Event-Based Social Network Discovery (ESONED) Using WiFi Access Points

Le Fang, Tong Guan, Wen Dong, and Chunming Qiao

Computer Science and Engineering
State University of New York at Buffalo
338 Davis Hall, Buffalo, New York 14260, USA
{lefang,tongguan,wendong}@buffalo.edu,qiao@computer.org

Abstract. Social network of a community is essential in lots of social research and applications. In this paper, we propose a novel framework to capture people's interactions and discover their social network. More specifically, our event-based social network discovery algorithm (ESONED) extracts social events from activity logs of WiFi access points (AP), and we infer social relationships between people from analyzing the common events they appear together. We find that information such as where, when, how long and how frequent two people appear together implies the nature of their relationship. Our approach is accurate and efficient in discovering a large-scale social network, as demonstrated by experiments on a real-world dataset collecting from 1,600 WiFi APs, covering 32,000 people for one month.

Keywords: ESONED, Social Network Discovery, WiFi Access Point, Social Event, Co-appearance

1 Introduction

Social network discovery is essential in a wide range of research, such as modeling the propagation of epidemics [7, 2], understanding the influence of social network on individuals' behavior [2, 4]. The conventional way to accomplish this task is through surveys. However, such survey data usually consumes lots of resources, and can not capture details of social interactions. As a result, the development of an efficient method is still of great interest.

In the last decade, many researchers have exploited various sensors to capture people's interactions. For example, some sensors are wearable devices that detect mutual proximity distributively through active radio frequency identification (RFID) [1]. Another type of sensors getting popular is based on mobile phones [5], as people habitually carry their phones and use them frequently. However, these approaches rely on extra equipments or software, and require users' consent on participation [1, 6]. Consequently, the collected data only covers a small population but consumes a lot of human labor. When discovering large-scale social network, these approaches are expensive and impractical.

In this paper, we propose an event-based social network discovery algorithm (ESONED) to discover dynamic social network by identifying and analyzing social events from activity logs of WiFi AP. In contrast to existing approaches to perform sensing on participants, our framework creatively uses WiFi APs as "sensors" and exploits the activity logs at AP side. We realize that nowadays smart-phones periodically broadcast packets containing unique Media Access Control (MAC) addresses, to actively scan for network services in the environment. Nearby WiFi APs can receive these packages and record where and when the smart-phones appear. Thus the activity log on WiFi AP serves as a perfect resource to indicate people's mobility trace and their co-appearances patterns. Researchers have already verified that connections exist between people's social ties and their spatial-temporal co-appearance [3]. Based on this idea, our approach systematically identifies co-appearance patterns as social events from APs' activity logs and infers relationships between people from analyzing the common events they appear together. The contributions of this paper are as follows: 1) we develop ESONED to extract social events from activity logs of WiFi APs. 2) we infer dynamic social network of individuals from analyzing these social events. One worked example is that we differentiate different social communities by setting thresholds on total amount of time people sharing in common events. 3) we validate our algorithm on a real-world large-scale dataset.

The paper proceeds as follows. In Section 2, we introduce our dataset and the ESONED. In Section 3, we present some experiment results and validate our algorithm. Finally in Section 4, we conclude this paper.

