temporal Data

To implement existing methods (i.e., baseline methods) described in Chapter 7 and compare them with our fake liker detection approach, we additionally collected temporal data of 1,400 (700 fake likers and 700 legitimate likers) out of 13,148 likers. Temporal data contain series of daily profile snapshots and categories of daily liked pages for 30 days between Dec 1, 2015 and Dec 30, 2015.

# CHAPTER 4
# ANALYSIS OF SELLERS, WORKERS, AND FAKE LIKERS

In this chapter, we analyze characteristics of Fiverr sellers, Microworkers workers, and their corresponding fake liker profiles.

## 4.1 Fiverr Sellers and Associated Fake Likers

First, we analyzed how quickly each Fiverr seller delivered fake likes. Did they deliver them on-time? Figure 4.1 shows the number of fake likes delivered by each seller over time. 0 in x-axis is the time when we requested fake liking. Almost all of fake likes were delivered within three days, and there were no significant changes in the like patterns after the three days. Interestingly, all the sellers delivered the fake likes with a bursty pattern, which clearly shows that the sellers used bots or softwares to manage their Facebook accounts to provide the fake likes to buyers.

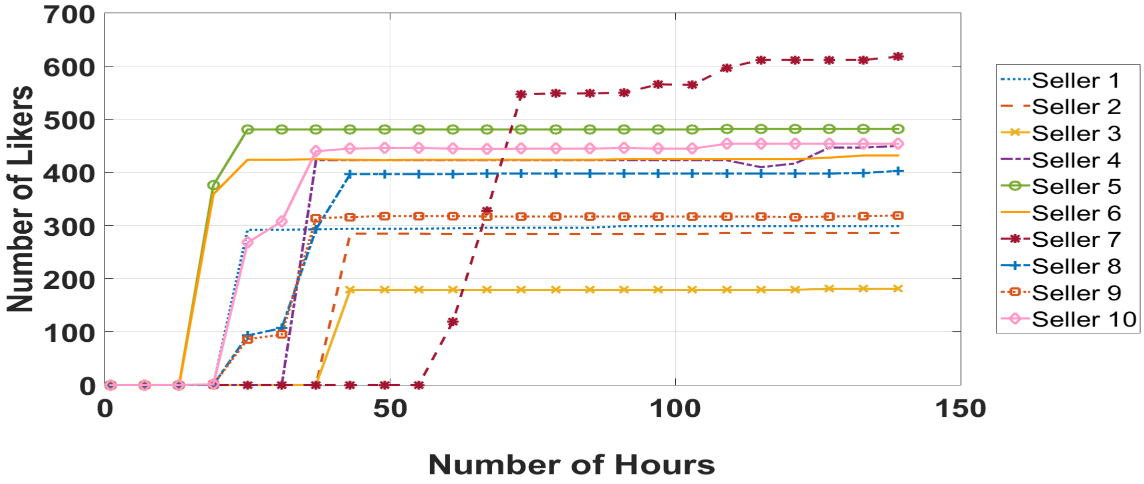


Fig. 4.1: Number of fake likes obtained from 10 Fiverr sellers over time.

Next we analyzed demographic information of fake likers associated with Fiverr sellers. In particular, we were interested in location information and other demographic information

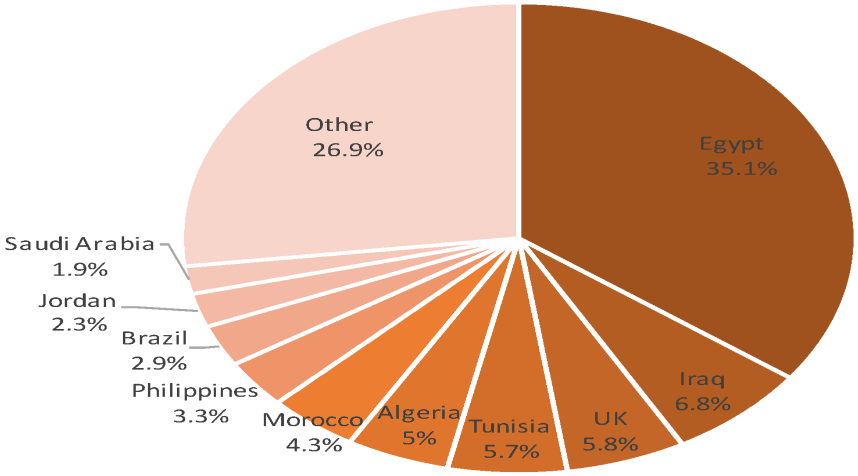
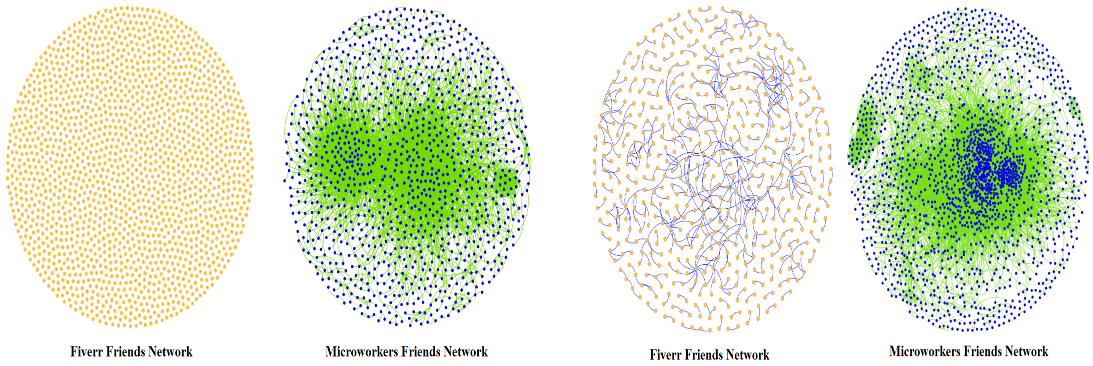


