Research Issues for Privacy in a Ubiquitously Connected World

Jason Hong (jasonh@cs.cmu.edu)
Associate Professor, Human Computer Interaction Institute, Carnegie Mellon University

1. Introduction: The Two-Edged Sword of a Ubiquitously Connected World

In just the past decade, we have seen the invention and adoption of wireless networking, smartphones, big data and predictive analytics, wearable computing, autonomous vehicles, sensor networks, social media sites, massive open online courses (MOOCs), and an array of many other wondrous technologies. We are now at a major inflection point, as computation, communication, and sensing are woven into our everyday lives.

A very likely scenario is that, *in the near future, our computing systems will know everything about us*. They will know how healthy we are and be able to detect the early onset of a variety of illnesses. They will know what our carbon footprint is and offer ways of being more green. They will even know what our information needs are, before we even know what we want. In short, there are tremendous opportunities here to improve all aspects of daily life and society in general, in domains as diverse as healthcare, safety, sustainability, transportation, education, urban planning, finance, and more.

A fundamental problem, however, is that these same technologies also introduce many new privacy risks, often at a rate faster than legal mechanisms and social norms can adapt. In a ubiquitously connected world, the costs of collecting, storing, inferring, searching, and sharing data are dramatically lowered. The privacy risks range from everyday ones—such as monitoring from overprotective parents, undesired social obligations with friends and family members, and overly intrusive marketing—to extreme ones, such as threats to civil liberties by governments as well as dangers to one's personal safety by stalkers, muggers, and domestic abusers. Every day, there is some new headline, interview, op-ed piece, blog post, research paper, or book describing people's concerns regarding the strong potential for abuse, general unease over a potential lack of control, and overall desire for privacy-sensitive systems. In some cases, these concerns have even led to outright rejection of systems, strongly suggesting that privacy may be the greatest barrier to creating a ubiquitously connected world.

I am a professor at Carnegie Mellon University, and I have been building sensor-based systems and studying the human dimension of privacy for over a decade now. My research has spanned a number of topics, looking at issues of usability and understandability, mobile computing (in particular smartphones), and sensors. My work has also looked at different parts of the ecosystem, including developers, marketplaces, operating systems, third parties, and end-users.

In this position paper, I sketch out several major research challenges for privacy. Parts of this position paper have been presented as a keynote talk at the Mobisys Workshop on Mobility and Cloud Computing, and the ideas within have been informed and refined through many discussions with privacy professionals, lawyers, friends, and fellow researchers.

Below, I present two different scenarios for privacy, probing the confluence of smart devices and inferencing, as well as issues in social media and big data. I also present some ongoing work my team is doing in investigating social influences on privacy, which I use as an example of how more social scientists can be brought on board to study issues of privacy.

2. Scenario: The Confluence of Smart Devices and Intelligent Agents

About 540 million years ago, a vast multitude of species suddenly appeared in the fossil record, along with major diversification of organisms. Paleontologists call this event the "Cambrian explosion." We are currently witnessing the computing world's version of the Cambrian explosion. In the past decade alone,

advances in sensing capabilities, screen technologies, solid state storage, battery life, and wireless networking have enabled the cheap manufacture of computers in all kinds of form factors.

If we imagine a pyramid split into three tiers, at the top of the pyramid we have just a few of devices that we interact with a lot. These are what we typically think of as computers, and include desktops, laptops, tablets, watches, and smartphones. In the middle tier are computational devices that we might only periodically interact with, but there are many more of these. Examples include refrigerators, thermostats, smart toys, game systems, and fitness trackers. At the bottom tier are computational devices that we are barely aware are there. Currently, there are few of these that everyday consumers use, but it is very likely that the numbers of these will dominate the other two. Today's examples include smart meters, radon detectors, carbon monoxide detectors, HVAC, security systems, and motion detectors. Tomorrow's examples might include smart toilets, tags and beacons to track where objects are, sensors that monitor where people are in the house and what they are doing, sensors that monitor the "health" of our houses and apartments (e.g. water in basement or water leaking through roof), and many, many more. Broadly speaking, these last two tiers can be considered the Internet of Things.

2.1. Example Scenarios of Use for Smart Devices and Intelligent Agents

One result of all of these smart systems is that far more data can be collected about us, enabling a rich set of analyses and inferences about an individual. As a simple example, Figure 1 shows screenshots from my smartphone. An intelligent agent (such as an advanced form of Apple's Siri or Microsoft's Cortana) could use information like this to infer who I know, what my relationships are with the people I communicate with, where I go, and what the important places in my life are. As an example vision of how far we might be able to push things, one large initiative we are undertaking at Carnegie Mellon University is InMind, a project to develop the next-generation of intelligent agents for smartphones. Given all of the rich information available on the web and on our smartphones, can we develop an agent that can continuously learn about users—for example, their tasks, who they interact with, where they spend their time, what they do in those places, what their information needs are, and what else is going on nearby—and use that information to improve interaction?

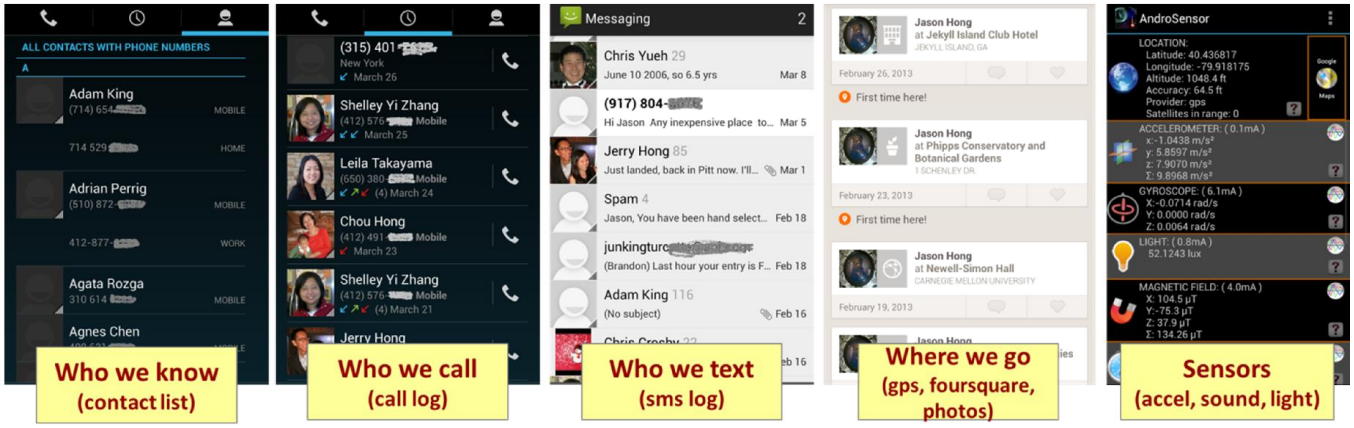

Figure 1. Even seemingly innocuous data, such as basic data logs and sensor data from our smartphones, can be aggregated together to form powerful inferences about our relationships with other people, the places we go, and the activities we do. These screenshots show example data from the author's smartphone.

In past work, my team used data like that shown in Figure 1 to infer whether a given contact was a family member, social relationship, or co-worker (Min, Wiese, Hong, & Zimmerman, 2013), and how close a person felt with a given contact given communication patterns (Wiese, Hong, & Zimmerman, 2015). Using the smartphone sensors, we have also been able to model a person's sleep quality and sleep quantity (Min et al., 2014). We are using these models of people's social behaviors, sleep, along

with physical activities (e.g. mobility patterns, motion) to infer the onset and exit of major depression in people (Doryab, Min, Wiese, Zimmerman, & Hong, 2015). All of these inferences are just from smartphone data. It is not hard to imagine that, as our homes, workplaces, cars, and clothes become instrumented, that inferences like these will become more commonplace and more accurate.

2.2. Privacy Challenges for Smart Devices and Intelligent Agents

The fundamental challenge for privacy is that, while these inferences can clearly be used for beneficial purposes, they can also be used in undesired and often in unexpected ways. In fact, we are already seeing privacy-intrusive behaviors on smartphone apps, and there is no reason to believe that the same won't happen for the Internet of Things.

