Towards a Conceptual Model for Provoking Privacy Speculation

Norbert Nthala
Dept. of Media & Information
Michigan State University
nthalano@msu.edu

Emilee Rader
Dept. of Media & Information
Michigan State University
emilee@msu.edu

Abstract

The proliferation of ubiquitous computing introduces several challenges to user privacy. Data from multiple sensors and users is aggregated at various scales to produce new, fine-grained inferences about people. Users of these systems are asked to consent to sharing their data without full knowledge of what data are recorded, how the data are and might be used, who has access to the data, and most importantly risks associated with sharing. Recent work has shown that provoking privacy speculation among system users, by visualizing these various aspects, improves user knowledge and enables them to make informed decisions about their data. This paper presents a conceptual model of how researchers can make inferences that provoke privacy speculation among system users and a case study applying the model.

1 Introduction

Ubiquitous computing relies on sensors and data aggregation to provide services to end users such as tracking one's healthy activities/status and driving behaviour. Despite its success, ubiquitous computing presents privacy risks [13]. People wear and carry with them sensor-based devices capable of collecting detailed information about where and when they spend their time. Widespread data collection and use of machine learning technologies makes it possible to infer new data about people that is based on or derived from other data that were collected. Raw or derived data can be privacy invasive or even be used in Internet-enabled discrimination [9].

However, evidence has shown that most people consent to sharing such data with limited or no understanding of the range of data that is collected or can be inferred, and potential privacy implications thereof [4, 7, 19]. In most cases, this is because it is hard for them to anticipate patterns that aggregation and machine learning can detect, and because platforms and systems are not required to tell people what inferences they make. Inferences are seen as "secondary or second-order data" [1], even by HIPAA (health privacy law in the U.S.) which permits free sharing of this deidentified data for research or commercial purposes[1].

Research has long sought to find ways to make users more aware of privacy risks [5, 12] or to nudge users to read privacy policies before giving consent [6, 17], all with limited success. Recent advances have revealed that users rely on mental theories regarding the types of data collected to make privacy decisions [14]. These theories allow users to put their privacy concerns into perspective and use them to identify privacy risks.

Emerging work shows that designing tools that can provoke speculation about data that systems collect and infer can help to improve the knowledge of users, their mental theories, their privacy decisions, and privacy-protective behaviours [4, 14, 16, 19]. Privacy speculation involves curiosity which prompts users to make explicit guesses and form lay theories about system behavior related to personal data collection and inferences that produce derived data. It allows users to explicitly question system behavior and search for information to guide their understanding and decisions.

The domain of *provocative design* focuses on using design as a mechanism for promoting questioning and reflection about existing beliefs and values, which can help people make more informed decisions about what products and services they should or should not use [10]. In this paper (first published at CHI 2020[2]), we adapt the provocative design approach to propose a conceptual model for how to provoke privacy speculation among users of ubiquitous computing systems. We demonstrate it's applicability in a case study. We

USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Boston, MA, USA.

[1] https://jamanetwork.com/journals/jama/fullarticle/2682916

[2] https://dl.acm.org/doi/abs/10.1145/3334480.3382815

postulate that developing a lightweight tool to give people clues about a range of possible inferences can provoke speculation about the kinds of information systems can collect and infer, and equip people to make informed privacy decisions.

2 Related Work

Previous research has focused on understanding how and what users think about their data, and how to help users become more aware of the kinds of inferences that systems can derive. These works provide evidence of the kinds of reactions people have when presented with various aspects of inferences made about them. They also present evidence of privacy speculation provoked by visualization of inferences and its potential to improve users' knowledge and decision making.

Zoonen [18] emphasized that ubiquitous computing systems often *combine and link data* from different sources to produce new inferences in service of the platform creators' goals. They argue that *personal data* (or personally identifiable data, e.g. individual location data) and *impersonal data* (data that cannot be linked to an individual person and are used for surveillance and control purposes, e.g. traffic data) can be used independently or combined for *service provision* or for *surveillance*. She described four areas of privacy-related concerns about data collection and inferences ranging from *hardly any* (impersonal data, surveillance) to *a lot* (personal data, surveillance purpose).

However, the framework she proposed is hypothetical and does not consider conflicting scenarios; for example, where users consider tracking of browsing important for personalization, but also consider it to be privacy invasive [19].

Rader and Slaker [14] qualitatively investigated folk theories about data collected by activity trackers and found that users conceptualized three types of data: *entered* by the user, directly *measured* by the tracker, and *calculated* from other data the tracker had collected. Users identified relationships across data by *visually watching* simultaneous changes of information in visualizations presented by the user interface. These connections, as well as perceived inaccuracies, encouraged users to *speculate* about how the tracker produced the data underlying the visualizations. The study revealed that users did not consider the fact that the data were estimates or inferences, which undermined their ability to speculate and reason about other possible uses of the raw data outside the context of activity tracking.