Fig. 4.2: Top 10 countries of fake likers from Fiverr.

(e.g., gender, age). Facebook provides the demographic information of likers to Facebook page owners. Since we created our own honeypot pages on Facebook, and requested the sellers to send fake likes to our pages, we were able to access the demographic information of the fake likers. Figure 4.2 shows top 10 countries of the fake likers, and about 73% of all the fake likers were from the top 10 countries, which are all developing countries except UK. The age and gender distributions of all the fake likers associated with Fiverr sellers are shown in Table 4.1. Most fake likers were in a range of 18-34 years old regardless of which seller they were associated with. Surprisingly, some fake likers were teenagers. There were more male likers than female likers.

Table 4.1: Age and gender distributions of fake likers from Fiverr.

| | | **Age Distribution (%)** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Sellers** | **Gender % F/M** | **13-17** | **18-24** | **25-34** | **35-44** | **45-54** | **55-64** | **65+** |
| **Seller 1** | 44/56 | 21.5 | 45.4 | 18.52 | 9.13 | 3.458 | 0.98 | 0.98 |
| **Seller 2** | 43/57 | 23.18 | 44.4 | 18.5 | 7.41 | 3.45 | 1.48 | 1.23 |
| **Seller 3** | 37/63 | 22.81 | 45.7 | 18.22 | 5.81 | 4.98 | 1.63 | 0.82 |
| **Seller 4** | 23/77 | 10.99 | 52.8 | 25.5 | 5.61 | 2.34 | 0.46 | 2.33 |
| **Seller 5** | 68/32 | 10.29 | 51.5 | 29.6 | 6.03 | 1.77 | 0.53 | 0.17 |
| **Seller 6** | 42/58 | 18.45 | 38.2 | 31.0 | 7.01 | 2.24 | 0.91 | 0.18 |
| **Seller 7** | 75/25 | 8.03 | 47.42 | 30.9 | 9.59 | 3.24 | 0.28 | 0.56 |
| **Seller 8** | 29/71 | 17.87 | 35.5 | 30.14 | 12.9 | 2.77 | 0.79 | 0.0 |
| **Seller 9** | 30/70 | 16.63 | 40.8 | 26.24 | 10.08 | 3.78 | 2.01 | 0.50 |
| **Seller 10** | 23/76 | 17.75 | 38.16 | 28.98 | 10.49 | 3.43 | 0.72 | 0.36 |



(a) Direct friendship relations in each community

(b) 2-hop (mutual) friendship relations in each community

Fig. 4.3: Social graphs of fake likers from Fiverr and Microworkers.

Another interesting question is "did the fake likers unliked our honeypot pages a certain period (e.g., a week or a month) later?". Researchers observed this behavior on Twitter in which spammers often followed and unfollowed users [12]. To answer this question, we checked how many fake likers unliked our pages in three months after they had liked our pages. Only 10% fake likers unliked our pages (i.e., dropped the likes).

## 4.2 Comparison of Fake Likers from Fiverr and Microworkers

We are interested to see how fake likers from Fiverr or/and Microworkers were connected via Facebook friendship network. Did fake likers from Fiverr (or Fiverr community) have different friendship network compared with fake likers from Microworkers (or Microworkers community)? To answer this question, we created two social graphs of each of the Fiverr and Microworkers communities as shown in Figure 4.3. In Figure 4.3(a), a node indicates a fake liker and if a pair of fake likers within a community has a direct relationship (i.e., friends in Facebook), we added an edge between them. The direct friendship connections among fake likers in the Fiverr community did not exist, whereas fake likers in the Microworkers community were densely connected. In particular, 1,099 out of 3,688 fake likers/nodes in the Microworkers community were connected via 5,239 friendship connections/edges with 9.5 degree and 0.446 clustering coefficient on average.

In Figure 4.3(b), a node indicates a fake liker and if a pair of fake likers within in a community has at least one common friend, we added an edge between them. Fake likers in the Fiverr community were sparsely connected whereas fake likers in the Microworkers community were densely connected. In particular, 570 fake likers in the Fiverr community were connected via 578 friendship connections with 2.0 degree and 0.233 clustering coefficient on average while 1,802 fake likers in the Microworkers community were connected via 48,173 friendship connections with 53.4 degree and 0.612 clustering coefficient on average. Then, we analyzed top 15 common friends of each of the Fiverr and Microworkers communities. Top 15 common friends of fake likers in the Fiverr community had 4,500 friends on average, but we did not find any evidence of crowdturfing activities on Facebook. But, by analyzing the posts of the Microworkers mutual friends, we observed that these users also shared crowdsourcing related posts on Facebook and had 650 friends on average. 20% of their friends were fake likers in the Microworkers community. This finding reveals that there may be many other fake likers in Facebook who may be friends of fake likers in our dataset.

