Poster: Understanding Event Attendance through Analysis of Human Crowd Behavior in Social Networks

Stavroula Karanikolaou
Department of Informatics
Athens University of Economics and Business,
Athens, Greece
karanikolaou@aueb.gr

Ioannis Boutsis
Department of Informatics
Athens University of Economics and Business,
Athens, Greece
mpoutsis@aueb.gr

Vana Kalogeraki
Department of Informatics
Athens University of Economics and Business,
Athens, Greece
vana@aueb.gr

ABSTRACT
Understanding human crowd mobility has found important applications in several commercial domains such as marketing, recommendation systems and resource planning. In this paper we investigate users' social activities and interactions developed in "human-centered participatory sensing" groups and perform an analysis to understand human crowd behavior. We exploit two popular real and heterogeneous datasets for our analysis: an event based social network (Meetup) dataset and a checkin-based geosocial network (Foursquare) dataset, to understand user attendance in social community events and provide insights into the factors that influence users to attend events.

Categories and Subject Descriptors
C.2.4 [**Distributed Systems**]

Keywords
Distributed Systems, Event Attendance, Social Networks, Human Crowds

1. INTRODUCTION
Over the recent years, the ubiquitous sensing capabilities, the prevalence of social networks and the widespread adoption of smartphones, are driving the development and adoption of applications and services that are changing the way we interact with the world and each other. We now enter an era where people actively participate in sharing aspects of their lives online, creating virtual social communities, turning people into producers of "personal data", in what is termed as "human-centered" or "participatory sensing" systems.

For a concrete example of this trend, consider the newly emerged social networks such as Twitter[3], Foursquare[1], Meetup[2], etc.; these have evolved into attractive platforms for exchanging ideas, initiating discussions, sharing experiences, or publishing group activities, from informal gatherings (such as movie or sport nights) to organizing formal activities (*e.g.*, business meetings). They have the ability to aggregate information, opinions and activities of diverse groups of people at relatively low cost. For instance, there are more than 40M users in Foursquare as of January 2014 in which users post their location at the venues they attend, while in the Meetup event-based social network, which has attracted over 15.86M users, users are able to publish social group meetings (events).

One important aspect of such services is that they can provide interesting insights into how members of the human crowd relate to each other in terms of the social events they attend, and also for identifying the factors that influence user behavior. Such information may be useful in several application areas, for instance, for commercial purposes (*i.e.,* advertising leaders of the field that attend a business event has a higher chance of attracting a larger RSVP list at the event), for making personalized recommendations (based on users' preferences for recommending similar groups and events or for suggesting "people the user might know", when they attend similar events), and for achieving better resource planning in urban environments (*i.e.,* when scheduling multiple and potential concurrent events and facilitate early attendance decisions).

At the same time, the large growth of location aware smartphone devices with ever-growing sensing, computation and communication capabilities, has enabled users to connect to their social networks from their smartphones and provide additional metadata to the social network posts, such as their current location. These metadata, including their spatiotemporal information, can be used to estimate whether the user has attended an event even if he/she has not responded to the group's invitation; this is achieved when the user shares social content with spatiotemporal information that matches with the time and place of an event (*i.e.,* in Foursquare, users have the ability to post their location at venues they attend, in Twitter and Facebook social posts can include geo-located metadata of the user's location).

In this paper, we study the human crowd behavior in social networks and provide insights on the factors that influence users to attend events. The questions that we aim to answer are: (1) what does the participation of users in a geosocial network reveal about user involvement in community events? (2) can we exploit the links between differ-

DEBS'14, May 26-29, 2014, MUMBAI, India.

http://dx.doi.org/10.1145/2611286.2611324.

ent social networks to provide insights on the factors that influence a user to attend an event? We exploit two heterogeneous datasets for our evaluation: an event based social network dataset (Meetup) and a location-based social dataset (Foursquare) and we provide a study to understand the form of these datasets and the behavior of the human members in the social networks.

2. SYSTEM MODEL

We assume a social network where users can create social events. We denote as an event $e \in E$, a pre-organized gathering of people that takes place in a certain location during a particular interval of time and the people participate in their own way. An event can be a music festival, a lecture at the university, a motorbike competition and so on. In our setting each event e has the form of $< id_e, lat_e, lon_e, time_e, dur_e >$, where id_e is the identifier for the event, lat_e and lon_e reflect the location of the event, $time_e$ represents when the event will initiate, and dur_e represents the duration of the event.