2 Our Dataset and Algorithm

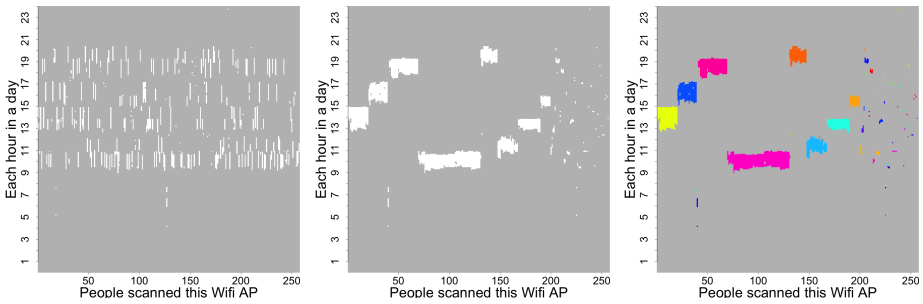
2.1 Our Dataset

Our dataset contains the complete activity logs of WiFi APs in a typical North American university from mid-March to mid-April, 2015. More specifically, each activity log records the basic service set identification (BSSID), the MAC address of the mobile device which broadcasts packages, the start time and end time for WiFi AP to receive packages from mobile phones, and the signal strength etc. Complete anonymization is applied to identifiable information such as user MAC address to protect people's privacy. The dataset totally contains over 5,400,000 records, involving 1,600 WiFi APs and 32,000 mobile phone devices. We assume that there is a one-to-one correspondence between these mobile phones and people in the campus. Our dataset implies people's location and mobility trace in a room-level spatial resolution.

2.2 The ESONED

The ESONED focuses on identifying and analyzing social events. Generally, events can be characterized by the co-appearance of participants—where and when a bunch of people accumulate and disperse. Therefore, we define *event size* θ_s as the number of people in a social event, and *event duration* θ_d as the amount

of time from the start of the event to its end. For example, a typical event at university is a class. People attending the same class tend to enter the classroom around the start time of class and to leave at its end. Furthermore, events are not necessarily purposely scheduled and involving a large amount of people—an event can be that two friends unintentionally encounter one another and chat for a while. Our algorithm works in several steps as demonstrated below.



(a) Feature vectors (b) Hierarchical clustering (c) CCL

Fig. 1: Extracting social events from co-appearance

Step 1: Feature Vector. As WiFi AP receives packages from mobile phones within several meters, spatial proximity of smart-phone owners is automatically ensured if we analyze one WiFi AP each time. Mathematically, for a WiFi AP W and phone device D, we can represent D's scanning activities of W as a "feature vector" $\boldsymbol{V}$: each element V_i is a binary number which represents the activity during the i-th minute. As Fig. 1a, we aggregate people's feature vectors into columns of a 2D heat image. The vertical axis represents the timeline and each row in the matrix is a snapshot representing devices scanning W during that minute. Yet we cannot get much information from these chaotic patterns.

Step 2: Identification and Extraction of Social Events. In order to identify the temporal coincidence of people around the same AP, we use distance measures (such as Euclidean distance) to evaluate similarity between their feature vectors. Intuitively, smaller distance indicates higher synchronicity. We then apply the hierarchical clustering algorithm on these feature vectors: feature vectors within higher similarity are grouped together. Then a 2D "clustering matrix" is formed as shown in Fig. 1b by reordering columns in terms of their hierarchical structure. Due to temporal coincidence of people in same event, certain blocks will be formed in the clustering matrix. In this way, each block is a candidate of social event. The width of the block is the number of devices in this event, i.e. θ_s, and the height gives the duration of the event, namely θ_d.

However, it is impractical and inefficient to manually extract these characteristics, θ_s and θ_d. Accordingly, we employ the so-called connected component labeling (CCL) algorithm to differentiate each block. CCL works to label subsets of connected components distinctively based on a given heuristic (Fig. 1c). Of course, not all blocks should be recognized as a social event. Feature vectors of people who coincidentally pass the same WiFi AP could form a block but there is no meaningful social interactions behind. Therefore to identify meaningful events, we must set specific thresholds and criteria for θ_s, θ_d, etc accordingly.

Step 3: Inference of Social Events. Finally, we infer people's social ties by analyzing the events they appear together. We believe that the nature of people's relationship can be characterized by when, where, how long and how frequent they interact. The inference is always based on statistics of multiple events, since single event is not plausible due to randomness. For instance, people who share a significant amount of time in common places and appear in multiple places can be reasonably interpreted as having a social connection.

3 Findings and Validations.

Direct validation on extracted social events. In order to validate our framework, we extract over 35670 events involving 28,678 people from our dataset. Fig. 2 shows the number of events versus event size θ_s in a logarithmic scale, we observe that those events matches "Power Law" distribution. In addition, we acquire the ground truth data such as the class schedule, location and number of students enrolled in each class. And we find over 80% of courses have been captured from extracted social events.