We also performed analysis to find out whether fake likers in the Fiverr community were directly befriended by the fake likers in the Microworkers community or vice versa. It turned out there was no friend relationship between fake likers of Fiverr and Microworkers. Then, we checked whether fake likers in the Fiverr community and fake likers in the Microworkers community had any common friends (i.e., the same users were commonly friends of both fake likers of Fiverr and Microworkers). We found that there were 672 mutual friends.

Next, we wondered whether there were cross-workers between Fiverr and Microworkers who had performed fake liking tasks in both sites. Therefore, we computed a pairwise similarity among the pages liked by Fiverr sellers and Microworkers workers. There was only a negligible amount of similarity between the pages liked by Fiverr sellers and Microworkers workers. This shows that there were no cross-workers. The small similarity was due to the verified pages liked by fake likers from both sites. By liking some verified pages, these fake likers could avoid being detected by the Facebook security team. Note that the verified

Table 4.2: Top 5 URLs shared or posted by fake likers, and number of likers who shared or posted the URLs.

| **Community** | **Top 5 URLs** | **# of Likers** |
|---|---|---|
| **Fiverr** | https://www.facebook.com/facebook/ | 215 |
| | http://apps.facebook.com/monsterlegends | 89 |
| | http://apps.facebook.com/dragoncity | 38 |
| | http://apps.facebook.com/stick_run | 37 |
| | http://apps.facebook.com/topeleven | 31 |
| **Microworkers** | https://www.krowdster.co | 537 |
| | http://www.indiegogo.com/projects/1240265 | 412 |
| | https://www.indiegogo.com/projects/1528609 | 312 |
| | http://www.fitneszszakuzlet.hu | 276 |
| | https://www.mykomms.com/ | 197 |

pages are the pages verified by Facebook (e.g., Amazon, Walmart and Disney). However, a pairwise similarity of pages liked by fake likers from each site/community was much larger.

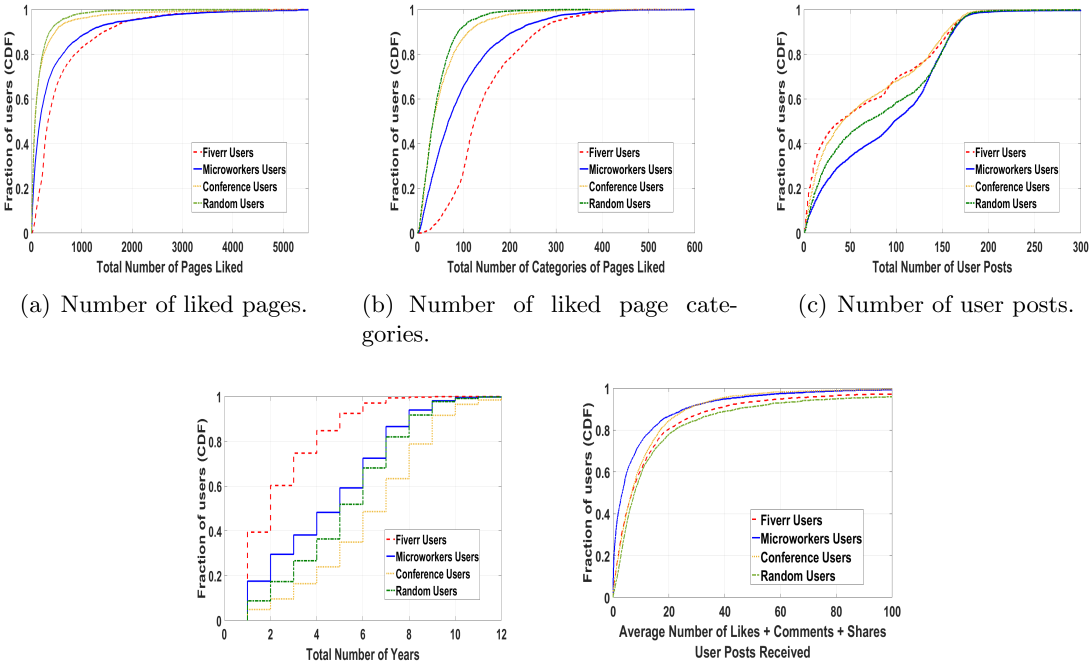
Next, we analyzed what URLs fake likers frequently shared or posted on Facebook. We extracted URLs from their posts and counted the number of posts containing each distinct URL in the Fiverr community and the Microworkers community. Table 4.2 presents top 5 URLs in each community. Fake likers in the Fiverr community posted URLs related to Facebook apps, games and etc. However, fake likers in the Microworkers community posted URLs related to advertisements of crowdfunding projects and other websites. Based on this URL sharing analysis, Microworkers workers used their Facebook accounts for not only fake liking activities, but also other crowdturfing tasks to earn more money.

