Smoke Screener or Straight Shooter: Detecting Elite Sybil Attacks in User-Review Social Networks

Haizhong Zheng
Shanghai Jiao Tong University
sjtu.zhenghaizhong@gmail.com

Minhui Xue
NYU Shanghai and ECNU
minhuixue@gmail.com

Hao Lu
Shanghai Jiao Tong University
luhao0522@gmail.com

Shuang Hao
University of Texas at Dallas
shao@utdallas.edu

Haojin Zhu
Shanghai Jiao Tong University
zhu-hj@cs.sjtu.edu.cn

Xiaohui Liang
University of Massachusetts Boston
xiaohui.liang@umb.edu

Keith Ross
NYU and NYU Shanghai
keithwross@nyu.edu

***Abstract*—Popular User-Review Social Networks (URSNs)—such as Dianping, Yelp, and Amazon—are often the targets of reputation attacks in which fake reviews are posted in order to boost or diminish the ratings of listed products and services. These attacks often emanate from a collection of accounts, called Sybils, which are collectively managed by a group of real users. A new advanced scheme, which we term elite Sybil attacks, recruits organically highly-rated accounts to generate seemingly-trustworthy and realistic-looking reviews. These elite Sybil accounts taken together form a large-scale sparsely-knit Sybil network for which existing Sybil fake-review defense systems are unlikely to succeed.**

In this paper, we conduct the first study to define, characterize, and detect elite Sybil attacks. We show that contemporary elite Sybil attacks have a hybrid architecture, with the first tier recruiting elite Sybil workers and distributing tasks by Sybil organizers, and with the second tier posting fake reviews for profit by elite Sybil workers. We design ELSIEDET, a three-stage Sybil detection scheme, which first separates out suspicious groups of users, then identifies the campaign windows, and finally identifies elite Sybil users participating in the campaigns. We perform a large-scale empirical study on ten million reviews from Dianping, by far the most popular URSN service in China. Our results show that reviews from elite Sybil users are more spread out temporally, craft more convincing reviews, and have higher filter bypass rates. We also measure the impact of Sybil campaigns on various industries (such as cinemas, hotels, restaurants) as well as chain stores, and demonstrate that monitoring elite Sybil users over time can provide valuable early alerts against Sybil campaigns.

I. INTRODUCTION

User-Review Social Networks (URSNs)—such as Dianping, Yelp, and Amazon—are often the targets of Sybil attacks, where multiple fake accounts, called Sybils, are used to generate fake reviews that masquerade as testimonials from ordinary people. The goal of the attack is to deceive ordinary users into making decisions favorable to the attackers. A recent evolutionary trend is a new type of Sybil attack in contemporary URSNs, which we call *elite Sybil attacks*. Elite Sybil attacks recruit highly-rated users (*e.g.*, "Elite" member on Yelp or "5-star" member on Dianping) who normally post genuine reviews, unbiased by financial incentives. Directed by organizational leaders, elite Sybil attackers mimic the behavior of real users by posting topically coherent content with temporal patterns consistent with real users. Because elite Sybil users' review behavior greatly resembles that of genuine users, elite Sybil attacks are extremely difficult to algorithmically or manually detect. Therefore, new approaches are needed to detect elite Sybil accounts rapidly and accurately.

Challenges. Previous work on defending against Sybil attacks in Online Social Networks (OSNs) aims to identify fake or compromised accounts mainly by two means: (i) investigating an account's social network connectivity [10, 21, 41, 49, 50] relying on the trust that is established in existing social connections between users; (ii) building machine learning classifiers with a set of identified features [13, 35, 52]. The literature on Sybil defense schemes mostly targets general OSNs, and almost no reasons are tailored toward a situational logic behind that attack, much less pay attention to Sybil defenses in URSNs, such as Yelp and Dianping. URSNs pose the following three unique challenges. (i) The nodes in URSNs do not exhibit tight connectivity as in general OSNs, rendering graph-connectivity based approaches less effective in URSNs. (ii) Elite Sybil attacks in URSNs are more professional, writing elaborate reviews and posting related pictures to imitate real reviews. Thus, Sybil attacks in URSNs are more difficult to detect than those in traditional OSNs. (iii) Since elite Sybil attackers only contribute to a small fraction of overall reviews, the existing Sybil detection approaches based on the similarity of aggregate behavior do not work well. To address all these challenges and deficiencies, a novel Sybil detection technique for elite Sybil users is highly desired.

ELSIEDET. In this work, we design a novel **El**ite **Sy**bil **Det**ection system, ELSIEDET, which can identify URSN Sybil users with elaborate camouflage. Different from previous studies, we focus our design on Sybil campaigns that have multiple Sybil workers colluding to perform a task (*e.g.*, posting positive reviews and high ratings for a specific restaurant) under the coordination of a Sybil leader. These campaigns have an active time period. Any user who posts during the active time period is suspicious to be part of the campaign. This user could either

Network and Distributed Systems Security (NDSS) Symposium 2018
18-21 February 2018, San Diego, CA, USA
ISBN 1-891562-49-5
http://dx.doi.org/10.14722/ndss.2018.23009
www.ndss-symposium.org

be a benign user who happens to visit the store and post her review in the campaign period, or a Sybil user who posts fake reviews specifically for the campaign. We build ELSIEDET based on the following empirical observations: A benign user posts honest reviews based on her real experience while a Sybil user always posts fake reviews during the active time period of the Sybil campaigns. Therefore, in the long run, the more campaigns a user gets involved in, the more likely she is a Sybil user.

ELSIEDET is designed with three stages: detecting a Sybil community (Phase I), determining the Sybil campaign time window (Phase II), and finally classifying elite Sybil users (Phase III). In Phase I, since Sybil users collaborate to post fake reviews in a Sybil campaign, ELSIEDET exploits this group behavior to cluster users and identify Sybil communities. In Phase II, ELSIEDET uses a novel campaign detection algorithm to automatically determine the start and end points of a Sybil campaign, while ruling out reviews not belonging to a Sybil task. Lastly, in Phase III, we propose a novel elite Sybil detection algorithm to separate out elite Sybil users from undetected users based on a new defined metric, *Sybilness*, which scores the extent a user participates in the Sybil campaign.

We implement ELSIEDET and evaluate its performance on a large-scale dataset from Dianping, by far the most popular URSN in China. Our dataset was crawled from January 1, 2014 to June 15, 2015 and includes $10,541,931$ reviews, $32,940$ stores, and $3,555,154$ users. We show that, of all the reviews, more than $108,100$ reviews are fake reviews, which were generated by $21,871$ regular Sybil users and $12,292$ elite Sybil users. These Sybil users belong to 566 Sybil communities, which launched $2,164$ Sybil campaigns. Our research shows that the current filtering system of Dianping is ineffective at detecting fake reviews generated by the elite Sybil users since less than 33.7% of the fake reviews have been filtered by Dianping. Finally, through manual inspection, we conclude that 90.7% of randomly sampled suspicious users are elite Sybil users, and 93.8% of the $1,000$ most suspicious users are elite Sybil users. We have reported all of our findings to Dianping, which acknowledged our detection results.

Findings. Our study reveals the following main findings about the operation logistics of elite Sybil attacks.

- Motivated by economic revenue on black markets (*e.g.*, an elite Sybil user can receive up to 20 times more income than a regular Sybil user for the same task), elite Sybil users have developed a series of techniques to evade the Sybil detection systems, including coordinating the posting time and crafting carefully-polished review contents and pictures.
- We evaluate the impact of Sybil attacks on different categories of industry. Surprisingly, cinemas, hotels, and restaurants are the most active in hiring Sybil users for promotions. In particular, 30.2% of cinemas, 7.7% of hotels, and 5.5% of restaurants are actively involved in Sybil campaigns.
- We observe that 12.4% of Sybil communities post fake reviews for chain stores, which is different from recent research performed on Yelp [27]. What is more interesting is that that overhyped chain stores with the same brand recruit the same Sybil communities for Sybil campaigns.
- We find that more than 50% of Sybil campaigns can be determined within the first two weeks by only observing activities of elite Sybil users, thereby allowing the URSN to defend against the attack while in progress.

Contributions. To the best of our knowledge, our work is the first to study elite Sybil detection in URSNs. In summary, we make the following key contributions:

1) We show that the Sybil organization of Dianping has evolved to a hybrid architecture, rather than a prevalent centralized or a simple distributed system [34, 45].
2) We identify a new type of Sybil users, elite Sybil users, which employ a sophisticated strategy for evading detection and have never been studied before.
3) We characterize the behaviors of elite Sybil users and propose an early-warning system to detect online Sybil campaigns.
4) We show that ELSIEDET complements the Dianping's current filtering system, which has been verified by both our own manual inspection and the feedback received from Dianping.

Roadmap. The remainder of this paper is structured as follows: Section II introduces the necessary background on Dianping and Sybil attacks while Section III defines elite Sybil attacks. In Section IV, we propose our Sybil detection system. Section V evaluates the experimental performance, whereas Section VI provides detailed measurements of elite Sybil users and Sybil communities. Section VII discusses applications and limitations of the study. Section VIII surveys the related work. Finally, Section IX concludes the paper.

A. Ethical Considerations

In this paper, we only collected publicly available review information and its relation with stores on Dianping. We do not crawl, store, or process users' privacy information including usernames, gender, small profile pictures, or tags that often accompany the user profiles. Furthermore, we did not craft fake reviews in order to ensure that our experiments do not have a negative impact on Dianping's services. Finally, we have alerted Dianping about the discoveries and results made in this paper. We are currently discussing possibilities of our system deployment at Dianping.

II. BACKGROUND

In this section, we first briefly describe Dianping. We then summarize traditional Sybil attacks and the recent trend on User-Review Social Networks (URSNs).

A. Dianping: A User-Review Social Network

Dianping is by far the most popular URSN in China, where users can review local businesses such as restaurants, hotels, and stores. When a user uses Dianping, Dianping will return to the user with a list of choices in order of overall quality-rating. The quality-rating of a restaurant review is typically scaled from 1 star (worst) to 5 star (best), mainly depending on the restaurant service. Users are also assigned star-ratings. These star-ratings vary from 0 stars (rookie) to 6 stars (expert), depending on the longevity of the user account, the number of reviews posted, etc. A higher star-rating indicates that the user is more experienced and more likely to be perceived as an expert reviewer. Similar to "Elite User" on Yelp, a senior

level user (*e.g.*, 4-star, 5-star, or 6-star user) is supposed to be a small group of in-the-know users who have a large impact on their local community. Dianping has established its user reputation system that classifies user reviews into "normal reviews" and "filtered reviews." The latter includes the uninformative reviews or the suspicious reviews that are potentially manipulated by the Sybil attackers, but the details of the algorithm remain unknown to the public.