To further validate these social events are meaningful, we consider both people in an event and people sampled randomly, and compare the total time they are found at other events. We sample 1000 comparative groups and the results are shown in a QQ plot 3. The values, i.e 0.1, 0.99 etc, mark respective quantiles in the two distributions. Fig. 3 shows that people from an event are remarkably more possible to be found together in other events than people randomly sampled. We know that people at each event co-appear at same time, and Fig. 3 further proves that people from the same event are not coincidentally gathered—which indicates the extracted events are meaningful to imply social relationships.

Validation and findings based on a real-world survey. To get ground truth of people's relationships, we conduct a simple survey over ten PhD students and let them name a bunch of people as their "friends", "classmates" and "lab members" during April 2015. For each people dyad, we obtain statistics of the common events they appear together, such as total duration (ST) and number of distinct WiFi APs (SN). As shown in Fig. 4c, we can find the nature of relationship is well differentiated by these statistics. Specifically, friends may interact less time durations but appear in most diverse places; classmates meet longer but rare places besides classroom; lab members stay together with longest

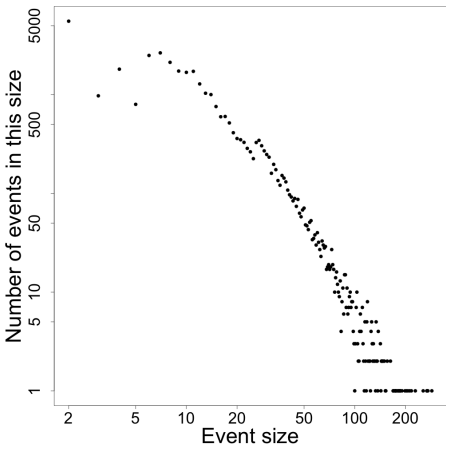


Fig. 2: Distribution of event size θ_s

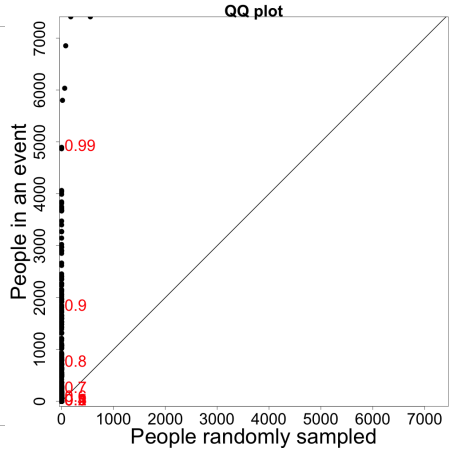


Fig. 3: QQ plot between people inside an event and people randomly sampled

time. Furthermore, friends are found together at very diverse places, such as lecture hall, restaurant, gym and even resident hall, but the total time differs greatly. The experiments which combine survey results and extracted events suggests ST as a good indicator to identify "lab members" and SN to "friend".

As a worked example, we utilize ST to differentiate social communities as following: for a set of WiFi APs in a comprehensive building, we set threshold on ST and visualize the consequent social network as a weighted graph in Fig. 4a. For instance, when threshold is 300 minutes during 2 weeks, which means only edges with "$ST >= 300''$ are interpreted as social ties (left of Fig. 4a)[1]. We observe 900 people in the graph and multiple communities shown with different colors. The sizes of communities vary from 2 to about 40, in which people are respectively inferred as friends or those attending same classes. When threshold is 800 minutes (right of Fig. 4a), totally 232 people are observed. Communities are relatively independent than that under 300 minutes threshold, which can be interpreted as people in same research labs or offices. People with "$T >= 800''$ are more likely to be professors, senior graduate students and staffs who used to stay there for long time. These populations agree with the real population very well. In this way, we demonstrate that people in different roles, i.e "students" or "researchers", could be differentiated by different thresholds on ST. To further demonstrate this, we set multiple ST intervals, discover social communities and plot the average community size (ACS) versus ST intervals(Fig. 4b). We can see ACS starts from lower values as red bars, reaches highest values as yellow bars and deceases as green bars afterwards. For red bars, people are usually friends or small study groups, who don't stay long time there; for yellow bars, more people are those attending same classes; for green bars, people are members in same

[1] For displaying nodes and communities clearly, only partial edges are drawn in 4a.