For example, the Pandora music service was under Federal investigation because their app surreptitiously gathered location data, sex, birthdate, and unique Android ID, and shared this information with advertisers. As another example, the Path and Facebook smartphone apps were discovered to upload a person's entire contact list to remote servers without the user's knowledge (Bohn, 2012; Hachman, 2012). These are not isolated examples either: many apps have highly unexpected behaviors. Among my team's more surprising findings were that many popular games (like Angry Birds and Fruit Ninja), the Brightest Flashlight app, and even a Bible app sent a user's location data to various remote servers.

There are four reasons why we need to address these kinds of unexpected behaviors and make sure that people feel like they are in control. First, *people might not install an app in the first place*. When talking to people about our work on detecting depression, many people ask if their phone will be spying on them. We take great care to emphasize that people can choose to install our app, perhaps because they are entering therapy and their psychiatrist asked them to install the app, or because the user wants to see their own trends and prevent possible recurrences of depression. Second, *when people are surprised to learn about an app's actual behaviors, they might delete the app*, negating any potential benefits of these technologies. I have given several talks about my team's research on app privacy, and many people told me that they deleted apps from their phone during my talk because they were so surprised. As one example, it has been reported that WhatsApp may have lost over 500,000 users after Facebook purchased it (Russell, 2014). Third, *a lack of perceived privacy can lead to a media backlash*, which can hurt many other apps that might have positive benefit. Fourth, *a lack of perceived privacy might lead to overreactions or overreach by legislators and regulators* that could inadvertently stifle the adoption of promising technologies.

2.3. Research Opportunities for Privacy, Smart Devices, and Intelligent Agents

Table 1 presents some opportunities for privacy research in this design space. We have broken up our analysis based on various entities in the app ecosystem, including:

- *Developers*, people who create apps
- *Third-party developers*, people who create reusable code and web services that facilitate other developers (advertisers fall into this category)
- *App store owners*, people who aggregate and distribute apps (today, this would be primarily Google, Apple, Amazon, and Microsoft)
- *OS and hardware manufacturers* (for operating systems, primarily Google, Apple, and Microsoft)
- *Government and third parties interested in privacy*, including the FTC, EFF, CDT, and others
- *End-users*, the people who use smartphones and apps

We believe that this framing as well as the research opportunities below may generalize well for other domains, for example the web and the Internet of Things. Furthermore, the opportunities presented in Table 1 embody two general philosophies on privacy. First, we should try to prevent as many privacy problems as possible, but also make it feasible to detect and respond to problems as well. Second, the

burden of managing one's privacy should not be placed primarily on end-users, primarily because there are too many hardware components, software applications, and data flows for any single person to understand. Instead, the heavy lifting should be shouldered by all of the other entities, based on how effective and how reasonable it is for each category to manage things.

Entity	Opportunities for Privacy Research
Developers	Simpler ways of conveying privacy issues, verifiable rationales for data use, better tools for privacy, design patterns, more education
Third-party developers	Clearer explanations of what their software libraries do and why (primarily advertising)
App Store Owners	More effective scans of potential privacy problems, simpler summaries of potential privacy issues with apps
OS and Hardware Manufacturers	Privacy-sensitive APIs, countermeasures for monitoring
Government and Third Parties	Checking apps for privacy-related behaviors, better tools to find potential privacy problems with apps
End-Users	Understanding changes in attitudes and behaviors over time, better ways of expressing preferences

Table 1. Sample research directions for various entities in the smartphone app ecosystem. Note that while we focus on smartphones, the entities and opportunities generalize to many other domains, for example, web privacy and Internet of Things.

2.3.1. Research Directions for Developers and Privacy

With respect to *developers*, one major challenge is in designing *simpler ways of conveying the privacy implications of an app*. Today, the state of the art is to offer a long privacy policy full of legalese that few people (aside from lawyers) actually read in practice. Computer-readable formats (such as P3P for web sites (Cranor, Langheinrich, Marchiori, & Reagle, 2000)) and visualizations (such as privacy nutrition labels (Kelley, Bresee, Cranor, & Reeder, 2009)) are possible alternatives, but it is not clear if they would actually help in practice, given that companies have little incentive to improve usability and understandability. It is also not clear if end-users would bother to read these either. We discuss the topic of better displays for privacy in the section on App Store Owners and OS and Hardware Manufacturers.

Verifiable rationales would also be an area for privacy research with great potential. Currently, app developers simply request the use of personal data in in the app's source code, and do not have to specify why that data is being requested or how it is used. Having verifiable rationales could be used to automatically generate displays of why an app uses one's information (e.g. "this app uses location data for ads" versus "this app uses location data"), could facilitate the development and use of various tools to verify those rationales, could be used to elicit people's potential privacy concerns over those uses, and could also help with access control as well (e.g. "only allow location data for maps"). This approach would require some new techniques for annotating source code to express these rationales, as well as developing an agreed-upon ontology for data use.

The last three areas of research are tightly coupled with each other: *better tools for privacy, design patterns,* and *more education*. Starting with education, in past work, we conducted interviews and surveys with app developers (Balebako, Marsh, Lin, Hong, & Cranor, 2014) and found that they had very little awareness of what they should be doing with respect to privacy, as well as of privacy guidelines set forth by organizations such as the Federal Trade Commission. One way to facilitate awareness would be the development of a generally accepted set of best practices for privacy. For example, for computer security, example best practices might include using SSL for network communications, hashing stored passwords, using well-known algorithms instead of rolling one's own, and instituting access control.

However, there is not a corresponding set of best practices for privacy, in terms of tools, system architectures, design patterns, or user interfaces.

Better tools for privacy would also help advance the current state of the art. One path would be better reusable components that offer useful properties with respect to privacy. To use an analogy, databases offer a set of properties known as ACID: Atomic, Consistent, Isolated, and Durable. A developer doesn't have to know anything about these properties, or anything about the internals of how a database works, to gain significant benefits from using a database. As such, one research direction would be enumerating a set of useful and desirable properties, as well as creating tools that either have or can support those properties. One possible starting point might be looking at data flows, starting from data being sensed or collected, stored, shared, and inferred. Another possible starting point might be to understand what kinds of difficulties developers are facing when they are trying to implement privacy features, and look for opportunities for reuse or generalizability.

Another possible path would be better ways of expressing desired properties or policies, and then weaving those into an application directly. Using a web analogy, HTML (content) is separated from CSS (presentation), but integrated together by the browser to form a unified user experience. This separation of concerns makes it easier to modify things, keeps things conceptually coherent, and lets people work independently (software developers can generate HTML, while visual designers can work on CSS). Examples for privacy might include direct support for privacy in programming languages (e.g. taint tracking or other kinds of metadata tags on variables), or making it easier to ensure that a set of policies is enforced as the code is compiled (Yang, Yessenov, & Solar-Lezama, 2012). In many ways, this line of work shares much in common with Aspect-Oriented Programming (Kiczales et al., 1997).

A key challenge for this research direction is coming up with a list of desired properties that are meaningful and feasible. For example, in security, researchers often talk of confidentiality, integrity, and availability. At a high level, some desired privacy properties might include anonymity, minimal interruptions, avoiding spam, control and feedback about data flows, avoiding behavioral advertising, preventing embarrassing content from being seen by friends and family, keeping people in different spheres of life separate, and more. The Fair Information Practices might also offer some other desired properties, such as notice, consent, recourse, and so on. However, it is not clear how many of these properties can be operationalized in hardware and software. For example, one of my close friends works on Human-Robot Interaction, and one thing she often hears from lay people is that robots should obey Asimov's Three Laws of Robotics. The problem is that these laws are like the Ten Commandments from the Old Testament, in that there is a very wide technical gap between these high-level declarations and the code needed to implement them.

Another key challenge is in ensuring that these properties are held by the entire system. For example, an app might run on a smartphone, be managed by an operating system, use different hardware and software components, and send data to remote servers. Can privacy properties be guaranteed for parts or even all of this system? This is an extremely difficult and open issue for cybersecurity and software engineering in general. To some extent, Android's approach to having manifest files is one step forward here. App developers have to state up front what sensitive data their app might access, in the form of permissions. Examples include "contact list" and "location". Can an approach like this be applied to Internet of Things? Can this approach be extended for distributed systems? Are there ways of verifying or even proving that an operating system or remote service is enforcing these permissions correctly?