Each user $n \in U$ in our system can be a member of a social network such as Foursquare, Twitter, etc. Users are able to inform these networks about their willingness for attending a social event or that they currently participate at the event. Whenever a user n provides feedback about attending a social event $e \in E$ we represent this user attendance with a tuple $< id_e, id_n >$, which encapsulates the user and the event's identifiers. The user attendance to an event can be either formally defined, meaning that the user responds to the event invitation and informs the respective social network for his attendance at event id_e, or informally defined when the user informs a social network about his presence on a location that coincides with the place and time of the event (lat_n, lon_n, $time_n$). This information about the event attendance is stored in user profiles. In this paper we consider only informally defined attendances since determining the user attendance from the user response is trivial.

3. REPRESENTATIVE NETWORKS

We have collected data from two popular social networks, Meetup and Foursquare. There are several reasons for choosing these datasets: First, Foursquare is one of the most popular location-based social networking websites for mobile devices and Meetup is one of the most popular event-based social networks. Second, most of the information shared in these networks is public, this allows us to crawl a large fraction of the network. Additionally, both networks provide an API to provide access to the data. In this section we introduce these social networks and discuss the datasets we used.

3.1 Foursquare

Foursquare is the leading location-based social networking website for mobile devices, such as smartphones. Foursquare allows registered users to post their location at a venue ("check-in") by selecting the location from a list of venues the application locates nearby. Simultaneously, users can also choose to post their check-ins to other social networks such as Facebook or Twitter. Moreover, users are encountered to be very specific with their check-ins by indicating their precise location or activity while at a venue. In this way, Foursquare collects important information from the user profiles and then supplies them with personalized recommendations and business deals. Every time users check into a place, Foursquare provides awards as incentives for checking in at locations with certain tags, for checking-in frequently at the same venue, or for other patterns such as the time of check-in. The users who check-in the most often to a venue become the "mayors" and users regularly vie for "mayorships". As of January 2014, there have been more than 5 billion check-ins with Foursquare from over 45 million people worldwide [1].

Foursquare API provides information for the users including *user id, firstname, lastname, friends, homecity, gender, checkins, etc.* For each checkin it provides: *id, timestamp, privacy level, user, venue, location, etc.* Our dataset from the Foursquare API contains 2073740 check-ins from 18107 users ranging from March 2010 to January 2011 [6]. For each user it provides the location and time of his check-ins and his friends in Foursquare.

3.2 Meetup

Meetup is the world's largest online social network of local groups. It facilitates group meetings in various localities around the world. Meetup allows members to join Meetup groups that derive from several categories such as politics, books, careers or hobbies. As of January 2014, the company claimed to have 15.86M members in 196 countries and 141,137 meetup groups [2], although these figures may include inactive members and groups. More than 9,000 groups get together in local communities every day. Each of these groups are associated with: *group id, category id, country, city, group urlname, latitude, longitude, organizer id, topic*. Additionally, each event contains information like *event id, description, group id, rsvp, time, venue id, latitude,longitude, etc.* We have crawled Meetup through its API and we have extracted events that took place from 13 March 2010 until 31 July 2011. These data contain approximately 90K events, which have been announced from 2578 Meetup groups. Each group corresponds to a topical category depending on the themes. There are 22 different categories defined in the Meetup network (table 1), where we present the amount of groups in each categories as well as the amount of events announced from the groups that belong to the respective category.

4. DATA LINKING

One of the many advantages of event-based social networks such as Meetup, is the fact that all events are available to the general public. However, in networks like Foursquare, the attendance is implicit, and thus, we implemented a Web-based crawler in order to extract the user attendance from their social posts (check-ins). Each check-in in Foursquare is represented as: $< id_{pn}, lat_{pn}, lon_{pn}, time_{pn}, content_{pn} >$, where id_{pn} is the unique identifier for the check-ins, lat_{pn} and lon_{pn} denote the geographical location where the user resides, $time_{pn}$ denotes the actual time of the post and $content_{pn}$ is the content of the post, such as text or image.

We assume that a user has attended an event indicated by Meetup, when there is a Foursquare spatiotemporally close check-in with a specific event, since the user would be located at the place of the event when it occurs, and thus we represent the user attendance with a tuple $< id_e, id_{pn} >$. Note, however, that an implicit response can be more accurate, since it ensures that the user was at the place of the event, while an explicit response (RSVP) does not guarantee that the user will attend the event.