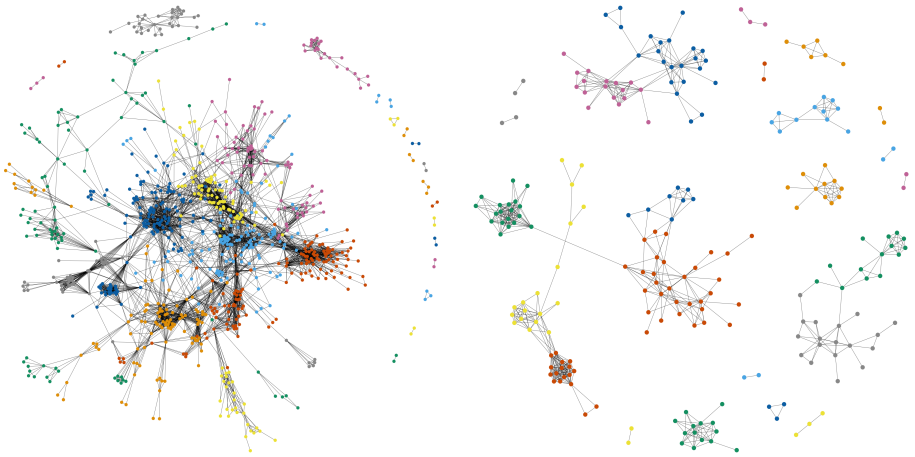
laboratory or colleagues in same office, who are more closely connected and form smaller communities.

4 Conclusion

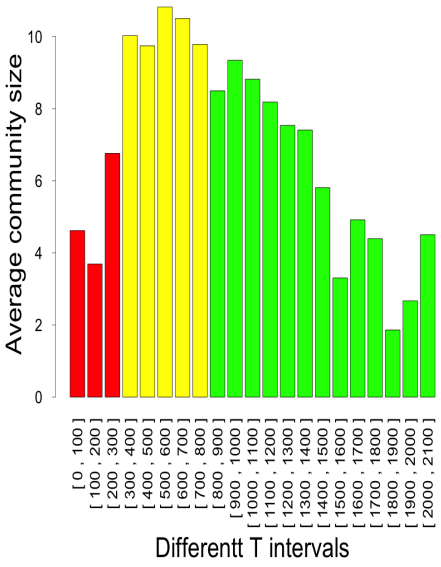
Based on our experiments, we have shown that our framework can effectively extract social events and social interactions. In addition, our framework is cost efficient compared with existing approaches, since distributed WiFi APs can passively receive probe packages from mobile phones.

References

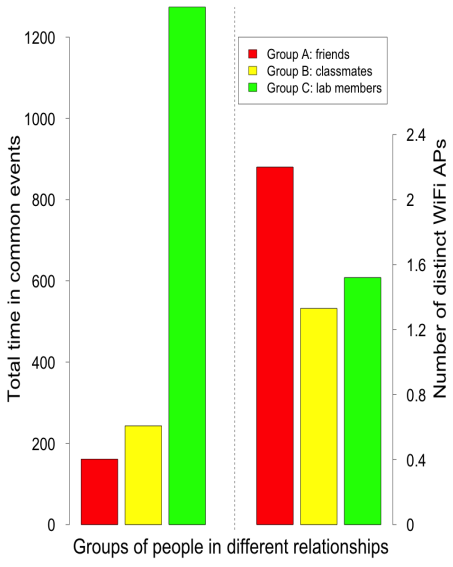
1. Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.
2. Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.
3. David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.
4. Wen Dong, Bruno Lepri, and Alex Sandy Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 134–143. ACM, 2011.
5. Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.
6. Daniel Olguın Olguın and Alex Sandy Pentland. Sociometric badges: State of the art and future applications. 2007.
7. Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.



(a) Social communities under two different thresholds. *Left:* threshold = 300 minutes, *Right:* threshold = 800 minutes



(b) Average community sizes under different ST intervals



(c) Statistics of people taking survey

Fig. 4: Interpret social ties and social communities