2.3.2. Research Directions for Third-Party Developers

With respect to *third-party developers*, we found that in past research, many smartphone apps used personal data due to third-party libraries in the app rather than custom code (Lin, Liu, Sadeh, & Hong, 2014). These third party libraries might offer services such as advertising, access to a social networking

site, or analytics of how users use an app. Interestingly, through interviews and surveys with app developers (Balebako et al., 2014), we found that many developers did not understand the privacy implications of using these libraries. It is also not easy for developers to do so either, as documentation is often sparse, there is no standard for helping developers understand what a library's actual behavior is, and today's tools offer no help.

Here is another way to think about this issue. Most app developers focus primarily on making their app work. However, these app developers also need features for advertising (to make money), analytics (to understand how users use their app), social networking (for sharing), and so on. The path of least resistance is to use libraries from third-party developers. Many developers don't intentionally mean to violate privacy, but the deck is stacked against them in terms of tools, expertise, and time.

As such, many of the research directions in the previous section would also apply here, though the users of these would be developers rather than end-users. For example, having simpler ways of conveying privacy implications of a library (beyond simple text) would be very useful for developers (e.g. "location data will be sent to our servers and used to serve up ads, the data will be retained for 12 months"). Verifiable rationales would also be helpful, as would having better ways of expressing and summarizing properties about a library. As an example of the latter, in the mid-1990s, the idea of proof-carrying code was developed, where pieces of code would come with proofs that could be quickly checked and offered certain guarantees about that code's behaviors (Necula, 1997). Is it possible to extend proof-carrying code for privacy? Is it possible that libraries could also come with proofs or other forms of verification, so that when they are integrated, we can still reason about overall properties of an app?

2.3.3. Research Directions for App Store Owners

One compelling research theme for app store owners is to develop better ways of scanning for possible privacy problems with apps. As noted by Gilbert et al, app markets are centralized repositories and can act as gatekeepers for apps (Gilbert, Chun, Cox, & Jung, 2011). If scanning techniques can be reliably developed, and if app store owners can be properly incentivized, many privacy problems could be prevented before end-users ever experience them.

There is unfortunately little documented information about what checks app store owners currently do. The conventional wisdom is that the major app markets tend to focus on malware, copyright, and quality control rather than privacy. Furthermore, while there is clearly some automated analysis being done, it is also clear that the process for vetting apps can be slow and labor intensive. For example, the web site appreviewtimes.com suggests that a typical iOS app takes about 12 days to be reviewed.

One research opportunity here is to develop better displays that can summarize the privacy behaviors of an app. Rather than making a judgment call as to what is or is not appropriate, an app store might seek to improve awareness and defer decisions to end-users. For example, one line of research my team is investigating has focused on people's expectations about an app, emphasizing those behaviors that do not match expectations. For example, most people don't expect the popular Fruit Ninja app to use location data. In contrast, most people do expect Google Maps to use location data. We view these expectations as a weak form of informed consent. For the latter case, users do not really need to be informed about the use of location data, but for the former case, they should. Figure 2 shows one sample user interface we have created, which shows people's level of surprise about an app's behavior (Lin et al., 2012). We used crowdsourcing to elicit these responses, which offer us some level of scalability. Essentially, we slice up the privacy behaviors of an app, farm out individual parts to the crowd, and then aggregate it back together.

One open question here is, what kinds of information should such a display show? Should these displays be presented as users are searching for apps, right before they install an app, as they use an app, or

after they use an app? Are there better ways of using the wisdom of crowds or crowdsourcing to help pinpoint “interesting” parts?

Another research opportunity is to develop techniques that can scan for narrow and well-defined scenarios. For example, the Child Online Privacy Protection Act (COPPA) requires that web sites and apps designed primarily for children must not collect personal information without parental consent. A system could scan apps, look for images and text that seem to be for children, and then scan the app to see if it collects data about the user. While false positives would certainly be an issue, today the process is manual and very labor intensive. Thus, anything that could help narrow things down could help.

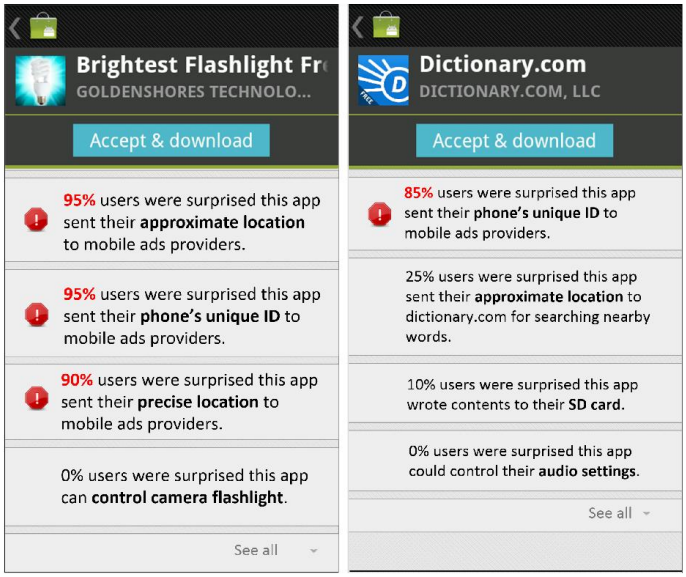

Figure 2. Example privacy displays our team developed (Lin et al., 2012). These displays modify Android's existing interface by including information about the level of surprise people had about an app's use of different permissions. We used crowdsourcing to evaluate the level of surprise.

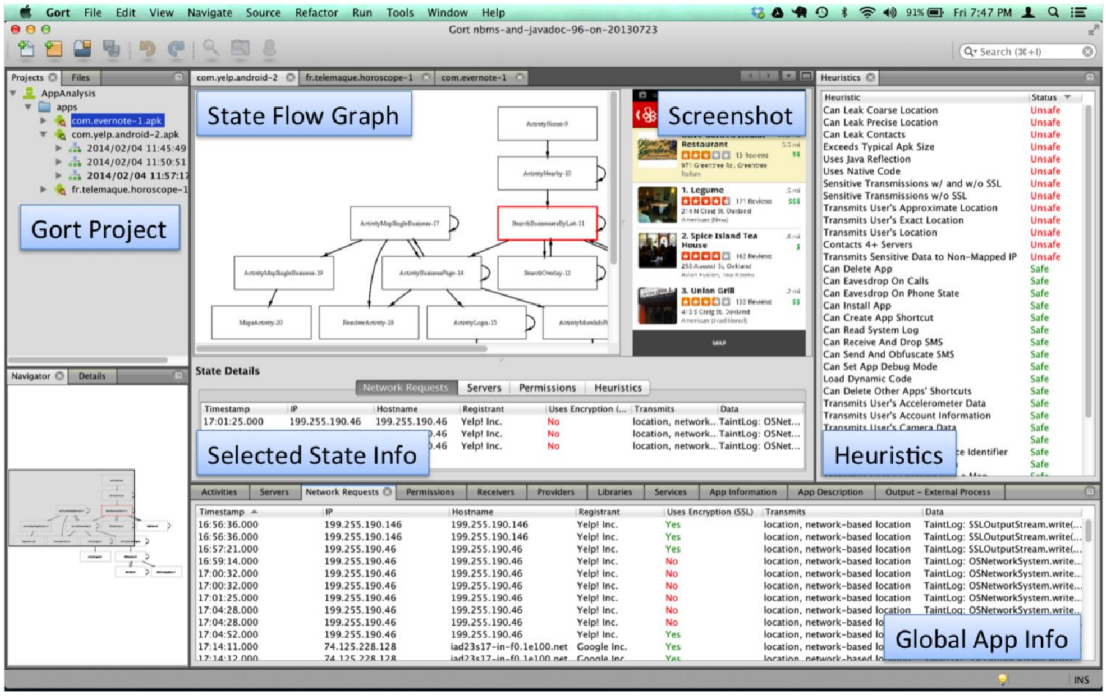

Figure 3. Gort is a tool we have developed to help non-experts analyze the behavior of smartphone apps. A core idea to Gort is to use heuristics to find possible red flags, thus helping direct analysts attention to salient issues and finding the right parts of the app to examine.

