Smartphone Data at Scale: Small Devices, Big Opportunities, Bigger Risks

Jason Hong, Carnegie Mellon University, Human Computer Interaction Institute

1. Introduction

In 2013, smartphones were close to 40% of all phones sold worldwide. These mobile devices come with an incredible array of capabilities. A commodity smart phone can sense location, sound, light, proximity, and motion. These devices also have access to our contact lists, call logs, SMS logs, pictures taken, email, browsing history, our activities on social networking sites, and more.

Altogether, these new capabilities offer us the opportunity to analyze real-world social networks and human behaviors at a fidelity and scale that previously was not possible. As such, mobile devices can be thought of as a new scientific tool for capturing and understanding 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. Below, I give one example of the work our team has been doing in this space in what we call *urban analytics*.

At the same time, these same capabilities pose *new kinds of privacy and security issues*. The risks are partly due to the widespread adoption of apps (roughly 500,000 available apps and over 40 billion downloads for each of Apple and Android's app markets), the level of intimacy we have with our devices (e.g. many teens report sleeping with their phones), and the growing range of accurate (and inaccurate!) inferences that can made. These are not just hypothetical risks either: overly intrusive or even malicious apps that misuse the data and the capabilities on our smartphones have already emerged. Furthermore, these risks will only get worse, as more powerful apps are deployed (for accessing our finances, our cars, and our homes), and as our data is spread out among multiple third parties (e.g. advertisers, insurance companies, doctors). Below, I give a brief overview of some of the risks that we have encountered in our work, as well as some potential solutions for using big data approaches to help mitigate these concerns.

2. Urban Analytics = Smartphones + Geotagged Social Media + Machine Learning

The forces that shape the dynamics of a city are multifarious and complex. Cultural perceptions, economic factors, municipal borders, demography, geography, and resources all shape the texture and character of urban life. However, it can be extremely difficult to study these intricacies. For example, classic studies (e.g. Whyte, Lynch, Milgram, Jacobs, and others) have found deep insights about city life. However, these studies also required hundreds of hours of observation, interviews, and analysis. These methods simply do not scale, and hence can only uncover a partial image of the inner workings of a city.

We argue that there is an exciting opportunity for creating new ways to conceptualize and visualize the dynamics, structure, and character of a city by analyzing the social media data that people generate on their smartphones. Towards this end, we developed Livehoods, our first urban analytics tool (see Figure 1). Our research hypothesis with Livehoods is that the character of an urban area is defined not just by the types of places found there, but also by the people that make it part of their daily life. To explore this idea, we crawled 18m check-ins from the location-based social network foursquare, and applied clustering algorithms that grouped nearby venues into areas (which we call "livehoods") based on geographic distance and the particular mix of the people who check-in to them.

To evaluate our work, we conducted interviews with locals, including city planners, business owners, and residents. We asked them to draw parts of the city that they were most familiar with (before they saw our maps), describe characteristics of the areas they were most familiar with, and offer feedback on the livehoods our system generated. In many cases, our livehoods matched the mental models of locals, and also provided new insights about how neighborhoods were organized and were changing over time.

Livehoods is our first urban analytics tool, but we believe there are many other rich opportunities here for location data and social media data, for helping urban planners, policy analysts, politicians, social

scientists, and businesses understand how people actually use a city, in a manner that is cheap, highly scalable, and insightful. Examples include understanding the behavior of different demographics in a city ("What do Chinese people do in Pittsburgh?"), economic development ("What was the spillover impact of the new Target store on other stores nearby?"), planning ("What kinds of services do residents of this neighborhood have to travel far for?"), sustainability and quality of life ("How many 'third places' are there in this neighborhood?" or "What is the location efficiency of this neighborhood?"), and measuring economic impact of events ("How many more people go to bars on football days?" or "How many fewer people go out because of the snow?").

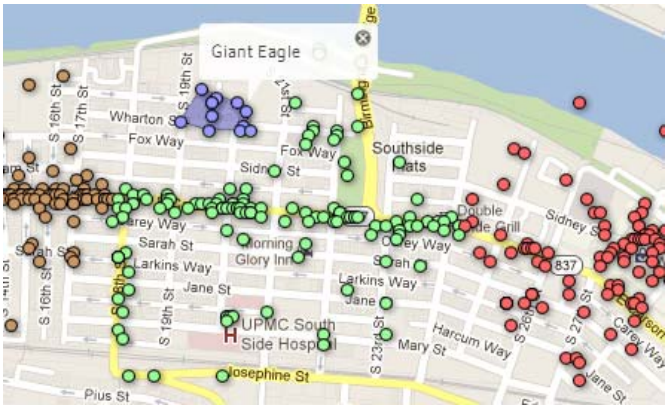

Figure 1. Four livehoods in Pittsburgh, based on clustering of foursquare data. These livehoods are based on geographic proximity of venues as well as "social proximity" based on co-occurrence of check-ins. The middle contains the only grocery store in the area. The right is a shopping mall. The left has many bars that students frequent.