Based on the analysis, we conclude that the Fiverr community was different from the Microworkers community in terms of direct relationship, mutual friends, job tasks they performed, and URL sharing patterns.

# CHAPTER 5

# CHARACTERISTICS OF FAKE LIKERS AND LEGITIMATE LIKERS

In this chapter, we try to understand characteristics of fake likers and legitimate likers. Can we find distinguishing patterns between them? To answer this question, we analyzed various properties of fake likers and legitimate likers, and show five representative cumulative distribution functions (CDFs) in Figure 5.1. Figure 5.1(a) clearly shows that fake likers from Fiverr and Microworkers performed more page liking activities than legitimate likers including random and conference users. In particular, 90% of random users and conference users liked at most 369 and 481 pages, respectively, whereas 90% of Fiverr's fake likers and Microworkers' fake likers liked at most 1,481 and 1,137 pages, respectively.



(a) Number of liked pages.

(b) Number of liked page categories.

(c) Number of user posts.

(d) Longevity (years).

(e) Average number of social attention per post.

Fig. 5.1: CDFs of fake (i.e., Fiverr and Microworkers users) and legitimate (i.e., random and conference users) likers in five properties.

When a user creates a Facebook page, Facebook requires the user to choose a category associated with the page. In general, people are interested in some certain topics/categories like electronics and furniture. Since fake likers like pages for earning money, can we observe that these fake likers like pages with various categories? In contrast, do legitimate likers like pages under certain categories? Figure 5.1(b) presents the number of categories associated with Facebook pages liked by fake likers and legitimate likers. Fake likers liked pages under more number of categories than legitimate likers. In particular, Fiverr's fake likers liked pages under the largest number of categories.

Next, we analyzed the number of posts created by fake likers and legitimate likers as shown in Figure 5.1(c). Interestingly, Microworkers' fake likers posted the largest number of posts whereas Fiverr's fake likers posted the smallest number of posts. This result indicates that fake likers from these two sites had different posting behaviors even though they both participated in fake liking activities. Conference users, a part of legitimate likers, created the number of posts closer to Fiverr's fake likers.

Another interesting question is when accounts of fake likers and legitimate likers were created. Can we observe a different longevity pattern between fake likers and legitimate likers? Figure 5.1(d) clearly shows that legitimate likers' accounts were created earlier than fake likers' accounts. In particular, conference users like tech savvy users created their accounts earlier than the other users. Fiverr's accounts were created most recently and 80% of them were created within four years.

Lastly, we measured how much social attention fake likers and legitimate likers received from other users. In particular, social attention was measured by the sum of average number of likes, comments and shares per post. Figure 5.1(e) shows that random users (legitimate likers) received the largest number of social attention followed by fake likers from Fiverr, conference users (legitimate likers) and fake likers from Microworkers. 80% of the random users and conference users received up to 23 and 15 likes+comments+shares per post, respectively, whereas 80% of Fiverr and Microworkers' fake likers received up to 19 and 13 likes+comments+shares per post, respectively.

Our analytical results confirm that fake likers were different from legitimate likers in terms of various characteristics like liking behaviors, longevity, posting activities and social interactions with friends. Based on the results, next we describe our proposed features which were used for building fake liker classifiers.

# CHAPTER 6
# FEATURE EXTRACTION AND SELECTION

In this chapter, we listed our proposed features, and grouped them by five categories: (i) profile features; (ii) posting activity features; (iii) page liking features; (iv) social attention features; and (v) temporal activity based features. Then we conducted feature selection. Note that we intentionally did not propose and use language-dependent features so that our models can be applied to any fake liker regardless what languages they used in online social networks.

## 6.1 Proposed Features

**(i) Profile Features :** Profile features consist of three features as follows:

- Number of lines in *About* section
- Longevity of an account (i.e., how many years it has been active)
- Number of friends

Our intuition is that fake likers would not spend much time to make rich profiles, and have short longevity and smaller number of friends.

**(ii) Posting Activity Features :** These features were extracted from each user's posts. There are two types of user posts: his own posts and shared posts. His own posts mean posts that were created by the user $A$. Shared posts mean posts that were created by another user $B$ and were shared by the user $A$. Posts include photos, pages, videos, text messages, etc. We extracted following features:

- Average number of posts per day
- Total number of posts created by the user himself
- Proportion of shared photos out of the total number of posts
- Proportion of shared posts out of the total number of posts

- Proportion of shared pages out of the total number of posts
- Maximum number of posts in a day: Given a time series of number of daily posted posts, we picked the largest number in a day
- Average number of links/URLs per post
- Skewness of daily posted posts: Given a time series of number of daily posted posts, we measured the skewness ($= \frac{3(mean - median)}{standard\ deviation}$)

**(iii) Page Liking Features :** We extracted two page liking features:

- Category entropy: In the previous chapter, we observed that fake likers liked pages under various categories. A larger category entropy indicates that a user randomly liked pages under various categories. Given a list of Facebook categories $C = c_1, c_2, c_3, \ldots, c_k$ and corresponding number of pages in each category liked by a user $u$, the user's category entropy is calculated as follows:

  $$CatEntropy(u) = -\sum_{i=1}^{k} \frac{p_i}{N} \log \frac{p_i}{N}$$

  , where $N$ is the total number of pages liked by the user $u$, and $p_i$ is the number of liked pages under a category $i$.
- Proportion of verified pages out of the total number of pages liked by a user. Some pages are verified by Facebook, and it means these pages are authentic.