B. Sybil Attacks

Social media platforms populated by millions of users present either economic or political incentives to develop algorithms to emulate and possibly alter human behavior. Earlier Sybil attacks include malicious entities designed particularly with the purpose to harm. These Sybil users mislead, exploit, and manipulate social media discourse with rumors, spam, malware, misinformation, slander, or even just noise [19, 20]. This type of abuse has also been observed during the 2016 US presidential election [2]. As better detection systems are built, we witness an arms race similar to what we observed for spam alike in the past. In recent years, Twitter Sybils have become increasingly sophisticated, making their detection more difficult. For example, Sybils can post collected material searched from websites at predetermined times, emulating the human temporal signature of content production and consumption [17]. In the meantime, the arms race has also driven the corresponding countermeasures [7, 11, 13, 40].

The evolutionary chain of Sybil attacks imposes a novel challenge in the most-up-to-date URSNs: They provide fake content among little pieces of their information, regardless of their accuracy, which is highly popular and endorsed by many high-level organizers, exerting great impact against which there are no effective countermeasures. In this paper, we characterize and detect a new type of Sybil attacks in URSNs, typically applying our methodology to Dianping as our case study.

III. Dissecting Elite Sybil Attacks

In this section, we first introduce some definitions. We then define a novel type of Sybil attackers, coined as *elite Sybil users*. We finally take an in-depth dive into the typical hierarchical architecture and the key actors playing in a Sybil organization.

A. Terminology

To formulate our problem precisely, we introduce the following definitions.

DEFINITION III.1. ***Store:*** *A Store S has an official website on Dianping that contains a large number of reviews of this particular store.*

DEFINITION III.2. ***Community****: A Community C is a group of users who post reviews in similar stores to rate and comment such stores.*

REMARK III.3. *In our paper, we categorize all communities into two types:* ***benign communities*** *and* ***Sybil communities****. We define a benign community to be formulated by all benign (real, normal) users and a Sybil community to be formulated by all Sybil (malicious) users. A set of users is also partitioned into two types: A* ***benign user*** *is a person who posts honest reviews and a* ***Sybil user*** *is a person who posts fake reviews to boost the prestige of stores. We will use the terms benign users and real users interchangeably.*

DEFINITION III.4. ***Campaign:*** *A campaign—denoted as (C, S, T_s, T_e), where C, S, T_s, T_e denote community ID, store ID, starting time, and ending time of a campaign—is an activity in which users of a Community C post reviews in Store S from T_s to T_e to boost the prestige of Store S.*

REMARK III.5. *For Sybil users in a given community, these Sybil users serve for various stores. Each of these stores has one particular campaign launched by this community. However, these stores can have other campaigns, but are launched by other communities.*

B. Elite Sybil Users

In a Sybil organization of Dianping, we find a new type of Sybil users, termed *elite Sybil users*. Different from *regular Sybil users* studied before, elite Sybil users post reviews not belonging to Sybil tasks, which can harm the accuracy of existing detection systems to a large degree. Elite Sybil accounts are mainly composed of two kinds of accounts: either (i) Sybil accounts created reviews not belonging to Sybil tasks (smoke-screening) in order to mimic genuine users purely for the use of campaigns; or (ii) accounts owned by benign users—usually with high-rating stars—that convert to Sybil accounts when fulfilling a Sybil task within a campaign in order to reap the rewards offered by Sybil organizations (The Sybil task is detailed in Section III-C2.). Although elite Sybil accounts belong to multiple users/entities, they are manipulated by a single entity (*i.e.*, Sybil leader). This satisfies the definition of Sybil attack that a malicious entity takes on multiple identities. Therefore we consider the attack performed by elite Sybil accounts as Sybil attack. By hiding behind massive reasonable reviews posted however deliberately or unwittingly, these reviews posted by elite Sybil users appear realistic as those posted by benign users. Compared with regular Sybil users, elite Sybil users are more active out of the Sybil campaigns, which enables elite Sybil users to have a much lower percentage of fake reviews in their posts and higher user-level star-ratings (see Section VI).

Black market and economic factors. Here, we try to explore the monetary reward for an elite Sybil user on Dianping. Table I shows hierarchical rewards for a specific Sybil organization into which we infiltrated recently. We see that the rewards depend on the ratings of Sybil accounts. Not surprisingly, the monetary rewards earned by each *submission* increase as the ratings of accounts increase. This is largely because users with higher ratings have a larger influence, their reviews are less likely to be deleted, and thus are more attractive to Sybil organizers. Likewise, the reviews from the highly-ranked users are more influential, and have a larger chance of being presented in the front page of a store, which can potentially attract more attention from customers.

TABLE I. HIERARCHICAL REWARDS FOR (ELITE) SYBIL WORKERS

Ratings of Accounts	Rewards per Submission
0-star, 1-star	\$0.30
2-star	\$0.75
3-star	\$1.50
4-star	\$3.74
5-star, 6-star	**\$5.98**

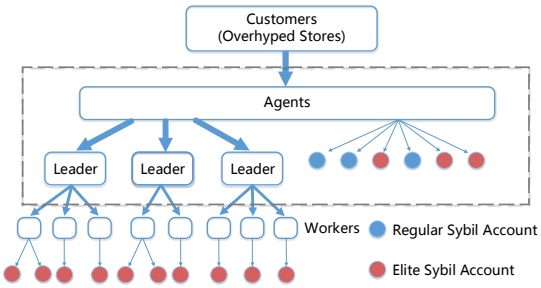


Fig. 1. The architecture of a Sybil organization

Fig. 2. An example of a fake review

C. Anatomy of Elite Sybil Attack Operations

Many review websites are suffering from review manipulation, which can be seen as a variant of Sybil attacks in URSNs. Similar to Yelp and TripAdvisor, Dianping is struggling with review manipulation as well. To investigate these organizations in depth, we impersonated Sybil users in order to investigate how the tasks are distributed and executed by the Sybil organizer. Note that, for the ethical considerations, we did not perform any real tasks in reality. In most cases, Sybil leaders regularly post contact information on social media (*e.g.*, Tencent QQ, WeChat, and Douban Group[1]) to attract Sybil workers. Specifically, we acquired contact information from Douban Group to reach out many Sybil organizations. During our month-long investigation, we found that the Sybil attacks on Dianping show a unique organization pattern.

1) A Hybrid Architecture: Sybil organizations usually show either a centralized or distributed architecture on Facebook or Twitter. The Sybil organization on Dianping, however, evolves to a hybrid architecture, which involves four key actors, as shown in Figure 1:

1) *Customers (or Overhyped stores):* Businesses that want to boost their scores rapidly on Dianping. Overhyped stores propose mission description and monetary rewards for a Sybil organization to launch Sybil campaigns. They are beneficiaries from Sybil campaigns.
2) *Agents:* Organizers are agents who are responsible for accepting the tasks from overhyped stores and upper management of a Sybil organization. Organizers take charge of launching the Sybil campaigns.
3) *Leaders:* Leaders take charge of recruiting Sybil workers and make arrangements for crafting reviews. Leaders distribute tasks to Sybil workers and payment.
4) *Elite Sybil workers:* Elite Sybil workers are Internet users, recruited by leaders, who post fake reviews for profit. These elite Sybil accounts are then manipulated by elite Sybil workers to post fake reviews. (Elite Sybil accounts, users, and workers are interchangeable in this paper.)

In this architecture, the leader plays a key role in task distribution and quality control of review comments for the following reasons: First, the leader himself/herself controls a certain number of Sybil accounts, and these facilitate the launch of a campaign. Second, to increase the impact of a campaign, the leader can also outsource a task to many elite Sybil workers, especially highly-ranked Dianping users. Finally, the leader actively participates in the review generation by directly generating the high-quality reviews by himself/herself or by closely supervising the review generation of workers. In summary, if elite Sybil workers are the puppets, then Sybil leaders are the masters who locally dominate the unique workflow of Sybil organizations on Dianping.

2) Typical Workflow: Each Sybil campaign is centered on a collection of *tasks*. For example, a campaign launched by an organization entails crafting positive fake reviews for a restaurant to boost ratings on Dianping. In this case, the owner of the overhyped store sets up the objects of a Sybil campaign, and the task is further distributed from organizers to Sybils. Each task would be "posting a single (fake) positive review online." Sybils who complete a task generate *submissions* that include screenshots of the fake reviews to be posted as evidence of his/her work (see Figure 2). The overhyped stores/agents can then verify if the work has been done to their satisfaction. It is important to notice that not all tasks can be completed because of some low-quality submissions.

The key feature of a Sybil organization on Dianping is that the Sybil leader is actively involved in the Sybil tasks. In particular, when receiving a task from the customer, a Sybil leader distributes this task to multiple elite Sybil workers and guides review generation, which is illustrated in step (1) in Figure 3.

- *Leader-supervised model*: In this model, the reviews are created by an elite Sybil worker (step 2.1) and the generated content and posting time must follow the leader's guidance and must be approved by the said leader (step 2.2).
- *Leader hands-on model*: In this model, it is the leader or the customer that generates the review comments first. The generated reviews are normally high-quality comments that include both favored comments and pictures of food or the store (step 2).

Given a certain review, the worker posts fake reviews of the specified stores (step 3). The leader will check if these crafted fake reviews exist for a period of 3 to 7 days (step 4). Once the existence of fake reviews is confirmed, the leader will pay the elite Sybil worker (step 5).

Through our investigation, we find that cultivating a 3-star elite Sybil account endorsing a tutorial offered by a Sybil organizer is priced at $6 per account. The tutorial provides details about the approach to boosting ratings of Sybil accounts and mimicking benign accounts. In concrete, (i) once an account is activated, its profile information, such as gender, date of birth, address, and profile picture, needs editing to

[1]Douban Group, being part of Douban, is composed of huge numbers of sub forums for users to post messages under various topics. https://www.douban.com/group

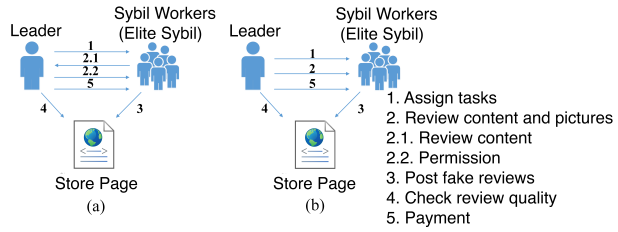


Fig. 3. The process to post fake reviews

the requirements of the tutorial. (ii) Before participating in the Sybil organization, a great number of reasonable reviews are also required to cultivate an elite Sybil account. Specially, through our sting operation in a Sybil organization on Taobao,[2] the largest C2C website in China, we find that an overhyped store, through several Sybil organizers, have collected approximately $100,000$ elite Sybil accounts. For the Sybil organization we have infiltrated, we observed 30 tasks which were assigned in three months. For a particular task, the store exploits some of these elite Sybil accounts to generate 500 fake positive reviews at the cost of around $3,000 in total. Moreover, the Sybil organization we participated in also provides an after-sales guarantee, meaning if fake reviews are deleted, it will launch a second-round elite Sybil attack. In addition, we also observe that rewards per submission on Dianping are many more than those on other Chinese websites, such as ZBJ and SDH [45]. The high monetary rewards incentivize the Sybil agents or leaders to develop sophisticated pyramid schemes to evade detection.

