This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

Understanding Human Behavior at Large Scales through Mobile Devices

Jason I Hong **jasonh@cs.cmu.edu**
Joy Zhang **joy@cs.cmu.edu**
John Zimmerman **johnz@cs.cmu.edu**
Carnegie Mellon University

Abstract

Smart phones, which come with a rich array of sensors and communication capabilities, are becoming more affordable and increasingly adopted by the general population. In this white paper, we argue that the increasing pervasiveness of these mobile devices offers researchers the ability to understand human behavior and real-world social interactions at a fidelity and scale that could not be accomplished before. We provide a brief case study of our own work in using location data from over 400 people to characterize aspects of both those people as well as the places they went. We also sketch out two major research challenges, in terms of the logistics of conducting this kind of research as well as managing privacy issues. We also argue that the costs of this kind of research are quite high, and that researchers need much more support from government and from industry to pursue this line of work, and that a large data collection and data sharing effort may be an effective way to proceed.

Keywords: mobile computing, sensors, social networking, location

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

1. Introduction

Today's **mobile devices** come with an incredible array of capabilities. A commodity smart phone can sense location, sound, proximity, and motion. These devices also have access to call logs, SMS logs, pictures taken, and email. These new capabilities offer us the opportunity to analyze **real-world social networks and human behaviors** at unprecedented fidelity and scale, in ways that previously were simply not possible. As such, mobile devices can be thought of as **a new scientific tool for computational social science** of real-world interactions, activities, and behaviors.

More specifically, many questions of human behavior which in the past were difficult to answer due to limitations of methods and issues of scale, will soon become possible to measure. This approach has already started to bear fruit. For example, Pentland and his research group have conducted a series of studies using wearable sensors, revealing new insights into how we influence and interact with each other using relatively simple wearable sensors [2].

1.1. Analyzing Location Data

As an example of this kind of research, we offer a short case study of our past work in analyzing location traces of **453 participants** of our Locaccino friend finder system, which amounted to **2.4 million location observations over the course of several months**. Based on this data, we adapted the idea of entropy (used in ecology as a measure of biodiversity) to characterize one aspect of the places people went. Using data about people and places, we (1) created a classifier that could predict whether two people were friends or not with 92% accuracy, using co-location data and where those co-locations happened [1]; (2) found correlations between a person's mobility patterns and the number of friends they have on Facebook [1]; and (3) found correlations between entropy and privacy, suggesting that we may be able to create effective privacy defaults for cities [3].