**(iv) Social Attention Features :** We extracted following three features:

- Average number of likes (selected by other users) per post
- Average number of comments per post
- Average number of shared per post

These features measure how much attention a user's posts get from other users. Intuitively, fake likers posts would get lower attention from other users since they would not socially interact with them.

**(v) Temporal Activity based Features :** As we mentioned in Section 3.2.3, we collected temporal data of 1,400 likers (700 fake likers and 700 legitimate likers) to compare existing methods with ours in the following section. From the temporal data, we extracted following two features:

- A change rate of the number of liked pages during 30 days: We extracted 30 values, each of which was the number of liked pages by the user as of the day. For example, a user liked 500 pages until day 0, and liked 5 pages in day 1 and liked 10 pages in day 2. Then day 1 and day 2's feature values were 505 and 515. Then, we measured a standard deviation of the 30 values.
- A change rate of category entropies during 30 days: In each day, we measured a category entropy. Then we measured a standard deviation of 30 category entropies.

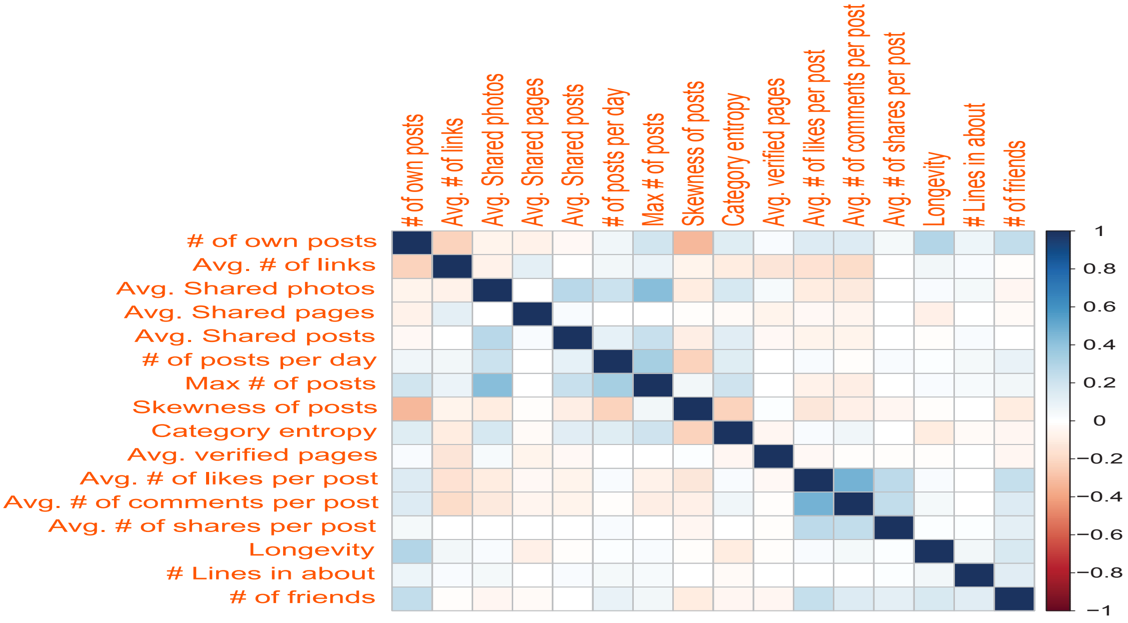


Fig. 6.1: Pearson correlation scores of all the pairs of 16 features.

## 6.2 Feature Selection

To make sure we did not use too similar or correlated features together, we measured Pearson correlation and ran Chi-square test for the first four feature sets except the temporal activity based features. Note that we did the same process for all the feature sets including temporal activity based features in a small dataset containing temporal data, but only show Pearson correlation results for the four feature sets in the entire dataset in this work. Figure 6.1 shows Pearson correlation results of all the pairs of features. The largest correlation score was still less than 0.5, so we kept all the features.

Table 6.1: Top 10 features and average feature values of fake likers and legitimate likers.

| **Features** | **F. Likers** | **L. Likers** |
|---|---:|---:|
| Category entropy | 6.35 | 4.47 |
| Longevity | 3.62 | 5.53 |
| Average # of posts per day | 0.31 | 0.12 |
| # of lines in *About* section | 4.02 | 4.08 |
| Proportion of verified pages | 0.23 | 0.26 |
| # of friends | 343 | 493 |
| Average # of comments per post | 1.69 | 1.95 |
| Skewness of daily posted posts | 9.10 | 11.81 |
| Average # of links per post | 0.18 | 0.16 |
| Maximum # of posts in a day | 8.58 | 5.56 |

Next, we ran the Chi-square test [22] to rank features in order of the largest distinguishing power. The larger Chi-square value a feature has, the larger distinguishing power it has. Table 6.1 shows ranked results, and average feature values of fake likers and legitimate likers. Category entropy, longevity, average number of posts per day, the number of lines in *About* section, and proportion of verified pages were the most significant features.