Based on the above discussion, it is clear that automatic detection of elite Sybil users is important to prevent Sybil attacks from URSNs. This motivated us to develop a novel framework of Sybil detection.

IV. ElsieDet: Design and Implementation

In this section, we will present three components of ElsieDet (see Figure 4): detecting Sybil communities, determining campaign time windows, and detecting elite Sybil users.

A. System Overview

ElsieDet is a three-tier Sybil detection system. In Phase I, we cluster communities based on collusion networks and perform a binary classification on detected-communities, echoing that a large number of fake reviews are usually posted by the Sybil community under the guidance of the Sybil leaders. In this phase, regular Sybil users will be clustered in Sybil communities, but most elite Sybil users are able to evade community clustering by covering up their collusion.

In Phase II, we extract time windows of Sybil campaigns from labeled Sybil communities. The rational behind the design is that a Sybil campaign has an active time period. A user posting a review towards the target store during the active time period is considered as a campaign-involved user. This user could be either a benign user who happens to visit the store and posts her reviews at that time or a Sybil user who posts fake reviews for the campaign benefits. We observe that a benign user posts reviews based on her real experience

[2]https://www.taobao.com/

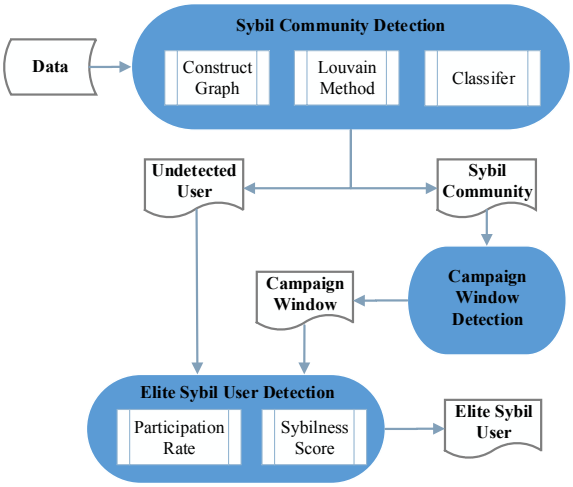


Fig. 4. System overview

while a Sybil user always posts reviews during the active time period of the Sybil campaigns.

In Phase III, followed by the undetected users and corresponding extracted Sybil campaign time windows, we first use the participation rate between users and communities to characterize the extent to which a user is related to a community; then we leverage a novel metric to determine elite Sybil users. The rationale behind the elite Sybil detection algorithm is that, through using elaborate reviews to obfuscate their fake reviews, the elite Sybil users are motivated to participate in multiple Sybil campaigns due to a high economic rewarding. Therefore, the more campaigns a user participates in, the more likely the user is an elite Sybil user.

B. Sybil Community Detection

In this section, we present the details of detecting the Sybil community. The first step of Sybil community detection is *constructing the Sybil social links* for the Sybil users belonging to the same Sybil community. It is defined that two users belonging to the same community have a Sybil social link if they have *collusive reviews*, which are similar reviews posted by two users according to the same store. Based on these virtual links, we further define a novel metric, *pairwise similarity metrics*, which measure the similarity among the users. Then, we adopt the Louvain method [6] to extract communities from the entire network. Finally, we perform classification to identify the Sybil community from the benign community.

*1) **Constructing Sybil Social Links via Collusive Reviews**:* To cluster and identify the Sybil community, the first step is to build the social links between the Sybil users, which are coined as *Sybil social links*. In general, two users belonging to the same community and having the *collusive reviews* posted in the same store or restaurant are defined to have a Sybil social link. Specifically, a tuple abstraction of a user's single review is referred to as (U, T, S, L), where U, T, S, and L represent user ID, review timestamp, store ID, and star-rating of a review, respectively. For users u and v, we derive **review sets** associated with u and v, respectively:

$$\mathcal{R}(u) = \{(U, T_1, S_1, L_1), (U, T_2, S_2, L_2), \cdots, (U, T_n, S_n, L_n)\};$$
$$\mathcal{R}(v) = \{(V, T'_1, S'_1, L'_1), (V, T'_2, S'_2, L'_2), \cdots, (V, T'_m, S'_m, L'_m)\}.$$

For all pairwise users u and v, and for a given k, $(U, T_k, S_k, L_k) \in \mathcal{R}(u)$, we define $P_u(k) = 1$ if there exists $(V, T'_l, S'_l, L'_l) \in \mathcal{R}(v)$ such that the following three properties are true:

1) The two reviews are posted in the same store: $S_k = S'_l$;
2) The two reviews are created within a fixed time slot ΔT: $|T_k - T'_l| \leq \Delta T$;
3) Both two reviews are 1-star or both of them are 5-star: $L_k = L'_l =$ 1-star or $L_k = L'_l =$ 5-star.

Otherwise, $P_u(k) = 0$.

Note that in previous research [11], Cao *et al.* simply defined two collusive reviews if they pertain to the same constraint object and their timestamps fall into the same fixed time slot, but these two collusive reviews defined are not mathematically equivalent.

Measuring similarity is key to grouping similar users. Different from the previous research [11, 40] using Jaccard similarity metric, we measure the similarity between pairwise users u and v as follows:

$$\begin{aligned} \text{Sim}(u, v) &= \frac{\sum_{k=1}^{n} P_u(k) + \sum_{l=1}^{m} P_v(l)}{|\mathcal{R}(u)| + |\mathcal{R}(v)|} \\ &= \frac{\sum_{k=1}^{n} P_u(k) + \sum_{l=1}^{m} P_v(l)}{n + m}. \end{aligned} \tag{1}$$

Note: $\text{Sim}(u, v) = \text{Sim}(v, u)$.

In summary, we model an Sybil community as an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $u \in \mathcal{V}$ is an user account and each edge $(u, v) \in \mathcal{E}$ represents a Sybil social link among users u and v if and only if $\text{Sim}(u, v) > \beta_{\text{Thre}}$.[3] Then users u and v are defined as *neighbors*.

*2) **Community Clustering via the Louvain Method**:* We then employ a community detection method, termed Louvain method [6], to detect communities on Sybil social links. The Louvain community detection method iteratively groups closely-connected communities together to improve the partition modularity. In each iteration, every node represents a community, and well-connected neighbor nodes are combined into the same community. The graph is reconstructed at the end of each iteration by converting the resulting communities to nodes and adding links that are weighted by the inter-community connectivity. The entire process is repeated iteratively until it reaches the maximum modularity. Each iteration has a computational cost linear to the number of edges in the corresponding graph and typically the process just requires a small number of iterations.

It is noted that community detection algorithms have been proposed to directly detect Sybils [41]. They seek a partition of a graph that has dense intra-community connectivity and weak inter-community connectivity. For example, the Louvain method searches for a partition with high modularity [6]. However, we find that it is insufficient to uncover massive Sybil users within Louvain-detected communities. In the following step, we apply supervised machine learning to Louvain-detected communities.

[3]The threshold β_{Thre} is tuned to optimize the following community classification in terms of accuracy. Community classification results obtained by multiple supervised learning techniques are not overly-sensitive to the different thresholds chosen.

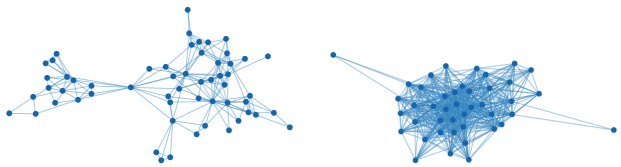
*3) **Sybil Community Classification**:* Next, we apply machine learning classifiers to discriminate Sybil communities from benign ones. The reason behind this is that some communities contain users who reside close-together or visit the same venues. To accurately characterize these observations, we apply eight features with respect to three types (tabulated in Table II) to our binary classifiers. The output is each community labeled either benign or Sybil. We validate this intermediate step in Section V-B.

TABLE II. TYPES OF FEATURES

Types of Features	Features
Community-based Features	Score deviation, Average number of reviews, Entropy of the number of reviews in each chain stores, Entropy of districts of stores
Network Features	Average similarity, Global clustering coefficient
User-based Features	Unique reviews ratio, Maximum number of duplication

(a) Community-based features. There are four types of Community-based features: score deviation, reviews per store, entropy of chain stores, and entropy of districts of stores. *Score deviation* and *Average number of reviews* are self-explanatory. To achieve the Sybil tasks, score deviation of reviews posted by Sybil users will become larger. *Entropy of the number of reviews in each chain stores* is the expected value of information contained in each of the chain stores by measuring the number of reviews occurred. We use this feature because some Sybil users post reviews only in chain stores. *Entropy of districts of stores* is a location-based feature to characterize mobility patterns of Sybil users that are driven by Sybil tasks. We therefore use *Entropy of districts of stores* to show this difference.

(b) Similarity-based network features. We redefine the network via Sybil social community construction since benign and Sybil communities have remarkable differences with respect to the graph structure (see Figure 5(a) and Figure 5(b)). We use *Average similarity* and *Global clustering coefficient* to show the difference according to the redefined graph structures. *Average similarity* is the average similarity between pairwise users in a community. Sybil users in a Sybil community are assigned tasks for similar stores, but users in a benign community randomly choose stores to post reviews. Hence, similarity values between Sybil users are greater than those between benign users. *Global clustering coefficient* is used to measure the degree in which nodes in a graph tend to cluster together. Sybil users have the characteristics of team working, so they are more likely to be clustered together.