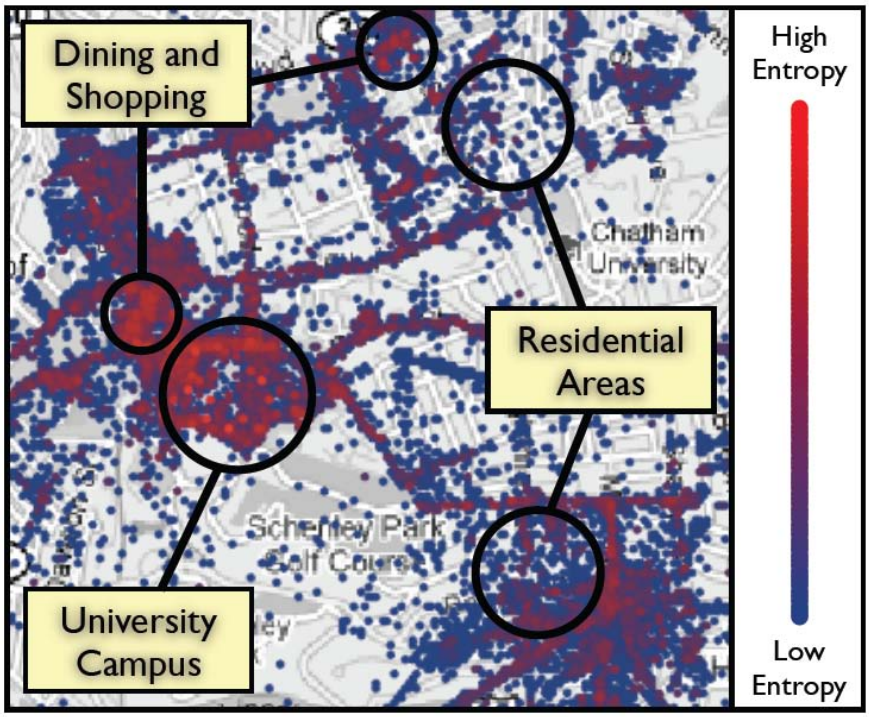

Figure 1. Entropy map for Pittsburgh, based on data we collected. Entropy is a measure of how many unique people were seen in a location over a given span of time. Entropy is also a weak approximation of how public or private a place is.

This kind of sensor data can be used to generate at least two kinds of graphs: (1) one of **people and places**, and (2) an inferred graph of **people and their connections to other people**. The former is a new kind of graph that opens up new opportunities for analysis. The latter is similar to existing social graphs like Facebook and MySpace. Together, these graphs can be a powerful new tool for a number of domains. Below, we present some examples of new research enabled by approaches like this:

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

- Using co-location data, call log data, and other data, we believe it is possible to computationally assess the **type of relationship** (e.g. parent-child, co-workers, supervisor-supervisee, friends, acquaintances) and **strength of a relationship** between individuals (a spectrum from weak tie to strong tie). This information could be used to automatically or semi-automatically organize one's social graph into meaningful clusters. These clusters could then be used as the basis for stronger privacy, filtering and prioritization of incoming messages (e.g. "prioritize messages from my close friends and family"), and appropriate sharing of photos, tweets, and Facebook updates.
- Using mobility data and call log data, it is also possible to computationally evaluate the **physical and mental health** of an individual and a population. We believe this line of research can also be done at very large scales. For example, using mobility data, we believe we can estimate if a person is an extrovert or an introvert. We believe we can also estimate how often a person spends time with friends, and how that changes over time (with a decline possibly signifying the onset of depression). We can also see if there is decline in mobility over time, possibly signaling the onset of physical difficulties that the individual is not yet aware of.
- Using sensed data, we believe we could evaluate **building architectures and urban design**. For example, what kinds of layouts in buildings lead to more interaction and collaboration between individuals (i.e., more ties and stronger ties between people that live or work in a place)? What kinds of layouts facilitate serendipitous encounters? Similarly, at a larger scale, what kinds of neighborhoods tend to have more interactions and friendships between neighbors? This kind of data collection and analysis could be replicated at scale across large numbers of buildings and cities, to empirically understand how to improve design rather than through the guesswork and intuition used today.

There are many other possibilities our group has discussed, including recommender systems for places, anomalous crowd detection, traffic monitoring and estimation (which is already being done by Nokia and UC Berkeley), ridesharing pairing recommendation, understanding what tourists and residents of a city do, creating a topic map of what people did this past weekend, detecting interesting events that happened in a city, generating fine-grained data on people in that city for economics and urban planning purposes (e.g. 58% of people in this city work 40 hrs a week, average wait time for buses is 12min), epidemic modeling (a popular example among researchers), and so on.

2. Fundamental Research Questions

Due to space, we highlight what we believe are the two biggest challenges in pursuing this kind of research.

2.1. Data Collection at Large Scales

Data collection at scale is a significant challenge. Gathering data from hundreds or even thousands of people is very expensive in terms of cost of hardware, software development costs, data network fees, recruitment, and incentives. We focus primarily on smartphones, given that these are the most promising platform for this kind of research in the near-term.

With respect to **hardware**, the costs may be in terms of developing new hardware that gathers the desired data, as well as in lost and stolen hardware. Typical smartphones today cost on the order of $500, making it expensive to scale these kinds of studies up. With respect to **software**, the challenges are in ensuring that the data being collected is accurate, and that power consumption is low. In some of our past work, we found that a typical smartphone lasts for about 6 hours if the GPS is left on, making it infeasible for everyday use. There are also major compatibility issues, which we discuss below.