# CHAPTER 7
# EXPERIMENTS

In this chapter, we conduct three experiments. In the first experiment, we build fake liker classifiers in a small dataset containing temporal data and compare the classifiers with three baseline methods (i.e., PCA [11], SynchroTrap [10] and CopyCatch [9]). In the second experiment, we build fake liker classifiers in the entire dataset without temporal data, and compare the classifiers with the three baseline methods. In the third experiment, we measure how robust our classification model is under two attack scenarios – (i) individual attack model and (ii) coordinated attack model.

## 7.1 Detecting Fake Likers in a Small Dataset

Out of 13,148 Facebook likers in our dataset presented in Table 3.1, we additionally collected daily snapshots of 1,400 likers (700 fake likers and 700 legitimate likers) for 30 days because PCA based approach, one of the baseline methods, used temporal data. Note that collecting daily snapshot containing which user liked what pages in each day is expensive and less scalable for billions of users, but in order to compare our approach with the three baseline methods, we did for a small sample of likers.

Then, we split the small dataset to training and test sets. The training set consisted of 1,000 likers (500 fake likers and 500 legitimate likers), and the test set consisted of 400 likers (200 fake likers and 200 legitimate likers). PCA based approach, CopyCatch and SynchroTrap used the training set as a validation set to find a threshold giving them the optimal results and then applied the threshold to the test set. In addition, we measured the upper bound of the baseline methods' performance by finding the threshold and corresponding optimal results within the test set without using the training set.

In our classification approach, we built fake liker classifiers based on 30+ machine learning algorithms (e.g., LogitBoost [23], Random Forest [24], XGBoost [25], SVM) using our proposed features described in the previous chapter in order to check which classifier produced the best result.

For PCA based approach [11], we extracted 30 temporal, 1,079 spatial/categorical and 30 spatio-temporal/category-temporal feature values from each liker's profile in both training and test sets as the authors did. To reduce the dimensionality of their features, we implemented the principal component analysis and determined the principal components. We observed 95% variance in the top 400 principal components in both sets. We then used the sets and projected them into the normal and residual subspaces in order to find fake liker's behaviors. We computed the L2 norm [26] and set the squared prediction error (SPE) as a threshold to find fake likers. If a user exceeds the SPE value, then the user is likely to be a fake liker. For determining the SPE, we changed a threshold value from 1% to 99% by increasing 1% each time in the training set. We found the optimal threshold, and applied the threshold value to the test set.

SynchroTrap [10] measures page liking similarity of each pair of users, and runs single-linkage hierarchical clustering. The output of the algorithm is a dendrogram structure, and the algorithm requires a cutoff threshold (i.e., similarity threshold) and a minimum size of a cluster to determine clusters of malicious accounts. Again, we found the best cutoff and size values from the training set and applied to the test set.

CopyCatch [9] discovers groups of fake likers by measuring page liking similarity of the groups in a specific time range, and outputs near bipartite cores/graphs which can be considered as malicious accounts. This method needs three input parameters: (i) the minimum number of users in a near bipartite core; (ii) the minimum number of pages in a near bipartite core; and (iii) how densely users are connected to pages (e.g., each user in a near bipartite core should like 90% pages). We varied these parameter values and found the best values from the training set and applied them to the test set.

In experiments, we measured accuracy, false positive rate (FPR) and false negative rate (FNR). FPR means the number of legitimate likers, who were misclassified as fake likers, over the total number of legitimate likers. FNR means the number of fake likers, who were

Table 7.1: Experimental results in a small dataset containing temporal data.

| Approach | Accuracy | FPR | FNR |
|---|---|---|---|
| PCA | 69.0% | 0.30 | 0.32 |
| SynchroTrap | 50.5% | 0 | 0.99 |
| CopyCatch | 56.5% | 0.85 | 0.02 |
| PCA - Test | 69.0% | 0.30 | 0.32 |
| SynchroTrap - Test | 63.5% | 0 | 0.73 |
| CopyCatch - Test | 65.5% | 0.46 | 0.23 |
| our LogitBoost | 87.5% | 0.11 | 0.13 |
| our Random Forest | 88.5% | 0.09 | 0.13 |
| our XGBoost | **89.7%** | 0.08 | 0.11 |

misclassified as legitimate likers, over the total number of fake likers.

Table 7.1 presents experimental results of the three baseline methods and our three most effective classifiers. Overall, our classification models outperformed the three baseline methods. In particular, XGBoost classifier achieved 89.7% accuracy, 0.08 FPR and 0.11 FNR, improving up to 39.2% (= 89.7 - 50.5) accuracy compared with the baseline methods. PCA, SynchroTrap and CopyCatch achieved 69%, 50.5% and 56.5% accuracy, respectively. Their upper bound results (i.e., PCA - Test, SynchroTrap - Test and CopyCatch - Test which found optimal threshold/parameter values within the test set) achieved 69%, 63.5% and 65.5% accuracy, respectively. These two different results show that a weakness of the baseline methods is hard to find optimal threshold or input parameter values. Second, even though we found the upper bound, they were still less effective than our classification models. One interesting observation is that SynchroTrap achieved 0 FPR. It means there was no misclassification of legitimate likers to fake likers. But, SynchroTrap only correctly identify 1% of fake likers (i.e., 0.01 recall).