(a) Graph structure of a benign community (community 7914)

(b) Graph structure of a Sybil community 7924 (community 7924)

Fig. 5. Comparison of the graph structure between a benign community and a Sybil community

(c) User-based features. Since community-based features may lose information of users, we then abstract the user-based

features of each user and aggregate them as a feature of the community. By analyzing Sybil communities, we observe that some Sybil users will repeatedly post reviews in the same store. We therefore define two features, *Unique reviews ratio* and *Maximum number of duplication*, to reflect this user-level behavior. Lastly, we do not use linguistic or contextual features because these features are not so effective in the URSN setting [30].

C. Campaign Window Detection

To detect the time window of a Sybil campaign, one potential approach is detecting sudden increases in rating, ratio of singleton reviews, and the number of reviews by leveraging a burst detection algorithm (*e.g.*, Bayesian change point detection algorithm [15]). However, on Dianping, Sybil campaign detection results based on burst detection may not be reliable in practice. For example, the sudden increases in ratings or the number of reviews may be contributed by some unexpected factors such as offline promotions. An observation is that a store tends to entice its customers to write favorable reviews as the return of a discount coupon in promotion seasons.

Different from the previous research, our proposed solution focuses on detecting the anomaly collaborative behaviors of Sybil community. We interpret the algorithm of campaign window detection in the following. The Algorithm 1 takes as input a list L_{review} that represents the number of reviews posted each week and does the following:

1) Initializes the start and end points of the campaign window (Line 1 through Line 2).
2) Iteratively finds and deletes sparse review intervals within the campaign window (Line 3 through Line 14).
   a) Finds the first left and right sparse review intervals within the campaign window. If none, the functions will return the entire campaign window (Line 4 through Line 5).
   b) If there is no sparse review interval on either side, breaks the loop (Line 6 through Line 8).
   c) Removes the sparse review interval. This can prevent deleting major parts of the campaign window (Line 9 through Line 13).

The output of Algorithm 1 is the start point and the end point of each Sybil campaign accordingly.

ALGORITHM 1: Detecting Campaign Time Windows

Input: A list L_{review} whose item $L_{review}[i]$ denotes the number of reviews posted in the ith week.
Output: The start point l and end point r of the campaign time window.
Initial:
1: $l \leftarrow 0$;
2: $r \leftarrow length(L_{review}) - 1$;
3: **while** (**true**) **do**
4: $I_{l,l'} \leftarrow find(left, l)$; {Find the first sparse interval $I_{l,l'}$ from left.}
5: $I_{r',r} \leftarrow find(right, r)$; {Find the first sparse interval $I_{r',r}$ from right.}
6: **if** ($l' = r$ **and** $r' = l$) **then** {There is no sparse interval.}
7: **break**;
8: **end if**
9: **if** ($|I_{l,l'}| \leq |I_{r',r}|$) **then** {Choose the interval with fewer reviews.}
10: $l \leftarrow l' + 1$;
11: **else**
12: $r \leftarrow r' - 1$;
13: **end if**
14: **end while**
15: **return** l, r;

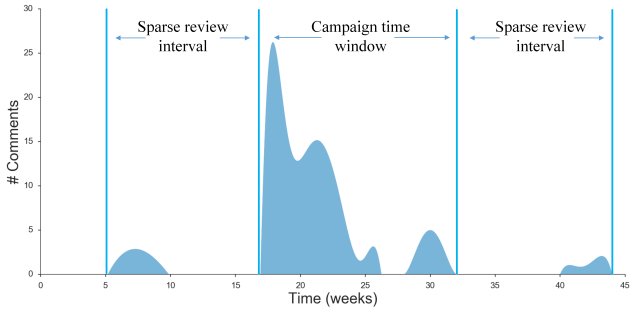


Fig. 6. An example of campaign time window detection

As shown in Figure 6, it is observed that a campaign period is comprised of multiple segment periods. We are interested in those segment periods in which the Sybil users are active and thus we need to filter out those periods when the Sybil users are inactive. To achieve this, we introduce the concept of Sparse Review Interval, which is used to indicate whether or not the users are active in this time period. In particular, *a sparse review interval $I_{i,j}$ (where i represents the start point of the ith week and j represents the end point of the jth week) is referred to as the period in which the number of weeks with at least one review is less than the number of weeks without any reviews.* As shown in Figure 6, with a long time period, the entire time interval can be seen as a sparse review interval. In order to avoid removing intervals with massive reviews, our strategy is scanning the time period from both left and right to find the first sparse review intervals respectively, and then removing the sparse review interval with fewer reviews. We repeat this process until there is no sparse review interval and the remaining period is the targeted campaign period.

D. Elite Sybil User Detection

Recall that elite Sybil users are those who often post reviews not belonging to Sybil tasks like a benign user but occasionally post fake reviews. The primary reason that the existing Sybil detection approaches cannot effectively detect elite Sybil users is that reviews not belonging to Sybil tasks decrease the similarity between elite Sybil users. Labeling all reviews of an elite Sybil user as fake reviews may misjudge some real reviews, which will take away the enjoyment of the service. In order to detect elite Sybil users, we take as input the time windows of Sybil campaigns and corresponding undetected users. Then we define the participation rate and *Sybilness* that is the perceived likelihood to output an elite Sybil user. Finally we use *Sybilness* to quantify each review.

Participation rate between users and communities. We first define participation rate between users and communities to characterize the extent to which a user is related to a community. Based on our observations, we assume that the more campaigns a user participates in, the more likely the user is an elite Sybil user. Given a community C, we define:

- $N_C(k)$: the accumulated number of reviews posted within the kth time window of community C.
- $N_C^{\max}$: the maximum number of reviews posted within all time windows of community C.

We then "normalize" the number of reviews in the kth time window by $P_C(k) = \frac{N_C(k)}{N_C^{\max}}$, for a given C. $P_C(k)$ will help indicate the importance of a time window, since the larger the number of reviews is within a time window, the more active

this campaign is in the community. Then for a given user u in community C, we can calculate the "weighted sum" of the number of reviews u posts by:

$$N_{u\in C} = \sum_k P_C(k) \cdot N_{u\in C}(k), \tag{2}$$

where $N_{u\in C}(k)$ represents the number of reviews u posted within the kth time window of C. We finally plug $N_{u\in C}$ into a standard sigmoid function to measure the participation rate $\rho_{u\in C}$ between u and C:

$$\rho_{u\in C} = \frac{1}{1+\exp^{-\frac{N_{u\in C}-\mu_C}{\sigma_C}}}, \quad \text{for any } u \in C, \tag{3}$$

where μ_C and σ_C are the mean and the variance of $N_{u\in C}$ for all users u in C.

Sybilness. *Sybilness* score is a perceived likelihood indicating if a user is an elite Sybil user. Since simultaneously participating in multiple communities leads to the large cardinality of C but small $N_{u\in C}$ and $\rho_{u\in C}$, just considering about the participation rate $\rho_{u\in C}$ will fail to tease out elite Sybil users. We then take $\rho_{u\in C}$ into consideration to construct the final index, *Sybilness*, to determine a specific user's legitimacy. To be specific, for assigning a *Sybilness* score f to each user u on Dianping, we take a weighted average method on $N_{u\in C}$ with respect to each of the corresponding coefficients $\rho_{u\in C}$, for all C, as shown below:

$$f(u) = \sum_C \rho_{u\in C} \cdot N_{u\in C}. \tag{4}$$

Eventually, we use the *Sybilness* score $f(\cdot)$ to determine the perceived likelihood that a user is an *elite Sybil user* or not (Note: *Sybilness* score here can be greater than 1.).

Annotating reviews posted by elite Sybil users. Since not all reviews posted by elite Sybil Users are fake, we annotate each reviews with a score defined as $\rho_{u\in C} \cdot P_C(k)$, for any k. This score can be used as a criterion to filter fake reviews or regulate the frequency of CAPTHCHAs.

V. Evaluation

We implement ElsieDet and evaluate it on a large-scale dataset of Dianping. Our evaluation covers the following aspects: Sybil community detection, elite Sybil user detection, and system performance.

A. Data Collection

In this section, we will introduce the datasets used and propose the methodology we use to gain the ground-truth data.

Dataset. We develop a Python-based crawler to analyze HTML structure of store pages and user pages on Dianping. All reviews were crawled by the web crawler from January 1, 2014 to June 15, 2015. Starting from the four hand-picked overhyped stores (the seed list) in the training set belonging to the same Sybil organization, which we discovered during our month-long investigation. We then crawled outwards—crawling one level down of all users who wrote reviews in these stores and extended the store list that was commented by these users. Second, we crawled all reviews appearing in these stores and collected all users of these reviews to form a user list. The web crawler repeated these steps until reaching $32,940$ stores on the store list. Eventually, our resulting data set has $10,541,931$ reviews, $32,933$ stores, and $3,555,154$ users. We will make all of our data used publicly available in the future. Furthermore, we categorize the stores crawled into 13 types (see breakdowns in Table III). In Table III, the 13 categories are shown in decreasing order in terms of percentage of overhyped stores. Followed by our detection methodology, surprisingly, we find that more than 30% overhyped stores are pertinent to cinemas. The main remaining overhyped stores are hotels, restaurants, and places of entertainment.

TABLE III. Breakdowns of stores

Type	# Stores	# Overhyped Stores	Percentage of Overhyped Stores
Cinema	235	71	$\mathbf{30.21\%}$
Hotel	1,738	134	$7.71\,\%$
Restaurant	22,474	1,244	$5.54\,\%$
Entertainment	1,384	73	$5.27\,\%$
Wedding Service	320	8	$2.50\,\%$
Beauty Store	1,460	35	$2.40\,\%$
Fitness Center	326	7	$2.15\,\%$
Living Service	863	10	$1.16\,\%$
Scenic Spots	1,243	14	$1.13\,\%$
Shopping	2,466	22	$0.89\,\%$
Infant Service	216	0	$0\,\%$
Car	148	0	$0\,\%$
Decoration Company	67	0	$0\,\%$

Ground-truth dataset. Similar to the previous research [13, 29, 40], we rely on manually labeled data for Sybil community detection. In order to classify the communities as benign or Sybil using supervised learning, a subset of the communities needs to be labeled. To carry out the labeling, we actively exchanged ideas with Dianping of how high-profile Sybil users resemble. Particularly, the final manual labeling considers the following three criteria. If two of them are satisfied, then a community is labeled as a Sybil community.

(a) Massive filtered reviews by Dianping signify that a large proportion of reviews posted in a community are filtered by Dianping's Sybil detection system. Reviews that Dianping has classified as illegitimate using a combination of algorithmic techniques, simple heuristics, and human expertise. Filtered reviews are not published on Dainping's store/user pages. If we find a great proportion of reviews existing in our dataset but missing on Dianping's main listings, this indicates that these reviews have been filtered. Although a review can be filtered for many reasons, such as overly-florid or low-quality reviews, filtered reviews are, of course, partial indicators of being Sybil. If massive reviews have been filtered in a community, then the community has a high possibility to be Sybil.