A third research opportunity is to develop better tools that can help a person analyze apps faster and more in-depth. For example, Gort (Amini, 2014) is a tool my team has been developing to help non-experts analyze the behavior of an app (see Figure 3). It first traverses all the screens of an app, traces what information is being requested and sent to where, applies heuristics to find possible red flags in an app's behavior, and presents the results to an analyst for further inspection. In ongoing work, we have been investigating how to use crowdsourcing techniques to further scale up analysis, having crowds of workers examine possible red flags and label these as safe or worth further investigation.

2.3.4. Research Directions for OS and Hardware Manufacturers

Manufacturers of operating systems and hardware can also take an active role in privacy. As a simple example, iOS rotates Bluetooth IDs to prevent people from easily tracking individuals. What other kinds of basic capabilities could OS and hardware manufacturers offer?

One research opportunity here is, can we develop better ways for developers to use the capabilities of these smart devices in a way that balances privacy and utility? For example, rather than sharing one's exact location, what if an app could simply know if a person is at "home" or "work"? Or, rather than knowing all of the people in one's contact list and full call log, what if an app could simply know that a person has "520 contacts, 20 of whom she contacts regularly"? Or, rather than getting full access to the microphone, what if an app could simply get one of "very loud", "loud", "quiet", "very quiet"?

These kinds of APIs could offer utility to app developers, while not giving access to the full amount of data. Another advantage is that very few app developers have experience with machine learning, and this approach could give app developers more flexibility and capabilities.

These APIs might also be placed into tiers as well, with the top tier being the most sensitive, and the lowest tier being the least sensitive. Grouping APIs into different tiers could help with automated analysis (e.g. "this app uses a lot of sensitive data") as well as with automated generation of interfaces summarizing what data is used and how worrisome they might be.

2.3.5. Research Directions for Governments and Third Parties

Two other entities to consider for research are government agencies (e.g. the Federal Trade Commission) and third parties interested in privacy (e.g. the EFF, CDT, or research teams at universities). People in both kinds of organizations may be interested in analyzing the behavior apps to find anomalous behavior. Tools like Gort (described briefly in Research Directions for App Store Owners) may be useful towards this end.

Another opportunity, in particular for third parties, is in scaling up the analyses and making the results public. For the analysis presented in Figure 2, it took us two weeks to crowdsource people's privacy concerns for 56 apps. Given that there are close to a million apps on the most popular app markets, we needed new ways of scaling up our analysis. We had two key insights here. First, we can infer many of the semantics of an app's behavior by examining what libraries an app uses. For example, if our analysis shows that an app uses location data only because of an advertising library, we can say that "this app uses location data for advertising." Second, rather than crowdsourcing people's concerns about each app individually, we can crowdsource people's concerns about general app behaviors. As a concrete example, we found that people were more concerned about "contact list used for advertising" than "contact list used for social networking." This approach let us create a general model of privacy concerns about app behaviors. We then crawled 750,000 Android apps, analyzed them using the techniques above, and are aiming making our results public on the web site PrivacyGrade.org (see Figure 4).

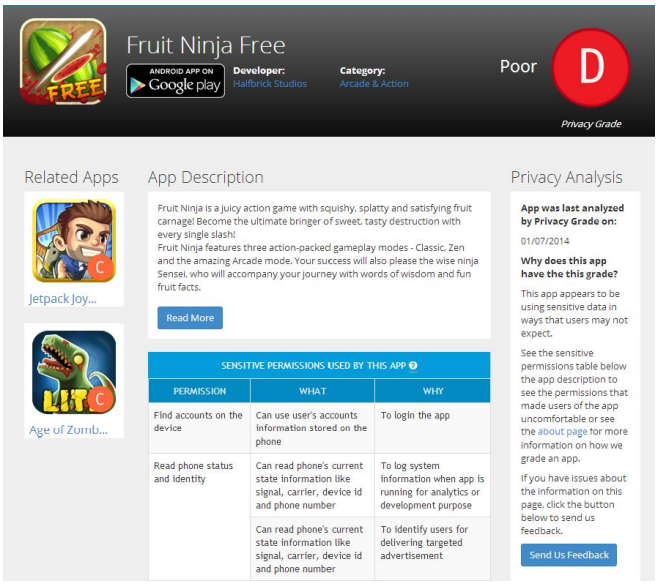

Figure 4. PrivacyGrade.org is a web site our team is creating for analyzing and summarizing the privacy behaviors of over 750,000 Android apps. PrivacyGrade combines automated analysis techniques with crowdsourcing to scale up our ability to analyze privacy concerns.

This line of work opens up at least two big questions for automated analysis of privacy. First, can we improve the range and richness of semantics that can be analyzed? Richer semantics would allow us to get more details as to what an app is doing with one's data. Our approach with analyzing libraries is fairly crude, and currently only captures simple things such as advertising, analytics, and social networking. Examples of richer semantics might include:

- This app uploads your contact list to a remote server
- This app only uses your contact list locally and does not share it with others
- This app only uses your location data to geotag photos
- This app only uses microphone data to detect how loud or quiet it is in a room

Our current technique is also limited in that it cannot analyze custom code that developers create. Improving the kinds of semantics that can be correctly inferred would greatly advance the state of the art. Note that this idea of inferring semantics is complementary to our previous idea of verifiable rationales, as described in Section 2.3.1.

Second, our technique heavily relies on crowdsourcing. Are there other ways of combining automated techniques and crowdsourcing to help with privacy in general? A major problem with automated techniques is that they cannot easily capture the context of how information is used, and context is often key to privacy. For example, location data being used by a Blackjack game is suspicious, whereas location data being used by a map application is expected.

We argue that crowdsourcing may offer the potential for significant advances for scaling up our ability to manage privacy. For example, people don't read privacy policies in general, because it is a lot of effort, there is a clear cost for unclear benefit, and because we want to use the service or app rather than read the policy. Crowdsourcing can solve all of these problems. By slicing up a privacy policy into smaller parts, each crowd worker only has to put in a little bit of effort. Crowd workers are also paid, offering a clear benefit. Crowd workers also are not using the service or app directly, so there is no strong incentive just to skip the policy. We are currently investigating these issues in a NSF SaTC grant, looking at better ways of using automated techniques and crowd workers to summarize privacy policies and terms of use.

2.3.6. Research Directions for Helping End-Users

The last entity in the app ecosystem, and arguably the most important, are end-users themselves. Given the vast diversity in demographics, skills, attitudes, and behaviors, end-users may also be the hardest group for research.

One possible line of work is developing better ways of *helping end-users express privacy preferences*. The most direct way of doing this is offering interfaces to specify rules. While I have worked on systems like this in the past (Tsai et al., 2009), I think that it puts too much burden on end-users, and also does not scale well if people have to deal with multiple smart systems. Furthermore, users might not know what they want. As Palen and Dourish (Palen & Dourish, 2003) have observed, users are often asked what their preferences are before even using a system.

A complementary approach is to develop techniques that can *cluster users who have similar preferences*. This approach sounds promising, but may be very difficult to deploy in practice, for similar reasons as above. For example, suppose a person gets a new smartphone. How would that person understand what they gain or lose by being placed in a certain cluster? What if that person's attitudes change over time, but they are still locked into a given cluster? Could they also understand why certain features are or are not working?

An alternative line of research that I believe would be more fruitful is *understanding how to design better interfaces that make people feel better about privacy*. One central issue that is not very well-understood is how the various value propositions affect people's attitudes and behaviors. For example, when ubiquitous computing was first reported to the public in the early 1990s, the researchers at PARC often talked about how the technology worked. The result was headlines like "Big Brother Pinned to Your Chest", "Orwellian Dream Come True: A Badge That Pinpoints You", and "You're Not Paranoid: They Really Are Watching You." Interestingly, as the team at PARC started talking more about what value it could offer people, how these technologies could help you rather than how they worked, the headlines became gentler, focusing more on invisible computing, and how it addressed problems that people had.

Brush et al found similar results when asking how willing people would be to share their location traces, based on various scenarios (Brush, Krumm, & Scott, 2010). When asked if they would be willing to share their data to help a city plan bus routes, 94% of their participants agreed. For weekly summaries of where they went, 59% agreed. For ads along one's intended route, 25% agreed.

