Inferring Significance of Meeting Groups in Human Contact Networks

Eiko Yoneki and Dan Greenfield

University of Cambridge Computer Laboratory
Cambridge CB3 0FD, United Kingdom
{firstname.lastname}@cl.cam.ac.uk

Abstract. The structure of human contact networks in the real world is time dependent and it is a complex task to describe its dynamics. This paper aims to identify dynamics of meeting groups in human connectivity traces, where meeting groups are expected to be a group interacting among the nodes in physical space. Thus, we define 'meeting group' differently from 'community'. We exploit statistical approach that provides quantitative attributes to uncover meeting groups. We identify the power law behavior of meetings that is important for supporting to understanding dynamics of information flow between meeting groups and building group oriented communication protocol.

1 Introduction

One of future visions of communications in the pervasive environment is called the Pocket Switched Network (PSN) [2]. A PSN provides intermittent communication based on physical proximity among dynamically connected mobile phones. We have demonstrated such communication paradigm in our previous work [5]. Efficient forwarding algorithms for such networks are emerging, mainly based on epidemic protocols where messages are simply flooded when a node encounters another node. Epidemic information diffusion is highly robust against disconnection, mobility and node failures, and it is simple, decentralized and fast. To reduce the overhead of epidemic routing, we have previously reported an approach that uses a logical connection topology, and that uncovers hidden stable network structures, such as social networks. In order to study social networks we have deployed a series of data collection and analysis from the human connectivity traces [8] [12]. We have shown improved performance by applying these extracted social contexts to a controlled epidemic strategy [7]. During this work, we have realized that further understanding of network models is essential, because the properties of human contact networks – such as community and weight of interactions – are important aspects of epidemic spread.

In our previous work, we have exploited community detection from the human connectivity traces by constructing weighted networks using characteristics of pair connections such as the duration of contact time and frequency of contacts also exploited spectral properties of the graph as well as Laplacian matrix [11]. Mostly, the approach we took is based on empirical and heuristic and the focus is finding a single aggregated logical network structure called 'community'. For dynamic graph mining, Berger-Wolf

et. al. show the study of community evolution based on node overlapping [1]. The evolution of subgraphs over time in biological networks has been discussed, however, these studies are based on static network setting and fairly small scale. The traces described in the next section are more complex with thousands of updates per day.

We define 'meeting group' differently from 'community'. Actual meeting among the member of the community may occur at certain time or location for possibly predictable duration. The number of participating members may not be 100% of the community members. Thus, it is important that the concept of community differs from 'meeting group'. Meeting groups can be the base of inferring the community. Tracking the dynamics of the meeting should show the inter-relationship of members within the community. Our goal is inferring dynamics of meetings in human connectivity networks based on the traces collected human connectivity by sensors and the work is in preliminary stage. We claim our contribution in this paper is two-fold: 1) our novel algorithm 'Cluster and Track (CAT)' to identify the significance of meeting groups, 2) as preliminary result, we show the power law behavior of meetings. This demonstrates duration of meetings for predicting network capacity or the limit of synchronisation mechanism. Even with noisy data, we believe that the result can lead to understanding dynamics of information flow between meeting groups.

The rest of this paper is structured as follows. We briefly introduce the expermental data collection process in Section 2. In Section 3, we analyse the duration of meetings. Finally, we conclude the paper with a brief discussion and future works in Section 4.

2 Human Connectivity Traces

The quantitative understanding of human dynamics is difficult and has not yet been explored in depth. The emergence of human interaction traces from online and pervasive environments allows us to understand details of human activities. For example, the Reality Mining project [4] collected proximity, location and activity information, with nearby nodes being discovered through periodic Bluetooth scans and location information from cell tower IDs. Several other groups have performed similar studies. Most of these [4] [5] use Bluetooth to measure device connectivity, while others [6] rely on WiFi. The duration of experiments varies from 2 days to over one year, and the numbers of participants vary. We have analysed various traces from the Crawdad database [3] and in this paper, we show the results using the following traces.

MIT: in the MIT Reality Mining project [4], 97 smart phones were deployed to students and staff at MIT over a period of 9 months.