(b) Duplicate user reviews mean that reviews posted by a user belonging to a community only serve one or two store(s) with similar content. To our observation, reviews posted by a benign user of a community are often evenly distributed in miscellaneous stores. The existence of *duplication* signifies that Sybil users are more addicted to boosting review ratings in only a few stores in a community. This feature is stricter than the collusive reviews defined in this paper.

(c) Spatio-temporal review pattern means that an unusual sudden jump with respect to the number of reviews of a target restaurant/store in a community is consistent with a collusive action of the Sybil community, by rule of thumb. Normally, the reviews of a store are evenly distributed since its inception. Hence, if many stores appearing in a community demonstrate unreasonable spatio-temporal patterns, then the community is highly likely to be Sybil.

To do this, we did not hire Amazon Mechanical Turk (AMT) to accomplish the tasks because scrutinizing those reviews requires deep familiarity with Chinese language and the Dianping platform *per se*. Instead, we hired 5 Chinese undergraduate students to classify communities as either benign or Sybil. For the rare cases where there was not a consensus, we used voting. For example, a community would be labeled as Sybil if and only if the 5 votes are SSSBB, SSSSB or SSSSS, with S representing Sybil and B representing benign.

B. Results and Detection Accuracy

Accuracy of Sybil community detection. For the dataset used, ELSIEDET detects in total 710 communities. By using the multiple criteria shown above, we randomly picked up 170 communities as ground truth and labeled 117 Sybil communities as well as 53 benign communities. The assumption that a community only takes a binary classification can be justified by the empirical percentage of Sybil (resp. benign) users taking up in the designated Sybil (resp. benign) communities. To justify this, we took a look at each of the $1,969$ users of 74 communities (54 Sybil vs. 20 benign) obtained from ground truth (which is more than 10% of the total amount of communities), still by following the above criteria to check each user in communities. We conclude that 96.85% of the users are designated to the correct community labels. With 8 features tabulated in Table II, we also compare several classifiers implemented by *scikit-learn* library [1]. We perform grid search to determine optimal parameters for each classifier and evaluate their performance on weighted *precision*, weighted *recall*, weighted *F1 score*, and *AUC* (Area under the Curve of ROC) using 5-fold cross-validation. As shown in Table IV, support vector machine (SVM) performs best among all classifiers with 96.45% F1 score and 99.42% AUC, using Gaussian (RBF) kernel with parameters chosen $\mathtt{C} = 18$ and $\gamma = 0.09$.

TABLE IV. CLASSIFICATION PERFORMANCE

Classifier	Precision	Recall	F1	AUC
Decision tree	93.80 %	92.90 %	93.60 %	92.83 %
SVM	96.74 %	96.47 %	**96.45%**	**99.42%**
GNB	94.21 %	93.44 %	93.57 %	97.64 %
KNN	96.75 %	96.47 %	96.50 %	97.45 %
Ada boost	93.84 %	93.54 %	93.60 %	97.92 %
Random forest	93.16 %	94.01 %	92.99 %	97.42 %

We then apply our trained classifiers to predict each community. As a result, ELSIEDET identifies 566 Sybil communities with $22,324$ users, and 144 benign communities with $5,222$ users. Surprisingly, detected Sybil communities significantly outnumber detected benign communities. It is perhaps because in the community clustering process, the constraints of posting time and review ratings pose limitations on forming benign communities. Most benign users are thereby pruned by applying the Louvain method.

Recall in Section II-A, we note that not all filtered reviews are fake reviews (some are viewed as useless reviews). Through our experiments, we confirm that the fake reviews classified are more likely to be filtered. As shown in Figure 7, we compare the percentage of filtered reviews in benign and Sybil communities, respectively. We observe that the percentage of filtered reviews of Sybil communities significantly outweighs that of benign communities with respect to the same cumulative probability. Specifically, we see that 80% of

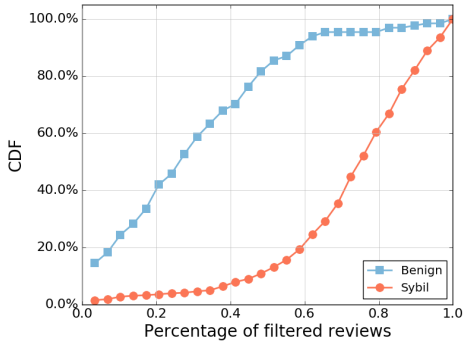


Fig. 7. Comparison between benign and Sybil communities by percentage of filtered reviews

Sybil communities (resp. benign communities) have more than 80% (resp. less than 50%) of reviews filtered. We conclude that filtered reviews are more likely fake, which validates the accuracy of our detection methodology.

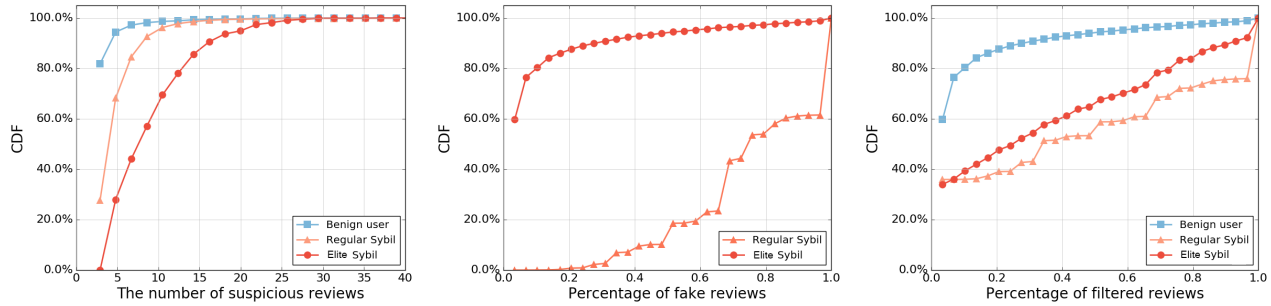
Accuracy of elite Sybil users detection. ELSIEDET considers a user u as an elite Sybil user if the following two conditions hold: (i) if the user u does not belong to any community; and (ii) the user participation rate $\rho_{u \in C}$ is larger than 0.5 (that is, the average participation rate of users in community C), for any community C. According to this criterion, we label $12,292$ elite Sybil users in total. Instead of binary classification, ELSIEDET ranks elite Sybil users according to the *Sybilness* score function (see Equation (4)).

To carry out the ultimate validation on elite Sybil users detected from ELSIEDET, we rely on human knowledge. In concrete, for each detected elite Sybil user, we manually categorize his or her reviews into two types, suspicious reviews and normal reviews, by inspecting Sybil campaign time intervals. The manual check then considers the following three criteria (by rule of thumb): *(i) this user is involved in vast Sybil campaigns*; *(ii) the intent of suspicious reviews is aligned with that of Sybil campaigns*. For example, in order to boost reputation in a Sybil campaign, the suspicious reviews should be 5-star; *(iii) suspicious reviews set apart from normal reviews in terms of spatio-temporal characteristics*. If a user satisfies all three criteria, we validate that he or she is an elite Sybil user. We emphasize that the criteria of manual validation are stricter than holding the two conditions carried out by ELSIEDET.

Finally, of all the top $1,000$ suspicious elite Sybil users that our system flags, through manual validation, we conclude that 938 are indeed elite Sybil users, which leads to a precision rate of 93.8%. We also randomly sampled $1,000$ flagged users to generalize the validation results, which also leads to a high precision rate of 90.7%.

C. System Performance

We evaluate the efficiency of ELSIEDET in a server with Intel CPU E3-1220 v3 @ 3.10GHz and 16G memory. Since ELSIEDET has to compute potential collusion set and the pairwise similarity between potential collusive users to construct Sybil social links, this step would be the bottleneck of efficiency. Instead, we implement a parallel program for this step based on the observation that the computation for each user is independent. Finally, we implement a single-threaded program to complete following steps. Specially, for Dianping's dataset, the step of computing the pairwise similarity takes



(a) Comparison on the number of fake reviews (b) Comparison on the percentage of fake reviews (c) Comparison on percentage of filtered reviews

Fig. 8. Comparison between elite Sybil users and regular Sybil users

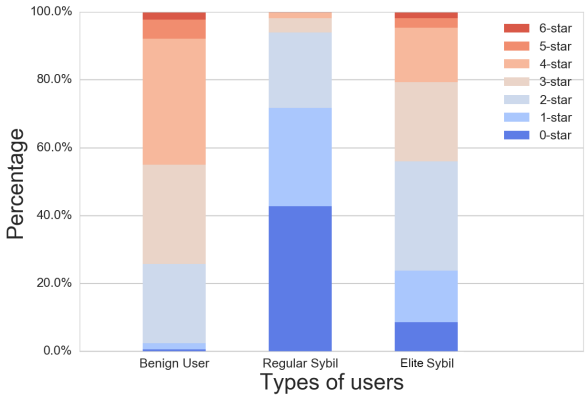


Fig. 9. Comparison on distribution of user-level star ratings

approximately 110 minutes and the remaining steps take approximately 22 minutes.

VI. Measurement and Analysis

In this section, we analyze the behavior of elite Sybil users and communities. First, we compare the behavior patterns among benign users, regular Sybil users, and elite Sybil users. We then discuss the relation between Sybil communities and elite Sybil users and review manipulation in chain stores. We study reviews, not belonging to Sybil tasks, posted by elite Sybil users to speculate their strategies to camouflage fake reviews. Finally, we demonstrate two temporal dynamics characterized by user posting period and Sybil campaign duration.

A. Comparison with Regular Sybil Users

Here, we try to explore the distribution of different types of user levels. Figure 9 shows that the distribution of levels of users is unevenly distributed for each type of users. As we can see from Figure 9, we find that the distribution of levels of benign users is almost symmetrically bell-shaped, centered at 3-star. In contrast, the levels of regular Sybil users are heavily skewed toward low-level. Based on our results, we observe that the levels of elite Sybil users detected are biased more toward high-level than those of regular Sybil users.

Comparing elite Sybil users with regular Sybil users at the micro level, we show that elite Sybil users post more fake reviews, are more spread out temporally, and have fewer reviews filtered by Dianping. Figure 8 compares the behavioral patterns among elite Sybil users, regular Sybil users, and benign users on Dianping.