Another related issue that we have touched on in previous sections is, *what is the best way of conveying privacy issues to people*? In my team's past work, we also found statistically significant differences in comfort levels for privacy displays (similar to those in Figure 2) that explained why data was being used versus those that just said that data was being used. In every case, even though the app's behavior was the same, people were more comfortable knowing that, for example, an app "uses your location data for ads" versus "uses your location data."

Another design issue, and one that raises many ethical questions, is the use of what I call *privacy placebos*. These are features that make people feel better about privacy, but offer little actual value. For example, in some of my team's past work (Hsieh, Tang, Low, & Hong, 2007), we found that people stated that they really liked logs that recorded how often someone checked their location data. However, in practice, we found that they were not used. A more vivid example of privacy placebos can be seen with Target's use of coupons, targeting pregnant women (Duhigg, 2012):

> We'd put an ad for a lawn mower next to diapers. We'd put a coupon for wineglasses next to infant clothes. That way, it looked like all the products were chosen by chance.

I believe that these kinds of designs might actually be effective in assuaging people's concerns, but also raise many questions as to whether they should be used.

Can we also gain a *better understanding of people's mental models and expectations of privacy*? For example, when I first arrived at CMU as an assistant professor, I was advising a group of undergrads for a capstone project in another faculty's research lab. It wasn't until the end of the semester that one of the students in the lab commented that there was a web cam in the corner of the lab that was broadcasting to the entire Internet. While there was nothing embarrassing that was broadcast, it did surprise me because I had no awareness of the web cam, and the lab felt like it was a closed and secure space. But the risk of embarrassment is certainly there. As sociologist Erving Goffman observed (Goffman, 1959), it is when a person thinks they are in one context but actually in another that embarrassment can occur, with his example being a man who doesn't realize that his zipper is down.

A simple solution to this problem would be to put a sign on the door, but this approach does not scale, especially as increasing numbers of these systems are deployed. Are there better ways of fostering certain mental models about privacy without requiring a person's full attention? Alternatively, are there better ways of building systems to match people's mental models? The challenge, of course, is that these systems break our everyday notions of space and time, making it hard for us to make good judgments about who will be able to see us and in what context (Grudin, 2001). Furthermore, we have not solved this problem for relatively easy cases. For example, it is not uncommon for politicians to speak candidly, unaware that they are in front of a hot mic.

A final line of research I will suggest here is *understanding how people's attitudes and behaviors change over time*. When Facebook's NewsFeed feature first came out, many people were livid about the change and demanded that Facebook remove the feature. They felt that it was a major violation of privacy. It is important to note that the NewsFeed feature did not add any new information, it simply aggregated all of one's friends status updates into a single place, making it easier to see what one's friends were up to. Over time, however, people started to see the value of NewsFeed, and I suspect that if the feature were removed, people would be very angry about it.

One can also find many other historical examples about fears of privacy that did not come true. In this seminal work on the adoption of telephones in the United States (Fischer, 1994), sociologist Claude Fischer observed that many people objected to having phones in their homes because it "permitted intrusion… by solicitors, purveyors of inferior music, eavesdropping operators, and even wire-transmitted germs." People were also forbidden to use Kodak cameras on several beaches, due to worries about being photographed in one's swimwear (Lindsay, 2004).

All of these stories are cautionary tales, and act as reminders that we as a community still have very little understanding about privacy. Privacy is a difficult topic that can be highly visceral and emotionally charged. Research that is only technical or only quantitative captures part of what everyday people actually experience in their daily lives. As such, what I would recommend is that we also encourage research that is qualitative in nature, research that can probe the intricate links between privacy, value, design, emotion, and intimacy.

3. Scenario: Social Media and Big Data

Every day, millions of pieces of social media are shared. Many researchers have used this kind of data to offer new kinds of insights about human behavior at a fidelity and scale that simply was not possible even ten years ago. This kind of data has the potential to be transformative in all aspects of life, including sustainability, transportation, healthcare, economics, politics, business, urban planning, and more.

In this section, we present our work on using social media to analyze the behavior of people in cities, and use our experiences to help frame a research agenda in this part of the privacy design space. Understanding who citizens are and what they do in cities is critical to effective planning. There are a number of methods used today, but these methods tend to be slow, labor-intensive, expensive, and lead to relatively sparse data. For example, the US census cost $13 billion in 2010 (The Economist, 2011), and is only collected once every ten years. The American Community Survey is collected annually, and cost about $170 million in 2012, but only samples around 1% of households in any given year (Griffin & Hughes, 2012).

3.1. Urban Analytics: Connecting Geotagged Social Media with Big Data

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 its residents already generate. Millions of people already use Twitter, Instagram, Foursquare, and other social media services to update their friends about where they are, communicate with friends and strangers, and record their actions. The sheer quantity of data is also tantalizing: Twitter claims that its users send over 500 million tweets daily, and Instagram claims its users share about 60 million photos per day. Some of this media is geotagged with GPS data, making it possible to start inferring people's behaviors over time.

We believe that this kind of geotagged social media data, combined with new kinds of analytics tools, will let a number of stakeholders explore how people actually use a city, in a manner that is cheap, highly scalable, and insightful. Figure 5 shows a screenshot of Livehoods, some of our early work in this space. We crawled foursquare check-ins via Twitter over several months, and used clustering algorithms to group venues together based on physical proximity (how close venues physically are) and social proximity (how similar those venues are based on the people who go there). We interviewed 27 local residents to understand their mental maps and to solicit feedback on our generated livehoods. These interviews, as well as subsequent feedback collected from people who have used our public website, indicate that the livehoods generated by our analysis resonate well with people's mental maps. They surface fine changes not captured in traditional neighborhood maps, including changes related to the mix of people frequenting certain areas at certain times of the day, the nature of their activities (e.g. shopping, entertainment, education, work), crime risk, architectural changes, and more.

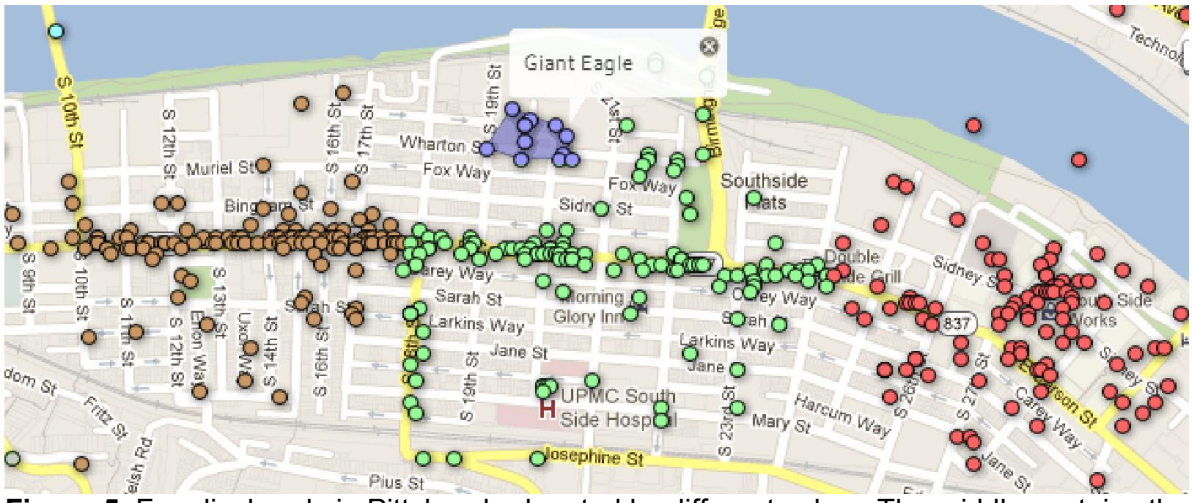

Figure 5. Four livehoods in Pittsburgh, denoted by different colors. The middle contains the only grocery store in the area. People travel from far away to shop here, since there are no others nearby. On the right is a shopping mall. The left has many bars that students frequent. Our algorithm also found a livehood in the middle. The street structure changes from short blocks on the left to long blocks in the middle. Our participants also commented that the transition area had no bars, and bar hoppers would turn around at this point.