In order to define the links among events and the geo-located posts we bound their spatial distance based on the haversine formula and their temporal distance based on the euclidean distance as follows:

$$s\_dist(id_e, id_{pn}) = 2 * R* \arcsin(\sqrt{x_1 + \cos(lat_e)\cos(lat_{pn})x_2}) < B_1 \quad (1)$$

where: $x_1 = \sin^2(\frac{lat_{pn} - lat_e}{2}), x_2 = \sin^2(\frac{lon_{pn} - lon_e}{2})$

$$t\_dist(id_e, id_{pn}) = \sqrt{((time_e + \frac{dur_e}{2}) - time_{pn})^2} < B_2 \quad (2)$$

where B_1 and B_2 are the spatiotemporal bounds and R is earth's radius. Thus, if both distance constraints are fulfilled then we assume that the user has attended the event.

The first bound B_1 is more complex since nearby venues might exist and thus we need to ensure that the user resides at the place of the event. Thus, in order to define bound B_1 we retrieve the closest venue compared to lat_e, lon_e, where the event takes place, defined as $clos$, excluding the venue of the event. Then we set B_1, using equation (1) as:

$$B_1 = s\_dist(id_e, id_{clos})/2 \quad (3)$$

Retrieving the closest venue can be achieved by extracting the venues that belong to that area through the Foursquare API and performing the kNN algorithm [9]. This metric ensures that the user will not be related with a different venue than the venue that the event belongs to, but the user will be linked with all the events that take place concurrently at the same location.

The second bound B_2 depicts the time interval we desire to select to bound the event, and thus it depends on the type of the event, since each event has a different duration. In order to support long lasting events we use the duration of the event dur_e and thus we compute the time distance of the user's presence to the middle time point of the event. The dur_e is a configurable parameter that allows us to capture the fact that users do not always arrive at the beginning of the events.

5. EXPERIMENTS

In this section we provide a study on the Foursquare social network. In figure 1 we present the frequency of the top-K locations per user, which are linked to an event. These are locations checked-in from Foursquare users. We only present up to top-30 since the frequency of the following top-k locations is too small (less than 0.5%). As can be observed, a lot of checkins for each user are associated with a few places. For instance 28% check-ins for the Foursquare data are located in one place, which means that users visit and check-in at a few places frequently. Note though, that check-ins are mostly public places since users do not typically announce their location when they are at home or work.

Figure 2 presents the Cumulative Distribution Function (CDF) of the distance among the "neighborhood area of the users" and the locations they check-in. In order to extract the neighborhood area of each individual user we extract the centroid of the locations that the user has checked-in. As can be seen users typically visit events close by, since 42% of the users visit events within 100km and 82% of the users visit events within 500km, but only 10% of the users announce their visits to places within 10km. This is because users want to share their location with their social environment when they are further away from their home (*e.g.*, for a trip abroad). Also, we note that such distant check-ins distort the "neighborhood area of the user", especially when users only check-in to distant places.

Id	Cat. Description	Groups	Events
1	Arts & Culture	134	4317
2	Career & Business	139	3068
3	Cars & Motorcycle	113	5746
4	Community & Environment	142	4636
5	Dancing	135	5316
6	Education & Learning	117	1931
8	Fashion & Beauty	59	574
9	Fitness	131	6366
10	Food & Drink	161	6487
11	Games	132	3887
12	LGBT	43	1215
13	Movements & Politics	147	7378
14	Health & Wellbeing	138	4033
15	Hobbies & Crafts	110	2120
16	Language & Ethnic Identity	138	4749
17	Lifestyle	37	493
18	Literature & Writing	140	3407
20	Movies & Film	143	5512
21	Music	144	4925
22	New & Age & Spirituality	161	5486

Table 1: Event Categories

In figure 3 we present the amount of attendances for the events. The events were extracted from the Meetup social network (table 1) that shows the respective amount of events per category, the amount of events does not always reflect the size of the attendance. For instance, although category 2 (Career and Business) has only a few events (3068), it has a higher attendance compared to other groups with more events such as Arts & Culture (4317). Figure 4 shows the amount of events per user. As the figure shows users do not follow too many events, with the highest attendance from a user being 24 events. This is due to the amount of check-ins that could be linked with events.

6. RELATED WORK

Social networks have attracted a lot of attention in the literature recently. Authors in [8] use Meetup to define event-based social networks (EBSN) and study their unique features. of such networks including network properties, community structures and information flow and show how EBSNs differ from conventional social networks without considering their links with other networks. Noulas *et al* in [10] exploit the Foursquare and Cellular data to predict the user activities that take place in a specific area and the most common activity in the area, but they do not investigate the user attendance to the activities beforehand. Wakamiya *et al.*[11] and Fujisaka *et al.*[5] have used Twitter geo-tagged data and their content to study crowd mobility. However, they do not exploit the links among different social datasets.