Figure 8(a) plots the CDF of the number of users in terms of the number of suspicious reviews posted. As can be seen in Figure 8(a), elite Sybil users post the most suspicious reviews among all. This demonstrates that elite Sybil users cater to market demand due to their potential larger impact on Dianping ranking and higher prices for the customers. For the regular Sybil users, their strategy is frequently changing their low-level accounts to evade the detection since it is easy to apply or buy with a low cost for low-level accounts.

Figure 8(b) plots the CDF of the number of users in terms of the percentage of fake reviews posted. As we can see, fake reviews are significantly more often generated by regular Sybil users than by elite Sybil users, which echoes our definition that elite Sybil users post massive reviews not belonging to Sybil tasks (smoke-screening) to mimic genuine users. Surprisingly, the distribution of regular Sybil users roughly follows the Pareto principle (also known as the 80-20 rule) that more than 60% of all the reviews posted by 20% of regular Sybil users are fake. In contrast, as we can see from Figure 8(b), we show that only 20% of all the reviews posted by more than 80% of elite Sybil users are fake, recognizing that the principle also applies in reverse.

Figure 8(c) plots the CDF of the number of users in terms of their percentage of filtered reviews. As can be seen from Figure 8(c), we show that the percentage of filtered reviews of regular Sybil users significantly outnumbers that of benign users with respect to the same cumulative probability. To be specific, we see that 80% of Sybil users (resp. benign users) have more than 90% (resp. less than 20%) of reviews filtered. This user-level observation is consistent with the community-level results shown in Figure 7. In addition, elite Sybil users have fewer reviews filtered by Dianping mainly because a large portion of their reviews are not assigned to any task.

B. Community Structure

Understanding the behaviors of elite Sybil users is important to reveal the characteristics of the (quasi) permanent workforce of Sybil organizations on Dianping. Looking at the macro level, communities of elite Sybil users form large-scale sparsely knit networks and their graph density is much lower.

Figure 10 shows an example of an induced network structure of elite Sybil users. In the figure, a dot represents an elite Sybil user, a square represents a Sybil community, an edge between a dot and a square represents that an elite Sybil user belongs to a community, and a red (resp. blue) dot

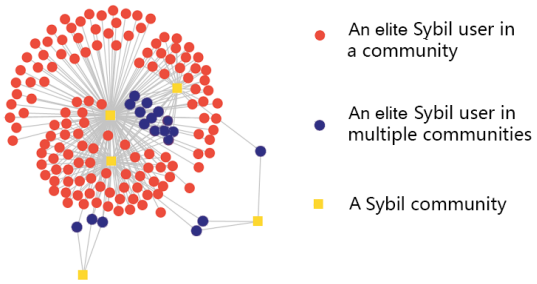


Fig. 10. Relation between elite Sybil users and communities

shows that an elite Sybil user belongs to a single community (resp. multiple communities). As can be seen, we observe that many elite Sybil users are correspondingly connected to a single community, forming a large-scale sparsely knit network. We also show that some elite Sybil users appear in multiple communities. Ranked by *Sybilness*, we pick up the top $1,000$ elite Sybil users out of all $12,292$ users in our collection. There are 824 elite Sybil users who participated in a single community, 160 who participated in two communities, and 16 who participated in at least three communities. Not surprisingly, we clearly show that these elite Sybil users are sparsely connected and their graph density is much lower than that of regular Sybil users.

C. Review Manipulation for Chain Stores

Recent research from Harvard [27] pointed out that it is less likely for chain stores to hire Sybil accounts to generate favorable reviews. Chain stores tend to depend heavily on various forms of promotion and branding to establish their reputation. This is because chains receive less benefit from reviews, and they may also incur a larger cost if they are caught posting fake reviews, destroying their brand image. However, our research contradicts this statement. We find that a series of chain stores leverage Sybil organizations to post fake reviews to manipulate their online ratings.

To be more specific, of all 566 Sybil communities in our dataset, it is observed that 12.37% of Sybil communities post fake reviews for chain stores listed on Dianping. The number of chain stores involved varies from 2 to 11. One possible explanation is that the chain stores hired the same Sybil agent, who recruited the same Sybil community for Sybil campaigns.

Figure 11 shows the main part of the entire network structure of Sybil communities and overhyped stores, pruned by a small portion of tiny networks. In the figure, a yellow square represents a Sybil community, a red dot represents an overhyped store, and an edge between a yellow and a red dot represents that a Sybil community connects to an overhyped store. As can be seen, almost all Sybil communities act as central nodes. This indicates that these Sybil communities not only launch campaigns for a single store, but also provide various services for a huge number of overhyped stores who are connected by the network. Furthermore, some overhyped stores connect to multiple communities, which indicates that they have employed Sybil communities more than once (A case study is detailed in Section VI-F.). We also label chain stores that have at least five branches with different colors other than red. These chains are connected to the same communities, respectively, possibly sharing similar reviews and having the same goal.

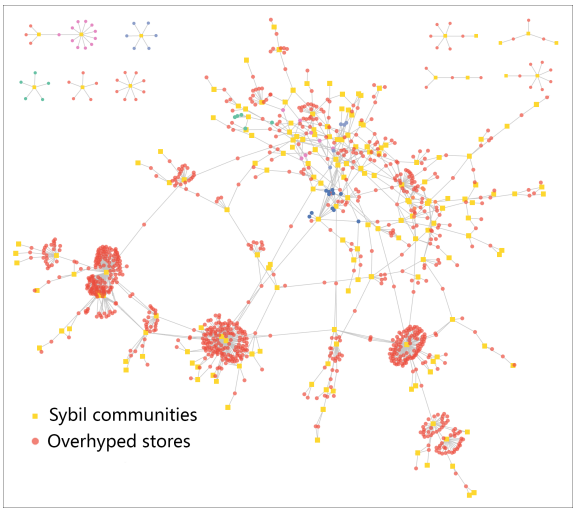


Fig. 11. Relation between Sybil communities and the overhyped stores

D. Early Alerts for Sybil Campaigns

In this subsection, we will show that it is feasible to uncover Sybil campaigns through monitoring our detected elite Sybil users. In particular, by continually monitoring the collusive behaviors of elite Sybil users, the social network operator can determine whether a Sybil campaign has been launched at the earliest stage, which serves as an early alert for a Sybil campaign.

Detecting Sybil campaigns via monitoring elite Sybil users. Our goal is to detect the presence of a Sybil campaign at the early stage based on identifying elite Sybil users via continually monitoring all elite Sybil users. To do this, we simply apply 7-day slide windows along the timeline to each store so as to detect campaigns. The rule of determining a Sybil campaign is more than a predetermined threshold number (*e.g.*, 7 in our experiment) of reviews that the elite Sybil users posted at the same store within a 7-day slide window. Our heuristic is that, in the non-campaign period, the elite Sybil users normally post reviews at different stores in similar ways as innocent users due to their different living habits, walking routines, or shopping preferences. However, only within the campaign period, the elite Sybil users collusively post reviews at the same stores to fulfill the Sybil campaign tasks. The evaluation results show that by scanning the activities of elite Sybil users during the entire campaign period, approximately 90.40% campaigns can be determined. This indicates that the campaign determination rule holds for almost all the Sybil campaigns.

Determining Sybil campaigns at the early stage. An interesting question is whether we can determine a Sybil campaign at the early stage. The benefits of early detection is that it can give a competitive advantage for the system operator to take countermeasures against Sybil campaigns. We run the campaign window determination algorithm by using the first $1/4$, $1/3$, and $1/2$ of the entire campaign period. The evaluation results show that 56.77%, 63.08%, and 75.14% of campaigns can be successfully detected correspondingly. Since the average Sybil campaign period is 68 days in our experiments, it indicates that more than 50% of Sybil campaigns can be determined within the first two weeks by only observing activities of elite Sybil users, thereby triggering lightning strikes on Sybil campaigns.

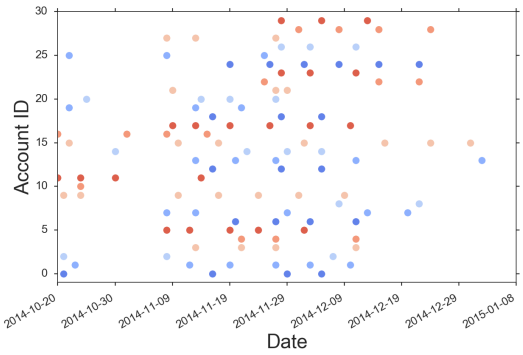


Fig. 12. Reviews posted by Community 4559 in Store 4112200

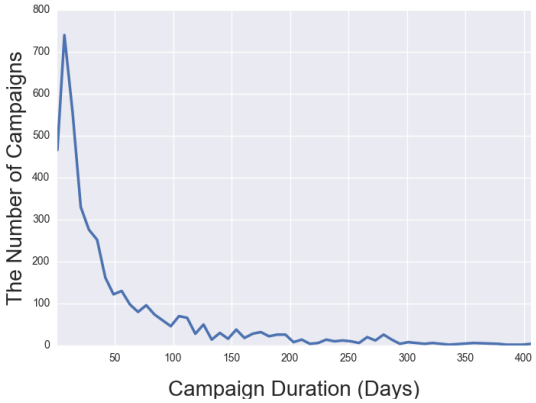


Fig. 13. The distribution of number of campaigns across campaign duration

E. Temporal Dynamics

We demonstrate two temporal dynamics characterized by user posting period and Sybil campaign duration.

User posting period. Figure 12 shows that elite Sybil users in Community 4559 repeatedly post fake reviews in Store 4112200. In the Figure, the x-axis shows the time when an account posts a review, and the y-axis is the account's ID. A dot (x, y) in the figure represents that an account with ID y posts a review at time x. We use staggered colors to encode reviews posted by different users. As we can see from Figure 12, 33 users in Community 4559 posted 127 reviews within a period of two months. Posting reviews by these users is much denser than by benign users. Apart from posting reviews within a short time period, these elite Sybil users also deliberately manipulate posting time of reviews. For example, some elite Sybil users even periodically (every week/month) post fake reviews. We emphasize that manipulation of posting temporal dynamics is key to orchestrating the evasive strategy.

Sybil campaign duration. By applying the campaign window detection algorithm, we finally obtain $4,162$ Sybil campaigns. Figure 13 shows the distribution of number of campaigns across campaign duration. As we can see from Figure 13, the distribution is unimodal with a sudden spike at 7 days for the x-axis, echoing our 7-day slide windows selected; then largely monotonically decreasing beyond 50 days. More remarkably, we observe there are 466 1-day ephemeral Sybil campaigns as shown by the y-intercept of Figure 13. In these campaigns, Sybil communities generally complete a task fleetingly.