## 7.2 Detecting Fake Likers in the Wild

Now we turn to detect fake likers in the entire dataset containing 13,148 users' profiles without using expensive temporal data/features. In this experiment, we conducted 10-fold cross-validation, creating 10 pairs of training and test sets. PCA method was applied to our 16 features excluding temporal activity-based features instead of the original features (i.e., 30 temporal, 1,079 spatial/categorical and 30 spatio-temporal/category-temporal features), which required temporal data. Like what we did in the previous experiment, PCA, CopyCatch and SynchroTrap found an optimal threshold in each training set and applied it to each test set. In addition, we measured the upper bound of the three baseline methods. In our classification approach, we chose XGBoost, the best method in the previous experiment, and used 16 features without temporal activity-based features.

Table 7.2: Experiment results in the entire dataset.

| Approach | Accuracy | FPR | FNR |
|---|---|---|---|
| PCA | 56.3% | 0.49 | 0.38 |
| SynchroTrap | 52.1% | 0 | 0.91 |
| CopyCatch | 60.1% | 0.80 | 0.04 |
| PCA - Test | 57.6% | 0.28 | 0.54 |
| SynchroTrap - Test | 62.0% | 0.01 | 0.71 |
| CopyCatch - Test | 66.9% | 0.09 | 0.55 |
| our XGBoost | **87.1%** | 0.10 | 0.14 |

Table 7.2 show experimental results in the entire dataset. Again, XGBoost based classifier outperformed PCA, SynchroTrap and CopyCatch, achieving 87.1% accuracy, 0.1 FPR and 0.14 FNR. The same weaknesses of the baseline methods occurred – (i) finding optimal threshold/input parameter values are hard; and (ii) the upper bound of the baseline methods is still lower than our classification model.

We further analyzed why XGBoost based classifier got 0.1 FPR and 0.14 FNR, focusing on misclassification cases. Out of 1,013 false negatives (again, who were misclassified as legitimate likers), 56 fake likers were from Fiverr and the remaining 957 were from Microworkers. It means that detecting fake likers from Microworkers were relatively harder than fake likers from Fiverr because some fake likers from Microworkers used their personal Facebook profiles for fake liking activities (i.e., their profiles look like legitimate profiles except fake liking activities). In false positive cases, some legitimate likers had similar behaviors with fake likers by posting many posts in a single day and having a high proportion of shared pages out of the total number of posts.

### 7.3 Robustness of Our Approach

In the previous two experiments, our classification approach significantly outperformed the baseline methods. What if fake likers learned what features legitimate likers have, and change their behaviors to avoid the detection approach? To measure robustness of our approach, we simulated two attack models: (i) individual attack model; and (ii) coordinated (group) attack model.

In the individual attack model, we assumed that each fake liker *independently* selects one of our features, and then change its value to a legitimate liker's feature value. Specifically, given a range of the feature values of legitimate likers, the simulator randomly choose a value which is used for a fake liker. We conducted 10-fold cross-validation and ran the simulation 10 times. Our XGBoost based classifier achieved 85.5% accuracy, 0.157 FPR and 0.133 FNR, decreasing 1.6% accuracy compared with our approach without the attack model (87.1% accuracy). It means our approach is robust under the individual attack model.

In the coordinated attack model, we assumed that all the fake likers choose the *same* feature or features, and change its value or their values to a legitimate liker's feature value or values. Compared with the individual attack model, these fake likers already know which feature/features they are going to manipulate. The simulator randomly assigns a legitimate feature value for each fake liker. We tested changing one feature to multiple features. Again, we conducted 10-fold cross-validation and ran 10 times.

Table 7.3 show experimental results under the coordinate attack model. Note that we only show results affected by the changes. The coordinated attack for single feature decreased accuracy to between 83.4% and 87%. When the fake likers targeted one of four feature sets, each of which consists of multiple features, classification accuracy decreased to between 82.4% and 86.1%. Compared with our approach without the coordinated attack model (87.1% accuracy), the coordinated attack model slightly affected the performance of our approach.

Based on the experimental results, we conclude that the coordinated attack model is

Table 7.3: Our classification results under the coordinated attack model.

| Features | Accuracy | FPR | FNR |
|---|---|---|---|
| Skewness of daily posted posts | 87.0% | 0.11 | 0.14 |
| Average # of links per post | 87.0% | 0.11 | 0.14 |
| # of friends | 86.9% | 0.11 | 0.14 |
| Proportion of verified pages | 86.8% | 0.11 | 0.14 |
| Average # of likes per post | 86.8% | 0.11 | 0.14 |
| Proportion of shared pages | 86.7% | 0.11 | 0.14 |
| # of lines in *About* | 86.0% | 0.11 | 0.15 |
| Longevity | 85.9% | 0.12 | 0.15 |
| Category entropy | 83.4% | 0.16 | 0.16 |
| Social attention features (3 feat.) | 86.1% | 0.12 | 0.15 |
| Profile features (3 feat.) | 84.2% | 0.13 | 0.17 |
| Posting activity features (8 feat.) | 84.1% | 0.13 | 0.17 |
| Page liking features (2 feat.) | 82.4% | 0.17 | 0.17 |

more effective than the individual attack model, but our classification model is still robust for identifying fake likers.