We believe there are numerous other opportunities for using this kind of geotag social media data. One compelling direction is *location efficiency*. A neighborhood with high location efficiency is one that has good mass transit and many resources (such as schools, shops, parks) nearby. Living in a place with high location efficiency can reduce one's transportation costs, and thus potentially reduce one's carbon footprint. Currently, location efficiency is estimated based on odometer readings from participants as well as analysis of streets and venues. This process could be automated using actual geotag data that can capture where people travel, how far they travel, and what they do. We also believe that similar models could help us understand how social a neighborhood is, and possibly be a major factor in understanding how the places we live affect our physical and mental health.

Similarly, this kind of data can also help *transportation engineers*. Travel information about where people go is typically gathered using GPS and diaries from a sample of the population once every ten or twenty years (Metropolitan Council, 2000). The problem is that planning and allocation of resources is based on this coarse-grained data. A possible alternative is to use Longitudinal Employer-Household Dynamics (LEHD) data, which is collected by the US Census and is derived from data from the Bureau of Labor Statistics and other Federal data sets (US Census Bureau, n.d.). However, this data set only captures commutes, and does not capture, for example, people driving to the grocery store or people biking to the park. Geotagged social media data offers an opportunity for modeling this kind of behavior at scale, potentially filling a large gap for transportation engineers.

Geotagged social media data could also help people understand business trends. For example, who are the people who come to my store, what percentage of people have repeat visits, how often do they come, where else do they go, and how does it compare to my competitors? Similarly, this kind of data could help us understand the effects of various events on a city. What was the impact on shopping due to the snow storm this past weekend? How many more people go to bars because of the football game? How were the stores on this street impacted due to the road repair?

3.2. Privacy Risks for Social Media and Big Data

While there is a great deal of potential in using big data techniques to analyze social media, there are also a large number of privacy risks.

One obvious privacy risk is *the range of inferences that can be made*. For example, using our data set, we believe it is likely that we could infer a great deal of demographics, such as religion (e.g. people who tweet or check-in to a church or synagogue), ethnicity (e.g. ethnic grocery stores and ethnic restaurants), and relative age (e.g. students). However, we believe that inferences can go even further, using more sophisticated clustering techniques. Figure 6 shows the results of applying the LDA topic modeling algorithm to venues in New York City. By manually examining each cluster, one could assign various labels to them. Surprisingly, we found one cluster was comprised primarily of gay bars, but also included several venues that weren't gay bars. It is important to note that in some cases, *automated algorithms might correctly classify a person*, but that individual considered that information as private and did not want to disclose that to others. In other cases, *automated algorithms might incorrectly classify a person*.

In some cases, these inferences are based not on anything that the individual has done, but rather what his friends have done. Jernigan and Mistree were able to create a fairly accurate logistic regression that could predict a person's sexuality based on the stated sexuality of their friends (Jernigan & Mistree, 2009).

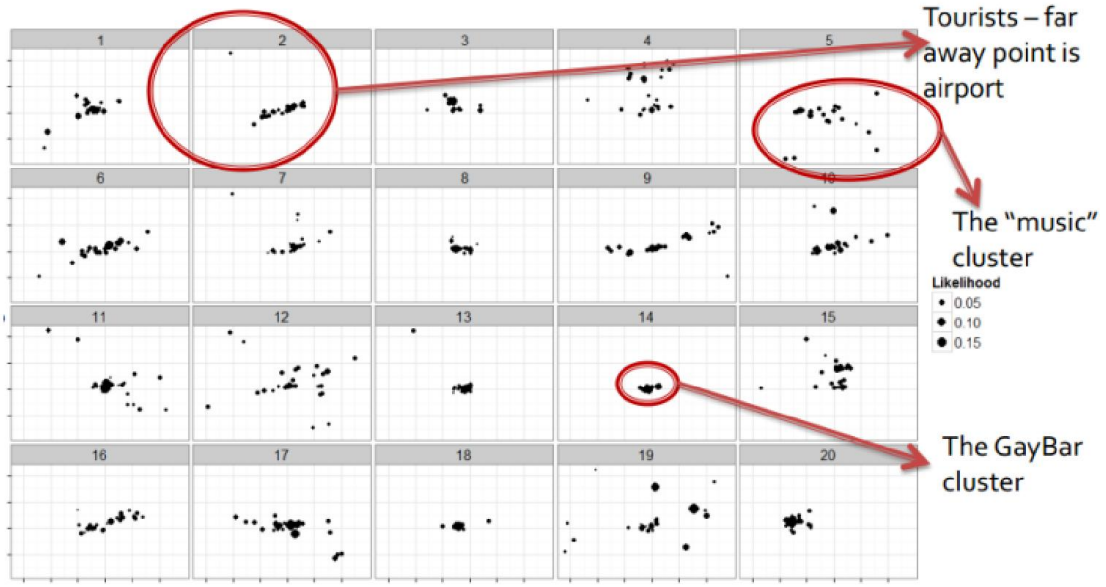

Figure 6. Visualization of clusters of venues as determined by applying the LDA topic modeling algorithm to geotagged data. Using data and algorithms like this, it may be possible to make many correct (and incorrect) inferences about individuals. The privacy risks are that, in both cases, a person might not have chosen to disclose this kind of information, and that this information might be used to discriminate against a person.

However, I would argue that the core privacy problem is not that these inferences are made, but rather that *an individual might be discriminated based on these inferences*. It is important to note that this concern is not hypothetical: these kinds of inferences are already being made, and there have already been instances of discrimination. One example of the former (correctly classifying a person) was seen in the New York Times Magazine article about how Target used algorithms to infer if a shopper was pregnant or not (Duhigg, 2012):

> [An analyst at Target] was able to identify about 25 products that… allowed him to assign each shopper a 'pregnancy prediction' score. [H]e could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Perhaps the most widely-reported example of discrimination using "risk-based algorithms" was with American Express (Cuomo, Shaylor, McGuirt, & Francescani, 2009):

> Johnson says his jaw dropped when he read one of the reasons American Express gave for lowering his credit limit: "Other customers who have used their card at establishments where you recently shopped have a poor repayment history with American Express."

It is easy to imagine that advertisers, credit card companies, insurance companies, banks, and other organizations using algorithms to create risk profiles for customers, and subtly discriminate against people all in the name of efficiency. These algorithms might appear to be objective, but might simply be a proxy for other socioeconomic characteristics. History has also shown that earlier techniques like the above, known as redlining, was used to keep blacks out of certain neighborhoods in Philadelphia. Essentially, lower income areas of Philadelphia were marked in red, and households and businesses in red zones could not get mortgages, economically crippling an entire segment of the population.

3.3. Research Opportunities for Privacy, Social Media, and Big Data

There are many research challenges for privacy in this design space of social media and big data. In the previous section on smart devices and inferencing, for the most part, the privacy issues focused on individuals, in terms of individuals choosing to install apps or understanding what the privacy problems

are. Here, the issues focus more on how companies and organizations can better manage privacy in the form of large quantities of personal data.

As such, one salient research issue is, *how can we build more secure systems to prevent breaches*? This is an area where cybersecurity and privacy clearly intersect. In recent years, there have been numerous instances of breaches affecting people's personal privacy. Perhaps the most vivid example is CelebGate, where cloud accounts of celebrities were breached and highly personal photos were made public. What kinds of mechanisms can organizations put into place to protect sensitive data? What kinds of tools could be developed to help detect breaches faster?

Another research issue is, *how should web sites with sensitive data be architected*? What kinds of design patterns are there for system architectures? What kinds of checks can be put into place to prevent loss of data? What kinds of tools can help check for possible leaks of data? For example, Digital Loss Prevention (DLP) tools can scan emails and web uploads to make sure that data with social security numbers or credit card numbers are not being sent outside of the network. Can tools like DLP be developed to help prevent other kinds of leaks of data or unwanted inferences? For instance, in some of the scenarios described above, there might be an interface that lets people query data. Can there be guarantees that people cannot re-identify specific individuals? If so, how can those guarantees be packaged into software components or tools that are easy to use and easy to deploy? In other words, as theory on algorithmic aspects of privacy advances, can everyday software engineers also use those ideas?

What kinds of design patterns are there for content and user interfaces? As an example of the latter, for Livehoods, we deliberately chose to show information about places rather than people, as the latter is very sensitive. As a counterexample (or a possible anti-pattern), Google Buzz was a social networking site that added everyone that a person had ever contacted as a friend. This led to problems like ex-spouses being added as friends. Are there other kinds of design patterns that can help guide developers of these systems?