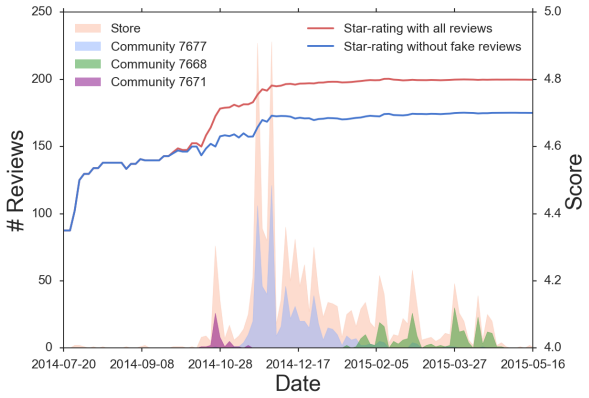


Fig. 14. Variation of star ratings and the number of reviews of a hotel

In this section, we will detail a case study of Sybil communities and campaigns and illustrate various strategies to evade the Dianping's Sybil detection system.

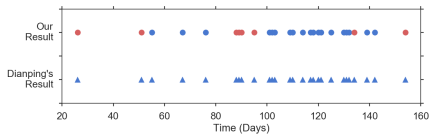
F. Sybil Communities and Sybil Campaigns

Recall in Figure 11, we show a part of stores employ several Sybil communities to increase their star ratings. Here, we zoom in and show the first case study that is about a hotel employing three different Sybil communities to post fake reviews. Figure 14 shows that the variation of the star rating and the number of reviews change over time. Orange represents aggregate reviews of a hotel; blue, purple, and green represent reviews coming from three respective Sybil communities, respectively. The red line denotes the star rating of a hotel and the blue line denotes the star rating without detected fake reviews.

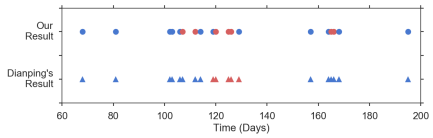
As we can see from Figure 14, many spikes occurred, generated by three Sybil communities, always correspond to the spike of the total number of reviews. This indicates that these fake reviews causing sudden spikes are taken into effect to raise the star rating of the hotel. In addition, as pointed out from Figure 14, red and blue lines are overlapping before the first spike; the red line then increases sharply afterwards but the blue line maintains a moderate growth. This indicates that these fake reviews posted by Sybil communities do have an impact on distorting the online rating. Figure 14 also implies that Community 7677 commits the largest-scale fake reviews and contributes most to increasing the star rating. However, Community 7668 launches a fairly long-term Sybil organization but takes a very "moderate" gain on the star rating. This is perhaps because the hotel has had accumulated a significant number of reviews previously. Another possible reason is that the secret ranking algorithm adopted by Dianping does not merely depend on the average rating of a store. Features, such as the number of reviews and the number of page views, are another factors to determine the rank of a store. Hence, although these reviews do not have a discernible impact on the average star rating, they may also affect ranking results on Dianping.

G. Evading Dianping's Sybil Detection System

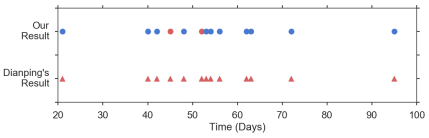
In this case study, we present three examples of elite Sybil users in the same community to attempt to illuminate the evasive strategy taken by elite Sybil users. We also compare the results processed by Dianping's filtering system with ours.



(a) No reviews filtered



(b) Partial reviews filtered



(c) All reviews filtered

Fig. 15. Different detection results for elite Sybil users

As shown in Figure 15, each subfigure corresponds to the reviews posted by each elite Sybil user. Each dot represents a review posted according to the timeline; the upper (resp. lower) dotted line of reviews represents the posting timeline generated ELSIEDET (resp. Dianping filtering system). In specific, each blue (resp. red) dot represents real (resp. fake) reviews labeled by ELSIEDET. Each blue (resp. red) triangle represents existing (resp. filtered) reviews according to Dianping. Through these three examples, first, we can analyze the evasive strategy taken by elite Sybil users. Elite Sybil users will post massive reviews to camouflage fake reviews and this strategy can evade most aggregate behavioral-based clustering approaches that rely on computing similarity of user activity. Second, these three users appear in the same community. They write fake reviews in similar stores and share with the similar behavioral patterns in the way to post reviews. However, there is one distinct difference from Dianping filtering system. For a given user, in Figure 15(a), we can see that no reviews have been filtered; in Figure 15(b), partial reviews have been filtered; in Figure 15(c), Sybil users are extremely sensitive to Dianping filtering system as all reviews have been filtered. This is perhaps because of his/her long negative-credit history.

In summary, we feel that Dianping filtering system is largely working on detecting regular Sybils, as shown in Figure 15(b). We feel that Dianping is being fairly opaque about its filtering system as most of the real reviews of an elite Sybil user have also been falsely filtered due to its high false alarm rate, as shown in Figure 15(c). Although our dataset is moderate in size compared with the Dianping database, it is large enough to allow us to gain meaningful insights and identifying factors that impact the results and limitations of conventional Sybil detection systems.

VII. DISCUSSION AND LIMITATIONS

In this section, we discuss the potential application of ELSIEDET and the limitations of the paper.

A. Application of ELSIEDET

We here show how ELSIEDET can be integrated to the existing Dianping's Sybil system to enhance its tolerance of elite Sybil attacks.

Mitigating Sybil attacks by changing the weight of reviews with respect to Sybilness. The ultimate goal of a Sybil campaign is to manipulate the ratings of stores by generating massive fake reviews and ratings. To mitigate the negative impact of Sybil attacks on stores' ranking, a potential approach is to tune the weights of reviews of the suspicious users according to their *Sybilness*. By assigning a lower weight to a highly suspicious user, it will significantly increase the difficulty of the Sybil organizations to manipulate the ratings and help alleviate the human labor required to verify massive number of users reported.

Monitoring the top elite Sybil users to predict Sybil campaigns. Detecting Sybil campaigns is critical for Dianping to limit the impact of Sybil attacks. In Section VI-D, we have pointed out that we can monitor elite Sybil users and exploit their group actions to identify the Sybil campaign in a real-time fashion. Note that, considering millions of stores and users, only monitoring a small set of suspicious users can significantly save the efforts and resources of the social network operators.

B. Limitations

First, although our detection system has strictly focused on Dianping, our results are applicable to a wider range of URSNs or any social media that relies on user-contributed comments. Examples include E-commerce (Amazon, Ebay, BizRate), movie-rating platforms (IMDB, Netflix, Douban), traveling services (TripAdvisor), and multi-agent systems (Advogato). In specific, in 2012, Yelp profile pages featured "consumer alerts" on several sneaky businesses which got caught red-handed trying to buy reviews, crafted by Yelp "elite" users, for these businesses [36]. TripAdvisor has also put up similar warning notices. These examples may have specific detection systems, and we leave their design and evaluation to future work. Second, we acknowledge if a Sybil community can minimize the involvement in multiple campaigns, it would be very likely to boost the chance to evade the detection; however, recruiting high-cost elite Sybil users to participate in limited Sybil campaigns contradicts the economic basis. Third, we do not study the relationships among reviewers on Dianping. For example, a reviewer can make friends and keep a friend list on Dianping. A reviewer can send a flower to another reviewer in order to present a sense of complement to the reviewer who posts a nice review. We think these social links among reviewers are weak, extraneous for characterizing elite Sybil users on Dianping. Instead, we exploit user-community as a zoom lens to take a particular micro-macro analysis of elite Sybil users without using any user profile information.

VIII. RELATED WORK

In this section, we survey the methodology used in previous research from four categories: graph-based approaches, feature-based approaches, aggregate behavioral-based clustering approaches, and crowdsourcing-based approaches. We review each of these approaches as follows.

Graph-based approaches. Graph-based detection views accounts as nodes and social links between accounts as edges. For example, Liu *et al.* [26] considered the dynamic change in the social graph. Much prior work [10, 18, 28] holds the assumption that in a social graph, there exist a limited number of attack edges connecting between benign and Sybil users. The key insights behind this is that it becomes difficult for

attackers to set up links to real users, and strong trusts are lacking in real OSNs, such as RenRen [47] and Facebook [5, 8, 12, 22]. Souche [46] and Anti-Reconnaissance [31] also rely on the assumption that social network structure alone separates real users from Sybil users. Unfortunately, this was proven unrealistic since real users refuse to interact with unknown accounts [37]. Recent research [7] relaxes these assumptions and takes a combined approach that first leverages victim prediction to weigh the graph and upper bound the aggregate weight on attack edges; then it performs a short random walk on the weighted graph and distributes manually-set scores to classify users. However, We argue that these methods do not hold on URSNs and the nodes in URSNs do not show a tight connectivity as those in general OSNs, which renders the social network graph-connectivity-based Sybil detection approaches less effective in URSNs.

Feature-based approaches. The advantage of behavioral patterns is that these can be easily encoded in features and adopted with machine learning techniques to learn the signature of user profiles and user-level activities. Different classes of features are commonly employed to capture orthogonal dimensions of users' behaviors [13, 24, 32, 34, 35, 43]. Other work [33, 38, 39] considers the associated content information, such as reviews context, wall posts, hashtags, and URLs, to filter Sybil users. Specifically, the Facebook immune system [35] detects Sybil users based on features characterized from user profiles and activities. COMPA [13] is designed to uncover compromised accounts via sudden change alerts according to the behavioral patterns of users. In addition to user profile, Song *et al.* [34] proposed a target-based detection on Twitter approach which bases on features of retweets. However, feature-based approaches are relatively easy to circumvent by adversarial attacks [4, 9, 42, 51]. Further work will also be needed to detect sophisticated strategies exhibiting a mixture of realistic and Sybil users features.

Aggregate behavioral-based clustering approaches. Recently, rather than classifying single users, much work [3, 11, 16, 29, 40, 43] focuses on detecting clusters of users. Specifically, CopyCatch [3] and SynchroTrap [11], implementing mixed approaches, score comparatively low false positive rates with respect to single feature-based approaches. For Dianping, the elite Sybil users, however, write elaborate reviews by mimicking the real reviews and intentionally manipulate the review temporal patterns within a Sybil campaign, so as to change the behavior features to bypass detection.