Several event detection techniques [7] also exist in the literature, including our own work in [4] where we exploited

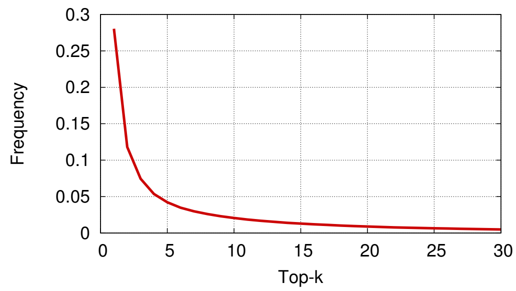

Figure 1: Frequency of Top-K locations

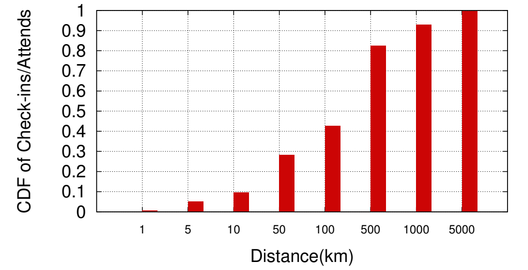

Figure 2: CDF of the Distance among Users and Check-ins

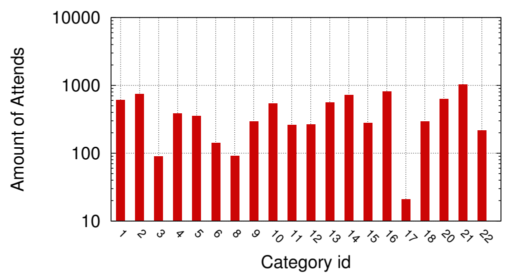

Figure 3: Amount of Users that attended events for each Category

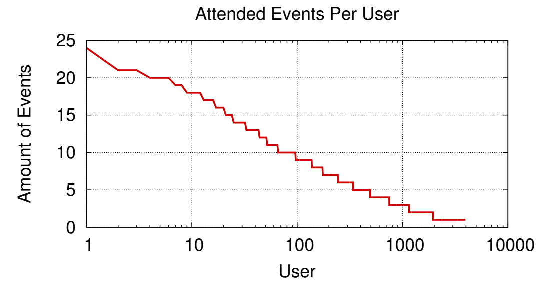

Figure 4: Amount of Events that each user has attended

GPS traces in order to identify real-world events efficiently using dynamic clustering and sampling techniques. These techniques can be used complementary to our proposed approach to link the events generated by these techniques to geo-located social data.

7. CONCLUSIONS

In this paper we perform a detailed analysis to understand human crowd behavior developed in "human-centered participatory sensing". We have observed that links among different social networks can be exploited to provide important insights for the individual user behavior. Our analysis show that (1) users typically visit frequently only a few places, (2) they announce their participation to events within 10-500km and (3) that the type of events plays an important role to the success of the event, in terms of participation. Finally we have shown that although users check-in frequently the amount of checkins that can be correlated with social events is small for each individual user.

Acknowledgments

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis - DISFER, Aristeia - MMD, Aristeia - INCEPTION Investing in knowledge society through the European Social Fund, the FP7 INSIGHT project and the ERC IDEAS NGHCS project.

8. REFERENCES

[1] Foursquare. https://foursquare.com/.
[2] Meetup. http://www.meetup.com/.
[3] Twitter. http://www.twitter.com/.
[4] I. Boutsis, V. Kalogeraki, and D. Gunopulos. Efficient event detection by exploiting crowds. In *DEBS*, pages 123–134, Arlington, Texas, USA, June 2013.
[5] T. Fujisaka, R. Lee, and K. Sumiya. Exploring urban characteristics using movement history of mass mobile microbloggers. In *HotMobile*, Annapolis, MD, February 2010.
[6] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *AAAI*, Toronto, Canada, July 2012.
[7] M. Keally, G. Zhou, and G. Xing. Watchdog: Confident event detection in heterogeneous sensor networks. In *RTAS*, pages 279–288, Stockholm, Sweden, April 2010.
[8] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: linking the online and offline social worlds. In *SIGKDD*, pages 1032–1040, Beijing, China, August 2012.
[9] S. A. Nene and S. K. Nayar. A simple algorithm for nearest neighbor search in high dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(9):989–1003, Sept. 1997.
[10] A. Noulas, C. Mascolo, and E. Frias-Martinez. Exploiting foursquare and cellular data to infer user activity in urban environments. In *MDM 2013*, Milan, Italy.
[11] S. Wakamiya, R. Lee, and K. Sumiya. Urban area characterization based on semantics of crowd activities in twitter. In *GeoS*. Brest, France, May 2011.