Are there also methods and tools for quickly evaluating potential privacy challenges before deploying such systems? From a software engineering perspective, it is very expensive to build and test an entire system, only to find out that it will fail due to privacy concerns. Are there better ways of doing rapid prototyping and evaluation, to uncover potential privacy problems earlier in the design process, when they are cheaper to fix? Are there better ways of involving potential end-users so that potential privacy issues are raised and addressed earlier in the process?

What about evaluation methods for later stages of an application or web site's life cycle? For example, are there better methods and tools for conducting large-scale A/B tests in a privacy-sensitive and ethical manner? This issue was raised in dramatic form with Facebook's study on emotional contagion (Kramer, Guillory, & Hancock, 2014), with the concern being that people felt that they may have been manipulated by Facebook. Given the prevalence and power of A/B tests, we need better ways of making sure that they are done in a way that the vast majority of people find acceptable. Can crowdsourcing or smaller samples of the population be used, doing tests with just a few people and then scaling up if participants think the study is ok? Are there other ways of eliciting potential concerns from end-users?

Lastly, *are there tools that can be built to detect or raise awareness of unexpected and undesirable uses of sensitive information?* As a simple example, pleaserobme.com was built to raise people's awareness of how much geotagged information existed in status updates and photos posted on social media. As another example, the Wall Street Journal has reported on how some web sites use price discrimination to offer different prices based on web browsing behavior, likely using big data techniques (Ozimek, 2013). While price discrimination is legal, it is unpopular, and many might consider that using sensitive personal data to do so a violation of privacy.

4. Opportunity: Social Influences on Privacy

The previous two sections focused on challenges posed by new kinds of technologies, in particular smart devices and big data. In this section, we focus more on a missing opportunity. More specifically, most research on privacy focuses on the computer itself. This might include better network protocols, static analysis techniques, distributed systems, or algorithms for encryption or private information retrieval. Some research on privacy focuses on individuals. This might include better user interfaces or probing people's attitudes and behaviors towards a particular aspect of privacy.

However, one very large gap in our knowledge is understanding how groups of people or organizations manage privacy issues. For instance, past research on the human side of privacy has tended to treat people as isolated individuals rather than as social actors acting within a web of relationships and social influence. There is ample evidence that social influence can be leveraged to encourage positive behavior change. In one powerful example, Nolan et al. (Nolan, Schultz, Cialdini, Goldstein, & Griskevicius, 2008) found that telling people that many of their neighbors were saving power increased conservation far more than other non-social interventions (e.g. telling them that saving power was good for the environment, that it would benefit society, or that it would save them money). This result held in both the short and long term. Goldstein et al. (Nolan et al., 2008) found that by simply telling hotel guests that the last person who had stayed in that room opted to reuse their towels increased hotel room towel reuse rate by 28%. The interventions above are very simple but significantly changed people's behavior.

As such, one line of research is: *can similar techniques be applied to influence people's awareness, knowledge, and motivation for privacy*? By awareness, we mean being aware that a certain feature or privacy problem exists. By knowledge, we mean knowing what specific steps to do to protect oneself. By motivation, we mean caring enough to actually protect oneself. All three of these are needed for individuals to protect themselves in practice.

There are many research questions in this space. How do people talk about privacy? What beliefs do they have that are correct, and what beliefs are incorrect? How did they acquire these beliefs? What causes people to act on these beliefs? In our work, we have been using social psychology as our primary lens. For example, social proof is a form of conformity in which individuals use the behavior of others as information to determine the 'correct' or 'right' way to behave. If a person believes that everyone else in an organization is active in protecting privacy, that person is more likely to do so. Perceived similarity may also be helpful. For example, a young woman probably would not be strongly influenced if she read a story about how an old man living in a different city got hacked, but probably would be if the story were about another young woman from the same city.

There are also other lenses that can be applied to understand social influences. For example, there is a body of work studying the diffusion of innovations, or, *why do people choose to adopt or not adopt certain beliefs or technologies*? One of the major findings in this body of work is that there are five factors that significantly influence people: (a) relative advantage, or how much better the innovation is over what exists; (b) compatibility with what one is already doing or already has; (c) complexity; (d) trialability, or how easy it is to try something; and (e) observability, or how easy it is to observe others gaining benefit from an innovation. How well do these factors apply to privacy?

In particular, we believe observability may be a particularly strong point of leverage for privacy, primarily because much of privacy is currently hard to observe. For example, I do not know what privacy settings any of my friends use on Facebook, or how many of my colleagues use Tor, or how many have chosen to delete an app or refused to install an app over privacy concerns. This lack of observability makes it difficult in terms of awareness, knowledge, or motivation, to be influenced by others. To probe this idea, we worked with data scientists at Facebook to design and evaluate interventions intended to

nudge people in adopting various security features (see Figure 7). We found that, as expected, social influences were statistically significantly better both in terms of people clicking on the interventions as well as people adopting the features.

As such, a natural follow-up question is: how can we increase observability of use of privacy features or behaviors? Similarly, what kinds of features or behaviors should we try to increase observability for?

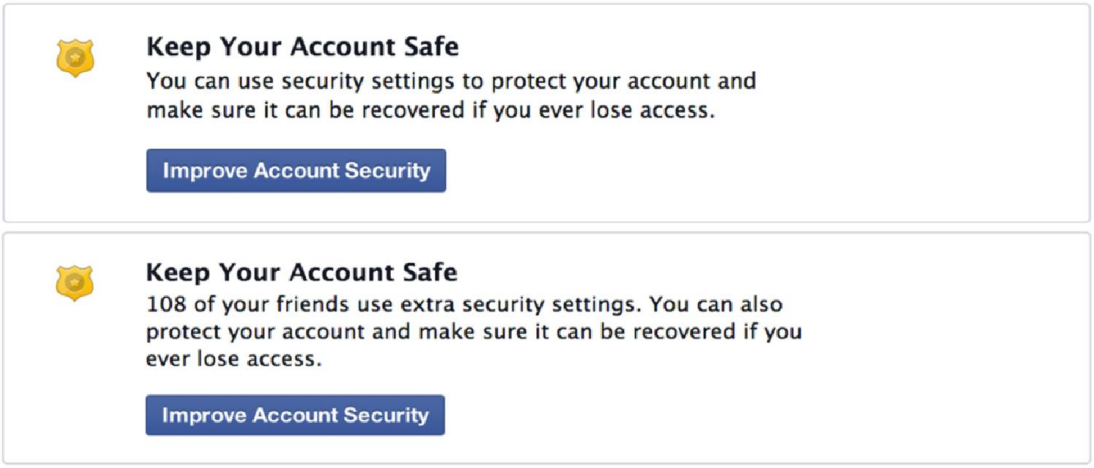

Figure 7. Two interventions that we tested on Facebook to understand the impact of social influences on individuals with respect to cybersecurity. We showed these interventions for three different kinds of security features on Facebook. In the best case, we found that about 4% more people clicked on the social interventions, and about 1% more people adopted them.

TIME SCALE OF HUMAN ACTION		
SCALE (sec)	SYSTEM	STRATUM
10^7 10^6 10^5		SOCIAL
10^4 10^3 10^2	Task Task Task	RATIONAL
10^1 10^0 10^{-1}	Unit Task Operations Deliberate Act	COGNITIVE
10^{-2} 10^{-3} 10^{-4}	Neural Circuit Neuron Organelle	BIOLOGICAL

Figure 8. Psychologist and computer scientist Allen Newell proposed different scales of time to understand different cognitive bands and different kinds of human action.

Speaking more broadly, our work on social factors of privacy is only one point in a much larger research space. Figure 8, which shows Newell's proposal of different cognitive bands and different times scales of action, offers another lens onto the issue. If we replace "biological" with "computational", the lowest

band is, for the most part, what computer scientists are focused on with respect to privacy. The next band, "cognitive", is what researchers focusing on human factors issues are investigating. However, currently, there is relatively little research at the other bands, but there are many crucial issues here with respect to privacy

For example, how do businesses work out cost-benefit for what data to collect or what privacy features to offer? How can those decision-making processes be improved? How does information about privacy guidelines and privacy policies get created and disseminated within an organization? How do people in a company handle sensitive personal information? What kinds of workflows are there? What kinds of workflows work better? What kinds of breakdowns are there? How are breakdowns handled? There are many research questions here that may be compelling for social and behavioral scientists, especially those in communications, social psychology, organizational behavior, and business.

