Consolidated Review of

Signals from the Crowd: Uncovering Social Relationships Through Smartphone Probes

1. Strengths:

The paper studies the SSID probes made by mobile phones in order to characterize social relationships. The authors build social graph from an affiliation graph, enabling them to connect between users that share similar advertised SSID.

An impressive and novel dataset, overall good analysis.

Clever way to track people and build a social network. Solid analysis of the resulting network.

I found the paper interesting. Previous work has leveraged SSID announcements, but as far as I can tell this seems the first to try to understand the social networks these phones belong to. The paper is in general well written. I always appreciate when papers give me an intuitive sense of the math and this one did that well.

2. Weaknesses

The method is similar to the MobiCom 2007 paper, and the primary difference is in the application. Some of the results presented as surprising seem well known: that a person can be tracked by mobile phone, that most social networks have similar form, that people at a location show regular patterns of arrival and departure.

3. Comments

This is a nice paper, however, it is very incremental relative to the work of Pang et al from Mobicom 2007 [28]. This paper provides a nice result, building a social network from observations of personal devices. (Although this result is not completely novel, see [9].)

I really liked the way you collected the data, but I find the results somewhat not insightful. This method of discovering a social network from a crowd seems clever, novel, and effective. The analysis of the resulting social network seems fairly typical, but carefully done.

Some issues:

1. The results that you show are for very specific events, so they cannot really be generalized to anything insightful (e.g., people in the political2 even use slightly more iPhones than people in the mall). The methods to infer these results are mostly applying existing graph processing techniques, and applying them to a new graph, which is derived from data similar to the one in [28]. So the overall applicability and novelty of this paper is limited.
2. A problem with the formation of the social network is the number of repeating SSIDs. When the SSID is "DLink", there is no real way to differentiate between two users that use the same DLink access point and users that are completely not related. I agree that the Adamic-Adar distance metric solves this problem, but I could not find a rigorous analysis of the "right" value of \tau (the threshold).
3. I think you overly stress the "sociological" aspects of the results. I do not really see any strong insights that can be generalized about human behavior in this paper.

The paper makes several statements about results that are "surprising", when they are often well known:

- "This scenario rises a natural question: `Can significant information on the owner of the devices be inferred by smartphone probes?"" But it's well established that mobile phones leak information (see paragraphs 2 and 3 of the related work section)
- "Quite surprisingly, despite the clear differences between the events, the structural properties of the underlying social graphs are very similar". But given the widespread analysis of social networks showing most have similar properties (including this week a report analyzing Homer's the Odyssey); it would be surprising if these social networks did NOT look similar.
- " a temporal analysis of the data collected in our long-term campus deployment showed a strong correlation between the frequency of the co-occurrence of devices in the same time slot and the strength of the relationship inferred by our methodology." Results of regular movement patterns of people and devices has been published before: N. Eagle and A. S. Pentland. Eigenbehaviors: identifying structure in routine. "Behavioral Ecology and Sociobiology" 63(7):1057–1066, May 2009, and "Data Muling with Mobile Phones for Sensornets", Park and Heidemann, ACM Sensys 2011. In addition, there is missing prior work that constructed social networks from chance meetings. See "A High-Resolution Human Contact Network for Infectious Disease Transmission", Salathe et al, Proc. National Academy of Sciences, 107 (51), Dec. 2010.

This is not to take away from this paper's contribution, but to help put it into perspective.

The dataset description in section 3.1.1 might better be done as a table. About figure 1: the y axis is a fraction (from 0 to 1), not a percentage (from 0 to 100). Please add a second y axis that shows the absolute count of devices. About figure 2: it would be nice to show these percentages as second bars over the bars on figure 1.

Cool dataset. Nice to see a form of the dataset will be released. "Vatican 1" ... "Vatican 2" ... nice! :)

Outside Expert: This is an interesting paper and there is some interesting analysis presented here. I would like to see a discussion on what can be done to fuzzy up and lessen the amount of information smartphones leak about themselves and their users. There is no reason why a local scope MAC address cannot be used in the Probe Request, for example. And what value do the vendors see in including Vendor information in the Probes? Why is there only ONE PNL on a device and not a role-based selection of PNLs? These are the sorts of questions I would hope that would come out from such a paper, rather than "What ELSE can we get phones and people to leak about themselves?

4. Summary from PC Discussion

The TPC felt its expertise was low, so one of the co-chairs solicited an outside review from one of the world's leaders in

SSIDs and their use. That review (at the end of the comments above) indicated the paper was novel, and so it was accepted.

5. Authors' Response

We thank the reviewers for their insightful comments. In response, we rewrote some parts of the text to help the reader better understand our contributions, while leaving the structure of the paper and the results unchanged.

Our study is focused on investigating whether probes requests collected in an event can provide hints about relevant sociological properties of the crowd that participated to it. We believe that this is a clear step in a novel direction with respect to [9] and [28]. Indeed, the aim of our study was neither that of discovering new, general properties of human behavior, nor that of identifying 802.11 device users by means of implicit identifiers (as in [28]). We achieve our goal by applying different well-known analysis techniques to large-scale datasets of probes collected in very specific, meaningful, and different scenarios. This adds a whole new dimension to previous studies, such as [9], that just focus on using probes requests to discover social links between pairs of people. According to our findings, there is a noticeable correlation between the sociological characteristics of the portion of society that attended to an event and the properties of the corresponding dataset. For instance, we found the distribution of the language of the SSIDs to vary according to the national or international nature of the events, and the vendors' popularity to match the observed attendees' social extraction and economic status. The sizes of our datasets, which are in the order of several thousand of devices, allow us to regard as significant even apparently small variations of, say, 5% in popularity of a given vendor. The soundness of our analysis is supported by the fact that we found similarities across different events of the same type (e.g., Vatican 1 and Vatican 2, Politics 1 and Politics 2). A larger collection of datasets would certainly help further validating and generalizing our ideas and methodology. Consider however that, to the best of our knowledge, we are the first to collect datasets of these sizes. We have added these comments in the text and improved presentation throughout the whole paper.

We made the objective of our study more clear in Section 1 and 4. Now the paper better highlights the hints and lessons that can be learned from the results of each step of our analysis. Section 4.3 better explains the purpose of our social network analysis: We agree it is not surprising that social networks emerging from different contexts share similar properties. Still, we are the first to show how probes requests allow uncovering meaningful social networks underlying large crowds of people. Comments on the results of Section 4.6 have been improved too: Our objective was not that of presenting well-known properties of human social behavior, but, rather, to show how a temporal analysis reveals other valuable information about the portion of society sampled in a dataset. Finally, following the reviewers' suggestions, we improved the readability of the figures and tables in the paper, and made our references more complete.