With respect to **data network fees**, smartphones require a data plan for them to be effective, which can cost on the order of $30 per month (not including regular voice plans). To mitigate these costs, one approach is to give people another phone that only uses WiFi. Data is collected and eventually uploaded by participants. The problem with this approach is that people are required to carry two mobile phones

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

(both their own and the one for the study), and keep two phones charged. In our experience, this approach works very poorly, since people are very likely to forget about the other phone. Another approach is to recruit people who already have GSM SIM cards, which can be transferred from one phone to another. However, this means that participants can only be recruited from a specific pool, excluding, for example, Verizon, since it does not support GSM. As such, this constraint locks out a number of people.

Incentives are also a major issue. If the software requires volunteers, then participants need a compelling reason to install the software on their mobile phones, which is no easy feat. It also means that the research software has to either target a specific smartphone platform (essentially Apple iPhone, Google Android, Nokia Symbian, or Blackberry), or has to be built on multiple platforms, which is a very challenging and expensive proposition for research groups, in terms of development and maintenance. Simply paying participants is somewhat viable, but gets very expensive quickly if one wants to study people at large scales.

These are all surmountable issues, but the costs are very high. One possible approach is for NSF to acknowledge these costs and offer more funds for this kind of research. Transformative research along these lines is barely viable on typical grant budgets, and only if researchers have a lot of supplemental support from industry. Another possible approach is for NSF and researchers to band together and work towards large-scale shareable test beds.

For example, NSF may opt to make a big bet in three or four projects along these lines, with the data to be shared with many researchers around the United States. The initial teams could be comprised of a core of social and technical scientists. The social scientists would participate by offering several ideas for what data to collect and what analyses to conduct, as well as to ensure that the data is gathered ethically and accurately. The technical scientists could ensure that data collection is power-efficient, secure, flexible, and robust to error. Additional data analysis could then be carried out by various teams around the country. This kind of work could also be done in stages, starting with dozens of people to demonstrate proof of concept and to mitigate risks. Afterward, the research community could push in efforts to scale up to hundreds, thousands, and eventually tens of thousands of people.

2.2. Managing Privacy

Sharing of data leads to the next issue, which is **privacy**. Gathering this quantity of information poses significant legal, social, and technical challenges.

Here, due to space, we will focus on only one issue, which is how to share the data with other researchers, while divulging minimal personal details? Using the Locaccino data set we described above, it would actually be fairly trivial to guess where a person works and lives. With larger data sets, it may also be easy to find potentially embarrassing things (or make incorrect inferences that may be potentially embarrassing).

As such, another research challenge is generating new ways of sharing this kind of collected data. One possibility may include sharing abstractions of the data (e.g., instead of sharing GPS locations, share names of places, like “home”, “work”, “café”, etc). Another possibility may be to have a single share server that other researchers can run software on and only get summary results, based on some constraints (e.g., no information about individuals are allowed, or information will only be allowed in buckets that have 50 people or more).

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

3. Summary

In this white paper, we outlined some of the opportunities for using the sensors and communication log data of mobile devices, to understand human behavior and real-world social interactions at a fidelity and scale that could not be accomplished before. We also sketched out two major research challenges, including the logistics of gathering this data as well as managing the privacy issues. We also argue that the costs of this kind of research are quite high, and that researchers need much more support from government and from industry to pursue these ideas, and that a large data collection effort and sharing may be an effective line of attack.

References

1. Cranshaw, J., Toch, E., Hong, J.I., Kittur, A., and Sadeh, N. Bridging the Gap Between Physical Location and Online Social Networks. *Ubicomp 2010*, (2010).

2. Pentland, A. *Honest Signals: How They Shape Our World (Bradford Books)*. The MIT Press, 2008.

3. Toch, E., Cranshaw, J., Drielsma, P.H., et al. Empirical Models of Privacy in Location Sharing. *Ubicomp 2010*, (2010).