It is important to emphasize, however, that any funding mechanisms need to be accompanied with a review panel with the right skill set. To give an analogy, in the CHI community, many researchers were unhappy with paper reviews because they were making a research contribution from one perspective (e.g. technology), but were being evaluated by researchers trained from other disciplines, with different values and methods (e.g. psychology). The upshot is that many well-qualified papers were rejected.

I would argue that NSF's approach to funding EAGER grants, combining a technical researcher with a social or behavioral researcher who have not previously worked together, is the right approach. It is relatively low effort for the researchers, and also allows NSF to make many small bets, foster new collaborations, and help grow a community that is still in its infancy.

5. Conclusions

We are currently at a crossroads, not just in computing, but in history. There is only one point in time when a global computing network will be created, and that time is now. There is only one point in time when the foundation is laid for how computation, communication, and sensing will be woven into our physical world, and that time is now. These technologies offer tremendous opportunities in terms of healthcare, safety, sustainability, education, and more. But this vision is possible only if we can find ways of addressing the privacy issues, if we can foster trust that the systems we build can respect people as individuals, if we can offer people tangible value, if we give people the right levels of control and feedback, and if these systems do what people expect them to do.

In this position paper, I've offered a lot of questions, but unfortunately do not have many answers. Instead, I'll end with one final question: *how can we create a connected world that we would all want to live in?*

References

Amini, S. (2014). *Analyzing Mobile App Privacy Using Computation and Crowdsourcing*.

Balebako, R., Marsh, A., Lin, J., Hong, J. I., & Cranor, L. F. (2014). The Privacy and Security Behaviors of Smartphone App Developers. In *Workshop on Usable Security*.

Bohn, D. (2012). iOS apps and the address book: who has your data, and how they're getting it. Retrieved from http://www.theverge.com/2012/2/14/2798008/ios-apps-and-the-address-book-what-you-need-to-know

Brush, A. J. B., Krumm, J., & Scott, J. (2010). Exploring end user preferences for location obfuscation, location-based services, and the value of location. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 95–104). Retrieved from http://doi.acm.org/10.1145/1864349.1864381

Cranor, L. F., Langheinrich, M., Marchiori, M., & Reagle, J. (2000). The Platform for Privacy Preferences 1.0 (p3p1.0) specification. W3C. Retrieved from http://www.w3.org/TR/P3P/

Cuomo, C., Shaylor, J., McGuirt, M., & Francescani, C. (2009). "GMA" Gets Answers: Some Credit Card Companies Financially Profiling Customers. Retrieved from http://abcnews.go.com/GMA/TheLaw/gma-answers-credit-card-companies-financially-profiling-customers/story?id=6747461

Doryab, A., Min, J.-K., Wiese, J., Zimmerman, J., & Hong, J. I. (2015). Detection of behavior change in people with depression. In *AAAI Workshop on Modern Artificial Intelligence for Health Analytics (MAIHA)*.

Duhigg, C. (2012). How Companies Learn Your Secrets. *New York Times Magazine*. Retrieved from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Fischer, C. (1994). *America Calling*. University of California Press.

Gilbert, P., Chun, B.-G., Cox, L. P., & Jung, J. (2011). Vision: automated security validation of mobile apps at app markets. In *Proceedings of the second international workshop on Mobile cloud computing and services - MCS '11*. New York, New York, USA: ACM Press. doi:10.1145/1999732.1999740

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York: Anchor, Doubleday.

Griffin, D., & Hughes, T. (2012). Projected 2013 Costs of a Voluntary American Community Survey. Retrieved from http://www.census.gov/acs/www/Downloads/library/2012/2012_Griffin_03.pdf

Grudin, J. (2001). Desituating Action: Digital Representation of Context, *16*(2-4).

Hachman, M. (2012). Path Uploads Your Entire iPhone Contact List By Default. Retrieved from http://www.pcmag.com/article2/0,2817,2399970,00.asp

Hsieh, G., Tang, K. P., Low, W.-Y., & Hong, J. I. (2007). Field deployment of IMBuddy: a study of privacy control and feedback mechanisms for contextual IM. *UbiComp*, 91–108. Retrieved from http://portal.acm.org/citation.cfm?id=1771592.1771598

Jernigan, C., & Mistree, B. F. T. (2009). Gaydar: Facebook friendships expose sexual orientation. Retrieved from http://firstmonday.org/article/view/2611/2302

Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009). A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*. New York, New York, USA: ACM Press. doi:10.1145/1572532.1572538

Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Lopes, C., Loingtier, J.-M., & Irwin, J. (1997). Aspect-oriented programming. In *ECOOP'97—Object-Oriented Programming* (pp. 220–242). Springer Berlin Heidelberg.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24). Retrieved from http://www.pnas.org/content/111/24/8788.full

Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., & Zhang, J. (2012). Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing. In *Ubicomp 2012*.

Lin, J., Liu, B., Sadeh, N., & Hong, J. I. (2014). Modeling Users' Mobile App Privacy Preferences: Restoring Usability in a Sea of Permission Settings. In *Symposium On Usable Privacy and Security (SOUPS)*.

Lindsay, D. (2004). The Kodak Camera Starts a Craze. Retrieved from http://www.pbs.org/wgbh/amex/eastman/peoplevents/pande13.html

Metropolitan Council. (2000). *2000 Travel Behavior Inventory: Home Interview Survey: Data and Methodology*. Retrieved from www.metrocouncil.org/planning/transportation/TBI_2000/TBI_Methodology.pdf

Min, J.-K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., & Hong, J. I. (2014). Toss “n” turn: smartphone as sleep and sleep quality detector. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI ’14* (pp. 477–486). New York, New York, USA: ACM Press. doi:10.1145/2556288.2557220

Min, J.-K., Wiese, J., Hong, J. I., & Zimmerman, J. (2013). Mining smartphone data to classify life-facets of social relationships. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW ’13* (p. 285). New York, New York, USA: ACM Press. doi:10.1145/2441776.2441810

Necula, G. C. (1997). Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL ’97* (pp. 106–119). New York, New York, USA: ACM Press. doi:10.1145/263699.263712

Nolan, J. M., Schultz, W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative Social Influence is Underdetected. *Personality and Social Psychology Bulletin*, *34*(913). Retrieved from http://www.csom.umn.edu/assets/118360.pdf

Ozimek, A. (2013, September 1). Will Big Data Bring More Price Discrimination? *Wall Street Journal*. Retrieved from http://www.forbes.com/sites/modeledbehavior/2013/09/01/will-big-data-bring-more-price-discrimination/

Palen, L., & Dourish, P. (2003). Unpacking “Privacy” for a Networked World. *CHI Letters*, *5*(1), 129–136. Retrieved from http://guir.berkeley.edu/projects/denim/denim-chi-2000.pdf

Russell, J. (2014). Chat app Telegram logs 5 million downloads in one day following WhatsApp sale. Retrieved from http://thenextweb.com/facebook/2014/02/21/whatsapp-lost-500000-users-to-telegram-but-most-others-seem-happy-to-stay/

The Economist. (2011). Costing the count. *The Economist*. Retrieved from http://www.economist.com/node/18772674

Tsai, J. Y., Kelley, P. G., Drielsma, P. H., Cranor, L. F., Hong, J. I., & Sadeh, N. (2009). Who’s viewed you?: the impact of feedback in a mobile location-sharing application. *Conference on Human Factors in Computing Systems (CHI)*, 2003–2012. Retrieved from http://portal.acm.org/citation.cfm?id=1518701.1519005

US Census Bureau. (n.d.). Longitudinal Employment-Household Dynamics. Retrieved from http://lehdmap.did.census.gov/datatools/aboutdata.html

Wiese, J., Hong, J. I., & Zimmerman, J. (2015). “You Never Call, You Never Write”: Call and SMS Logs Do Not Always Indicate Tie Strength. In *Proceedings of the 2015 conference on Computer supported cooperative work - CSCW ’15*.

Yang, J., Yessenov, K., & Solar-Lezama, A. (2012). A language for automatically enforcing privacy policies. *ACM SIGPLAN Notices*, *47*(1), 85. doi:10.1145/2103621.2103669