Crowdsourcing-based approaches. Wang *et al.* [44] tested the efficacy of crowdsourcing (such as leveraging humans, both expert annotators, and workers hired online), at detecting Sybil accounts simply from user profiles. The authors observed that the detection rate for hired workers drops off over time, although majority voting can compensate for the loss. However, two drawbacks undermine the feasibility of this approach: (i) This solution might not be cost effective for large-scale networks, such as Facebook and Dianping; (ii) exposing personal information to external workers raises privacy issue [14]. We observe that some recent work discusses how to identify the regular Sybil users in URSNs (*e.g.*, Yelp and Dianping) by exploiting crowdsourcing-based approaches [23, 32, 34], or model-based detection [25] that limits their broad applicability. Most recent work leverages Recurrent Neural Networks (RNNs) to automate the generation of synthetic Yelp reviews [48]. However, we emphasize that ELSIEDET is immune to the AI attack for two reasons: (i) ELSIEDET does not accommodate any contextual features that RNN-based attack is centered around. (ii) The attack dataset used in [48] does not take in any human-crafted fake reviews, which presumes that the proposed defense [48] cannot well identify the fake reviews written by elite Sybil users defined in our paper. We believe that our research is the first to define, characterize, and perform a large-scale empirical measurement study toward the elite Sybil attack in URSNs. We thus hope that our results may serve as a supplement to other traditional Sybil detection schemes and shed light on the novel Sybil detection system for uncovering other evolved Sybil users.

IX. CONCLUSION

This paper illuminates the threat of large-scale Sybil activities in User-Review Social Networks. We first demonstrated that Sybil organizations of Dianping utilize a hybrid cascading hierarchy to orchestrate campaigns. An in-depth analysis of elite Sybil users leads us to several important conclusions: elite Sybil users are more spread out temporally, craft better-edited contents, but have fewer reviews filtered. We showed that most Sybil campaigns can be determined within the first two weeks by only monitoring detected elite Sybil users. Strikingly, we also showed that a series of chains leverage Sybil organizations to distort the online rating, rendering previous research outdated. We emphasize that sophisticated manipulation of temporal patterns is key to orchestrating the evasive strategy. Finally, we demonstrated that ELSIEDET is both highly effective and scalable as a standalone system.

Although our study and experiments focus on Dianping, we believe that the anti-Sybil defense as examined in this paper provides an opportunity for all URSNs to stop the spread of elite Sybil users in a way that has never been visible on Dianping or other social networks like it.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of China, under Grants 71671114, 61672350, and U1405251. Corresponding author: Haojin Zhu.

REFERENCES

[1] (2017) Scikit-learn. [Online]. Available: http://scikit-learn.org/
[2] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," National Bureau of Economic Research, Tech. Rep., 2017.
[3] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "Copycatch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Networks," in *Proc. WWW*. ACM, 2013, pp. 119–130.
[4] B. Biggio, G. Fumera, and F. Roli, "Security Evaluation of Pattern Classifiers under Attack," *IEEE TKDE*, vol. 26, no. 4, pp. 984–996, 2014.
[5] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All Your Contacts are Belong to Us: Automated Identity Theft Attacks on Social Networks," in *Proc. WWW*. ACM, 2009, pp. 551–560.
[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.
[7] Y. Boshmaf, D. Logothetis, G. Siganos, J. Lería, J. Lorenzo, M. Ripeanu, and K. Beznosov, "Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs," in *NDSS*, vol. 15, 2015, pp. 8–11.

[8] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The Socialbot Network: When Bots Socialize for Fame and Money," in *Proceedings of the 27th ACSAC*. ACM, 2011, pp. 93–102.

[9] M. Brückner, C. Kanzow, and T. Scheffer, "Static Prediction Games for Adversarial Learning Problems," *JMLR*, vol. 13, no. Sep, pp. 2617–2654, 2012.

[10] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the Detection of Fake Accounts in Large Scale Social Online Services," in *NSDI 12*, 2012, pp. 197–210.

[11] Q. Cao, X. Yang, J. Yu, and C. Palow, "Uncovering Large Groups of Active Malicious Accounts in Online Social Networks," in *Proc. CCS*. ACM, 2014, pp. 477–488.

[12] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq, "Paying for Likes?: Understanding Facebook Like Fraud Using Honeypots," in *Proc. IMC*. ACM, 2014, pp. 129–136.

[13] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting Compromised Accounts on Social Networks," in *NDSS*, 2013.

[14] Y. Elovici, M. Fire, A. Herzberg, and H. Shulman, "Ethical Considerations When Employing Fake Identities in Online Social Networks for Research," *Science and Engineering Ethics*, vol. 20, no. 4, pp. 1027–1043, 2014.

[15] C. Erdman, J. W. Emerson *et al.*, "BCP: An R Package for Performing a Bayesian Analysis of Change Point Problems," *Journal of Statistical Software*, vol. 23, no. 3, pp. 1–13, 2007.

[16] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and Characterizing Social Spam Campaigns," in *Proc. IMC*. ACM, 2010, pp. 35–47.

[17] S. A. Golder and M. W. Macy, "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures," *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.

[18] N. Z. Gong, M. Frank, and P. Mittal, "Sybilbelief: A Semi-supervised Learning Approach for Structure-based Sybil Detection," *IEEE TIFS*, vol. 9, no. 6, pp. 976–987, 2014.

[19] A. Gupta, H. Lamba, and P. Kumaraguru, "$1.00 per rt# bostonmarathon# prayforboston: Analyzing Fake Content on Twitter," in *eCRS, 2013*. IEEE, 2013, pp. 1–12.

[20] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges," *IEEE Internet Computing*, vol. 11, no. 6, 2007.

[21] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social Spammer Detection in Microblogging," in *Proc. IJCAI*. AAAI Press, 2013, pp. 2633–2639.

[22] M. Ikram, L. Onwuzurike, S. Farooqi, E. De Cristofaro, A. Friedman, G. Jourjon, D. Kaafar, and M. Z. Shafiq, "Measuring, Characterizing, and Detecting Facebook Like Farms," *TOPS*, 2017.

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

[24] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns," in *ICWSM*, 2015.

[25] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, "Bimodal Distribution and Co-Bursting in Review Spam Detection," in *Proc. WWW*. ACM, 2017.

[26] C. Liu, P. Gao, M. Wright, and P. Mittal, "Exploiting Temporal Dynamics in Sybil Defenses," in *Proc. CCS*. ACM, 2015, pp. 805–816.

[27] M. Luca and G. Zervas, "Fake it Till You Make it: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 2016.

[28] A. Mohaisen, A. Yun, and Y. Kim, "Measuring the Mixing Time of Social Graphs," in *Proc. IMC*. ACM, 2010, pp. 383–389.

[29] A. Mukherjee, B. Liu, and N. Glance, "Spotting Fake Reviewer Groups in Consumer Reviews," in *Proc. WWW*. ACM, 2012, pp. 191–200.

[30] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp Fake Review Filter Might be Doing?" in *ICWSM*, 2013.

[31] A. Paradise, R. Puzis, and A. Shabtai, "Anti-Reconnaissance Tools: Detecting Targeted Socialbots," *IEEE Internet Computing*, vol. 18, no. 5, pp. 11–19, 2014.

[32] M. Rahman, B. Carbunar, J. Ballesteros, G. Burri, D. Horng *et al.*, "Turning the Tide: Curbing Deceptive Yelp Behaviors," in *SDM*. SIAM, 2014, pp. 244–252.

[33] A. Ramachandran, N. Feamster, and S. Vempala, "Filtering Spam with Behavioral Blacklisting," in *Proc. CCS*. ACM, 2007, pp. 342–351.

[34] J. Song, S. Lee, and J. Kim, "Crowdtarget: Target-based Detection of Crowdturfing in Online Social Networks," in *Proc. CCS*. ACM, 2015, pp. 793–804.

[35] T. Stein, E. Chen, and K. Mangla, "Facebook Immune System," in *Proc. SNS*. ACM, 2011, p. 8.

[36] D. Streitfeld, "Buy Reviews on Yelp, Get Black Mark, The New York Times," 2012. [Online]. Available: http://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html

[37] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting Spammers on Social Networks," in *Proc. ACSAC*. ACM, 2010, pp. 1–9.

[38] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and Evaluation of a Real-time URL Spam Filtering Service," in *IEEE S&P*. IEEE, 2011, pp. 447–462.

[39] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended Accounts in Retrospect: An analysis of Twitter Spam," in *Proc. IMC*. ACM, 2011, pp. 243–258.

[40] K. Thomas, F. Li, C. Grier, and V. Paxson, "Consequences of Connectivity: Characterizing Account Hijacking on Twitter," in *Proc. CCS*. ACM, 2014, pp. 489–500.

[41] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An Analysis of Social Network-based Sybil Defenses," *ACM SIGCOMM CCR*, vol. 40, no. 4, pp. 363–374, 2010.

[42] F. Wang, W. Liu, and S. Chawla, "On Sparse Feature Attacks in Adversarial Learning," in *ICDM*. IEEE, 2014, pp. 1013–1018.

[43] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are How You Click: Clickstream Analysis for Sybil Detection," in *USENIX Security*, 2013, pp. 241–256.

[44] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao, "Social Turing Tests: Crowdsourcing Sybil Detection," in *NDSS*, 2013.

[45] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Serf and turf: Crowdturfing for Fun and Profit," in *Proc. WWW*. ACM, 2012, pp. 679–688.

[46] Y. Xie, F. Yu, Q. Ke, M. Abadi, E. Gillum, K. Vitaldevaria, J. Walter, J. Huang, and Z. M. Mao, "Innocent by Association: Early Recognition of Legitimate Users," in *Proc. CCS*. ACM, 2012, pp. 353–364.

[47] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering Social Network Sybils in the Wild," *ACM TKDD*, vol. 8, no. 1, p. 2, 2014.

[48] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated Crowdturfing Attacks and Defenses in Online Review Systems," in *Proc. CCS*. ACM, 2017.

[49] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "Sybillimit: A Near-optimal Social Network Defense against Sybil Attacks," in *IEEE S&P*. IEEE, 2008, pp. 3–17.

[50] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: Defending against Sybil Attacks via Social Networks," in *ACM SIGCOMM CCR*, vol. 36. ACM, 2006, pp. 267–278.

[51] F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung, and F. Roli, "Adversarial Feature Selection against Evasion Attacks," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 766–777, 2016.

[52] Y. Zhao, Y. Xie, F. Yu, Q. Ke, Y. Yu, Y. Chen, and E. Gillum, "BotGraph: Large Scale Spamming Botnet Detection," in *NSDI*, vol. 9, 2009, pp. 321–334.