CAM: in the Cambridge Haggle project [9], 36 iMotes (see Section 2.1) were deployed to 1st year and 2nd year undergraduate students for 11 days. iMotes detect proximity using Bluetooth.

INFC06: 78 iMotes were deployed at the Infocom 2006 conference for 4 days [2].

2.1 Proximity Detection with Bluetooth

Bluetooth is a low-power open standard for Personal Area Networks (PANs) and has gained its popularity due to its emphasis on short-range, low-power and easy integration into devices. The platform used in the Haggle experiments is the Intel Mote ISN100-BA (known as the iMote). The iMote runs TinyOS and is equipped with an ARM7TDMI processor operating at 12MHz, with 64kB of SRAM, 512kB of flash storage, and a

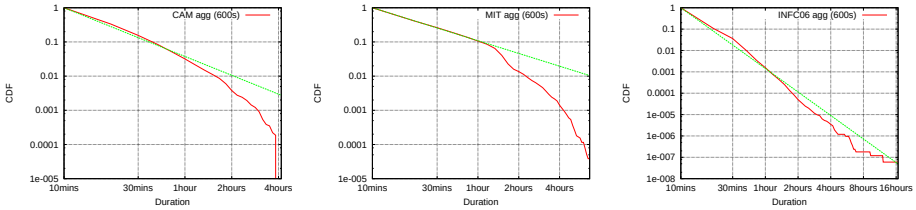


Fig. 1. Distribution of Meeting Times - Aggregated all K results

multi-colored LED, and a Bluetooth 1.1 radio. The specifications lists the radio range to be 30 meters.

It is a complex task to collect accurate connectivity traces using Bluetooth communication, as the device discovery protocol may limit detection of all the devices nearby. Bluetooth uses a special physical channel for devices to discover each other. A device becomes discoverable by entering the inquiry substate where it can respond to inquiry requests. The inquiry scan substate is used to discover other devices. The discovering device iterates (hops) through all possible inquiry scan physical channel frequencies in a pseudo-random fashion. The power consumption of Bluetooth also limits the scanning interval, if devices have limited recharging capability. The iMote connectivity traces in Haggle use a scanning interval of approximately 2 minutes, while the Reality Mining project uses 5 minutes. The ratio of devices with Bluetooth enabled to the total number of devices is around only an average 15% of population.

3 Inferring Significance of Meeting Time

It is important to make the distinction between the dynamics of meetings and the evolution of social networks. While the dynamics of meetings govern the physical connectivity over time as meetings are formed and dissolved, whereas the evolution of social networks involves changes to the social connectivity between people over time across multiple meetings. So the former operates at the timescale within a meeting and between consecutive meetings, the latter operates over the timescale of many recurrences of meetings. In this section, we show results for the analysis of meeting dynamics from duration of meeting time. We have used various community detection algorithms in our previous work [7] and found K-CLIQUE [10] shows stable results for different types of human contact traces. Thus, we demonstrate inferring physical group meetings using K-CLIQUE algorithm in this paper. The inter-contact time is the time interval between two contacts. Inter-contact time is the duration from when one contact finishes and the next one begins, it determines how often a communication is possible. Shorter inter-contact time means that the two people see each other quite often. The number of such contacts and the distribution of contact durations is an important factor in determining the capacity of PSNs. It gives insight on how much data can be transferred at each opportunity. This concept of pair node relationship can apply on the relationship between meeting groups, which helps building group oriented communication protocols.

3.1 K-CLIQUE

Palla et al. define a k-clique community as a union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k-cliques

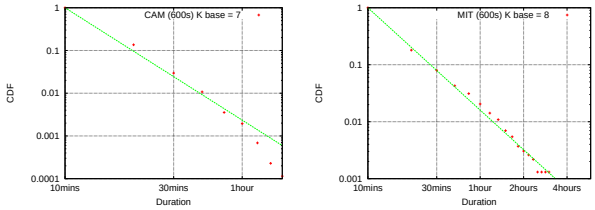


Fig. 2. Distribution of Meeting Time without duplication

[10]. Two k-cliques are said to be adjacent if they share k - 1 nodes. This definition is based on their observation that an essential feature of a community is that its members can be reached through well-connected subsets of nodes, and that there could be other parts of the whole network that are not reachable from a particular k-clique, but they potentially contain further k-clique communities.