# CHAPTER 8
# CONCLUSION

In this work, we conducted a comprehensive analysis of fake likers on Facebook collected from Fiverr and Microworkers by using the linkage and honeypot approaches. We compared how fake likers from two different sources were different. We found that fake likers from Microworkers were more densely connected than Fiverr in terms of direct friendship and mutual friends. In the comparison between fake likers and legitimate likers, we found that fake likers were different from legitimate likers in terms of liking behaviors, longevity, posting activities and social interaction with friends. Based on the analysis, we proposed 5 types of feature sets toward building our classification models. In experiments, we thoroughly evaluated them against three baseline methods (i.e., PCA, SynchroTrap and CopyCatch). Experimental results show that our models significantly outperformed the baseline methods, achieving 87.1% accuracy, 0.1 FPR and 0.14 FNR. To measure robustness of our models, we simulated the individual attack and coordinated attack models. Under both attack models, our approach consistently achieved over 82% accuracy.

# REFERENCES

[1] Conner, C., "Amazon Sues 1,114 Fake Reviewers On Fiverr," www.forbes.com/sites/cherylsnappconner/2015/10/18/amazon-sues-1114-fake-reviewers-on-fiverr-com/, October 2015.

[2] De Cristofaro, E., Friedman, A., Jourjon, G., Kaafar, M. A., and Shafiq, M. Z., "Paying for likes?: Understanding Facebook like fraud using honeypots," *IMC*, 2014.

[3] Liu, Y., Liu, Y., Zhang, M., and Shaoping, M., "Pay Me and I'll Follow You: Detection of Crowdturfing Following Activities in Microblog Environment," *IJCAI*, 2016.

[4] Motoyama, M., McCoy, D., Levchenko, K., Savage, S., and Voelker, G. M., "Dirty Jobs: The Role of Freelance Labor in Web Service Abuse," *USENIX Conference on Security*, 2011.

[5] Pham, N., "Vietnam admits deploying bloggers to support government," http://www.bbc.co.uk/news/world-asia-20982985, January 2013.

[6] Stringhini, G., Egele, M., Kruegel, C., and Vigna, G., "Poultry Markets: On the Underground Economy of Twitter Followers," *Workshop on Online Social Networks*, 2012.

[7] Thomas, K., McCoy, D., Grier, C., Kolcz, A., and Paxson, V., "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse," *USENIX Conference on Security*, 2013.

[8] Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., and Zhao, B. Y., "Serf and turf: crowdturfing for fun and profit," *WWW*, 2012.

[9] Beutel, A., Xu, W., Guruswami, V., Palow, C., and Faloutsos, C., "CopyCatch: stopping group attacks by spotting lockstep behavior in social networks," *WWW*, 2013.

[10] Cao, Q., Yang, X., Yu, J., and Palow, C., "Uncovering large groups of active malicious accounts in online social networks," *CCS*, 2014.

[11] Viswanath, B., Bashir, M. A., Crovella, M., Guha, S., Gummadi, K. P., Krishnamurthy, B., and Mislove, A., "Towards detecting anomalous user behavior in online social networks," *USENIX Conference on Security*, 2014.

[12] Lee, K., Eoff, B. D., and Caverlee, J., "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," *ICWSM*, 2011.

[13] Stringhini, G., Kruegel, C., and Vigna, G., "Detecting spammers on social networks," *ACSAC*, 2010.

[14] Lee, K., Webb, S., and Ge, H., "The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing," *ICWSM*, 2014.

[15] Boshmaf, Y., Logothetis, D., Siganos, G., Lería, J., Lorenzo, J., Ripeanu, M., and Beznosov, K., "Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs." *NDSS*, 2015.

[16] Song, J., Lee, S., and Kim, J., "CrowdTarget: Target-based Detection of Crowdturfing in Online Social Networks," *CCS*, 2015.

[17] Facebook, "Improvements To Our Site Integrity Systems," https://www.facebook.com/10151005934870766, 2012.

[18] Facebook, "Keeping Facebook activity authentic," https://www.facebook.com/10152309368645766/, 2014.

[19] Lee, K., Tamilarasan, P., and Caverlee, J., "Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media." *ICWSM*, 2013.

[20] Facebook, "Facebook - Annual Report - Investor Relations," http://investor.fb.com/secfiling.cfm?filingID=1326801-14-7, 2014.

[21] Viera, A. J., Garrett, J. M., et al., "Understanding interobserver agreement: the kappa statistic," *Fam Med*, Vol. 37, No. 5, 2005, pp. 360–363.

[22] Yang, Y. and Pedersen, J. O., "A Comparative Study on Feature Selection in Text Categorization," *ICML*, 1997.

[23] Friedman, J., Hastie, T., Tibshirani, R., et al., "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, Vol. 28, No. 2, 2000, pp. 337–407.

[24] Breiman, L., "Random forests," *Machine learning*, Vol. 45, No. 1, 2001, pp. 5–32.

[25] Chen, T. and He, T., "xgboost: eXtreme Gradient Boosting," 2015.

[26] Wolfram Research, I., "L2 Norm," http://mathworld.wolfram.com/L2-Norm.html, 2016.