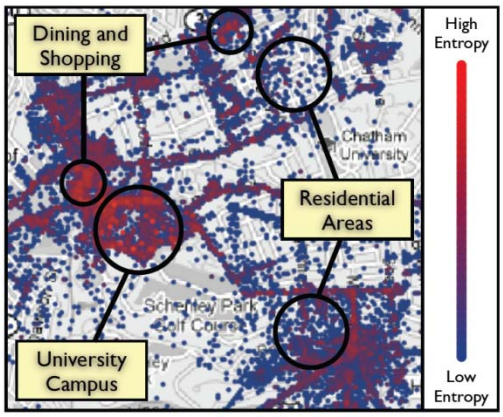

Figure 2. Entropy map for Pittsburgh, based on location data we collected. Entropy measures how many unique people were seen in a location over a given period of time. Entropy roughly describes the social quality of a place, in terms of how public or private a place is.

3. Smartphone Privacy Risks

Urban analytics is just one example of the kinds of benefits that data gathered from smartphones can offer. However, this same smartphone data poses new kinds of privacy risks. On an individual level, one significant challenge we have seen is a lack of understanding by end-users of what kind of data can be gathered by smartphone apps, as well as what data is *actually* being gathered.

In our own work, some of the more surprising behaviors we have seen include a flashlight app that requires Internet access and phone number, a backgrounds app that uses contact list, and a Bible app that uses location data. One of the main reasons why so much data is being collected is for advertising: there are currently only a few viable business models for apps, with advertising being one of the main approaches. However, advertising networks have a strong incentive to collect more data about customers, so as to tailor ads and increase relevancy and clickthrough rates.

We believe big data techniques can help address some of these privacy issues. In one research thrust, we operationalized privacy as expectations, probing the gap between what people *think* an app does and what it *actually* does. For example, most people don't expect Angry Birds to use one's location data, but in reality it does. This mismatch represents a big surprise to people. On the other hand, most people do expect Google Maps to use one's location (and it does), meaning that users have more awareness of what the app is doing.

To find these mismatches in a scalable fashion, we use crowdsourcing. Approaches that rely solely on static or dynamic analysis cannot understand people's perceptions of privacy. However, with the advent of crowdsourcing markets, we can dissect the behaviors of apps, have crowd workers analyze those behaviors, and then aggregate those analyses into a summary of an app's privacy-related behaviors. In

our studies, we found that this combined approach was effective in flagging unexpected behaviors of smartphone apps. Our current work looks at scaling up this approach to hundreds of thousands of apps, by building models to predict potential privacy concerns based on ground truth on a core set of apps. An example output might be “games that use location data -> -0.7 points to the app’s privacy score”.

As another example of using big data techniques to help with privacy, some of our past work suggests there are patterns in people’s mobility patterns that can be used to help tailor location sharing preferences. Our team has developed a friend finder system named Locaccino to study how people use location-based systems and manage privacy concerns. Locaccino runs on laptops and smartphones, and sends location data every 5 min to our servers. We analyzed location data from 453 users of Locaccino, which amounts to 2.4 million location observations. One analysis we did was to model *place entropy*, which characterizes how many unique people have been seen in a given place. High entropy places tend to be public places, whereas low entropy places tend to be residences (see Figure 2). Entropy is one example of a kind of analysis that can be done only once there is enough location data.

Interestingly, we found that place entropy is related to people’s comfort in sharing their location with others (see Figure 3). We had 28 participants use Locaccino over a period of four weeks, and had them rate the comfort level of 12 specific locations randomly sampled from their data. We found that as place entropy increases, comfort level also increases. If this finding holds in general, then this approach could be a new way of setting default privacy policies for when data should be shared in general.

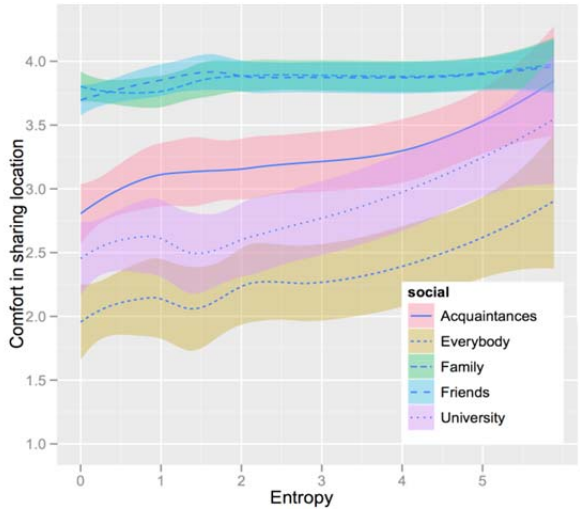

Figure 3. This figure shows people’s self-reported comfort levels in sharing their current location with others in different places. Interestingly, as place entropy increases, comfort level also increases, even for the “everybody” category. Colored regions represent the error bounds.

4. Concluding Remarks

Smartphone data at scale offers big opportunities for understanding human behavior, for addressing serious problems that we are facing in energy, sustainability, urban planning, and more. However, this same data poses potentially bigger risks with respect to privacy, in terms of the intimacy of data that can be gathered, the kinds of inferences that can be made, and the impact on all facets of our lives. We hope that this position paper has opened up new possibilities for smartphones, big data, and privacy, and offered food for thought as to directions for our community.

Acknowledgments

Special thanks to Lorrie Cranor, Norman Sadeh, and the members in the CUPS and CHIMPS research groups. Also, thanks to all of the people over the years who have helped refine the arguments in this position paper. The opinions in this paper are solely those of the author.