A problem with using k-cliques for detecting meetings is that of uniqueness. A cluster of nodes might have more than one overlapping k-clique embedded in it. For example, the largest k-clique size in a cluster might be five, but there can still be two overlapping 5-cliques, and by definition, each of those 5-cliques has embedded within it 5 overlapping 4-cliques and so on. Thus one must not only be careful about counting only one of the two 5-cliques, but also not to count the 4-cliques and 3-cliques that are subsets as well.

The figures 1-3 demonstrate how the overlapping problem affects the statistical analysis. In Fig.1 shows the CDF distribution of meeting times, where detected meetings from K=3 to K=max are aggregated. These thus include overlaps of each clique, thus a 5-clique would also be counted five times as a 4-clique, and so on. The distribution is then for the sum of these K-clique counts. In more detail, Fig.3 shows the individual distribution with all clique of size k=4 including overlaps. We note that larger cliques exhibit more power-law behaviour in the MIT dataset, whilst for the INFC06 dataset it is more power-law in behaviour for lower clique sizes. Fig.2 shows the distribution, where overlapping cliques are removed. We try to remove overlapping sub-cliques by working backwards from a high-clique base and removing all sub-cliques from the results. For example, when the base K=7, in CAM data, all subgraphs in K=6 to K=3 are removed for the duration of a 7-clique meeting. This ensures that cliques of size k are only counted once as size k. The figure shows a more clear power law distribution with a reduced rms error in the fit. We note however that, in the presence of noise, removing subgraphs is not a straightforward process without knowing the detection failure rate. For example, nodes a-b-c may form a 3-clique from times 1-10, then a 4-clique a-b-c-d from times 11-15, then b-c-d from times 10-20. The clique elimination used here was to count higher cliques as having precedence. Thus the duration of the cliques in this example detects times 11-15 as a 4-clique, and separately detects times 1-10 and 15-20 as 3-cliques.

3.2 Cluster and Track (CAT)

In this section we introduce our novel algorithm called ‘CAT’ (Cluster and Track), which builds clusters at each time slice based on the edge connectivity. The algorithm works iteratively and greedily by growing a cluster. Fig.4 helps illustrate the algorithm.

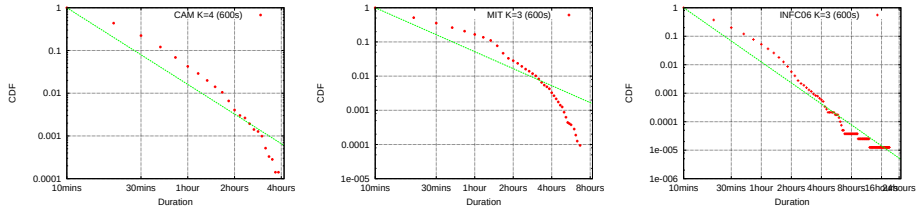
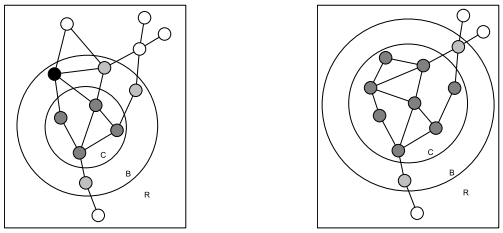


Fig. 3. Distribution of Meeting Time - Case K=4

Let the cluster C denote the set of nodes in the cluster. Let B be the set of nodes that are immediate neighbours of the cluster C. Let R be the set of remaining nodes. For each node in B, it compares the number of edges to C to the number of edges in R. If there are more edges to C, it adds the node to the next version of the cluster. In Fig.(4a) only the black node in the boundary region B satisfies this criteria and is added to the cluster. The process is repeated until the cluster cannot grow any further. The end result of the cluster growth is seen in Fig.(4b) - notice how the number of nodes in the cluster has grown to eight. By trying each node as a seed node, this greedy growth algorithm generates candidate clusters around each node.



(a) (b)

Fig. 4. CAT Algorithm

The next stage of the algorithm then greedily selects a smallest covering set of clusters. Determining the smallest covering set is helped by utilising dominance relations between clusters. This is where a candidate clusters can be eliminated because there is typically another larger cluster candidate than fully contains them thus dominating them. The clusters in the covering set are assigned an ID at each time slice. Each cluster is compared against the covering set of clusters from the preceding time-slice. If half or more of the nodes from the previous slice's cluster is seen in a current cluster then the ID is preserved in this cluster. When two clusters merge, the cluster ID from the larger number of preserved nodes is kept. When clusters split, again the cluster with the largest number of preserved nodes retains the ID and the rest are assigned a new ID. This, then, comprises a simple scheme for determining clusters and tracking them through time. This approach allows one to estimate meetings in the form of clusters, and track their duration. Such meetings are considered their own separate entity, and it is interesting to see if they exhibit any interest statistical behaviour. Fig.5 depicts the distribution of meeting time using the CAT algorithm, which exhibit more significant power law characteristics.

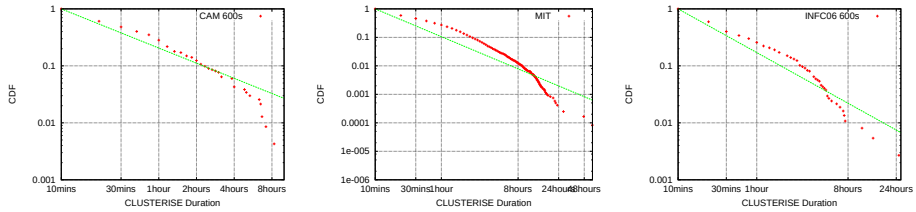


Fig. 5. Distribution of Meeting Time - CAT

4 Conclusions and Future Works

In this paper, we have shown the dynamics of meeting and the presented work has wide future extensions. Uncovering temporal and dynamics of meeting groups can be used as a signature for constructing synthetic network generation with the information of subgraph structure and dynamics. Most importantly extracted model must be validated in some way of real experiments. The iteration of modelling and experiments will uncover further understanding of time-dependent complex human connectivity networks. Future works include: identifying the significance of meeting using transitivity, classifying behavior of nodes in the core/transient meetings, and dynamics of flow between meetings.

Acknowledgments

The research is part funded by the EU grants for the Haggle project, IST-4-027918, the SOCIALNETS project, 217141, and the EPSRC DDEPI Project, EP/H003959.

References

1. T. Berger-Wolf and J. Saia. A framework for analysis of dynamic social network. In *Proc. KDD*, 2006.
2. A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. In *Proc. INFOCOM*, April 2006.
3. Dartmouth College. A community resource for archiving wireless data at dartmouth, http://crawdad.cs.dartmouth.edu/index.php, 2007.
4. N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, V10(4):255–268, May 2006.
5. EU FP6 Haggle Project. http://www.haggleproject.org, 2010.
6. T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. In *Proc. Mobicom*, 2004.
7. P. Hui, J. Crowcroft, and E. Yoneki. BUBBLE Rap: Social Based Forwarding in Delay Tolerant Networks. In *MobiHoc*, 2008.
8. P. Hui, E.Yoneki, S. Chan, and J. Crowcroft. Distributed community detection in delay tolerant networks. In *Proc. MobiArch*, 2007.
9. J. Leguay, A. Lindgren, J. Scott, T. Friedman, and J. Crowcroft. Opportunistic content distribution in an urban setting. In *ACM CHANTS*, 2006.
10. G. Palla, I. Dereny, I. Farkas, and R. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
11. E. Yoneki. Visualizing Communities and Centralities from Encounter Traces. In *ACM MobiCom - CHANTS*, 2008.
12. E. Yoneki, P. Hui, S. Chan, and J. Crowcroft. A socio-aware overlay for multi-point asynchronous communication in delay tolerant networks. In *Proc. MSWiM*, 2007.