Weinshel et al. [19] investigated how transparency about online tracking impacts users' knowledge, perceptions, and attitudes. They developed a browser extension that collected data about users' browsing behaviour, and displayed detailed longitudinal and inference-level information back to the users. The study found that participants were *surprised* by the amount of tracking displayed, details of data collected by trackers, and inferences made. By visualizing tracking, participants identified a lot more *new information* than they *already knew*. In a post-usage survey, the participants reported increased understanding of tracking.

While none of these studies explicitly focused on provoking privacy speculation, they did present evidence of it in users' reactions to deviations from their expectations around technology. These studies provide a solid reference for the way in which people categorize data types, and how they develop mental models of how ubiquitous technology works. Our analysis of the literature revealed four categories of reactions to visualizations of data collected, inferences made from the data, and how the inferences were made.

1. **Expected:** Obvious output from a given service.
2. **Unexpected, but not surprising:** New or more information in addition to expected output.
3. **Surprising, but OK:** Not expected, but considered not privacy intrusive.
4. **Surprising and shocking:** Not expected and considered to be privacy intrusive.

Visualizations that were either *Surprising, but OK* or *Surprising and shocking* appeared to provoke more and in-depth speculations and led users to reconsider their privacy protective behaviours (e.g. see [19]). We use this work as a foundation for exploring ways to provoke privacy speculation among users of sensor-based technologies.

3 Approaches for Making Inferences

We argue that there are three broad approaches for making inferences from raw data that designers must consider when thinking about possible uses of data: combining data types for a single user, comparing data across people, and integrating different sources of data. We briefly discuss each of the three below.

3.1 Combining data types within a user

Companies can infer things about a user by combining various data types from that user, creating dependencies between data types. These are by far the most common inferences that companies show to users of ubiquitous devices. For instance, activity trackers make inferences about the amount of calories burned [14].

3.2 Comparison across people

Data from various sensors and users can be compared to produce inferences that rank or categorize users in relation to others. One example of this in the financial industry is credit scores.

3.3 Integration across sources

Data from disparate sources can be integrated to produce new inferences [1], e.g. data about daily weather conditions in a city can be combined and correlated with data about daily car accidents in the city to make inferences about how weather conditions affect road safety.

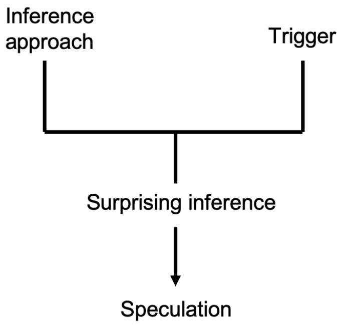

Figure 1: Provoking speculation

4 Provoking Privacy Speculation

To provoke speculation, designers must choose an approach (from the three discussed above) from which they prefer to make a provocative inference. Designers must also choose a *trigger* for the intended speculation. From our analysis of the empirical studies [2, 4, 8, 11, 14], we identified three categories of triggers (aspects of raw data) which can be varied in the chosen approach (i.e. within a user, across people, or across data sources) to make inferences which provoke privacy speculation: (1) time, (2) kinds and amount of data collected, and (3) data management. As discussed earlier, *surprising, but OK* and *surprising and shocking* inferences spur more and indepth speculations. For the purposes of this research, we define *surprise* as a reaction that may be followed by an emotional response that occurs when one encounters an event or piece of information that violates an existing belief, formulated from previously obtained information, about the relationship between elements of a system or the purpose of such a relationship.

We discuss the three categories of triggers and how they can be applied in the three approaches of making inferences to provoke privacy speculation.

4.1 Time

Frequency of recording data: The frequency at which data about a user is recorded plays an important role in their perception of privacy risks [3, 4, 8] Evidence shows that when a user of a tracker knows that a device records their location (latitudes and longitudes) in real time, it makes them feel "too watched" (and "too listened to" for voice-enabled devices) [8]. Showing users inferences of such data provokes thoughts and discussions of how they would want the device to operate, whether they wish to continue using the device, possible implications or risks that can arise from using such data (e.g. undermining one's physical security) and how they can be more in control of their data.

Within a user, provocative inferences can be made from various combinations of recorded data that seem to be privacy invasive and/or surprising. For instance, continuously recorded location data can be used to infer how many, when, and to where a user made overnight trips away from home. This can make some people *uncomfortable*, e.g. where someone has a "really controlling partner" [4].

Retention period of recorded data: Users develop perceptions of how long their data is and should be kept based on the service used [2, 8]. Creating inferences which depict various aspects of longitudinal tracking can provoke speculation, particularly because it will more likely lean towards surveillance. For instance, longitudinal recording of location data can be used to infer one's home or place of work, which provoked privacy concerns and discussions in [8].

In comparison across people, longitudinal data can be used to make inferences that compare and rank users of a system in relation to each other. Being ranked according to other aspects of their behavior is likely unexpected output and would provoke thoughts about why such comparisons are done, how they are done, what such comparisons are used for, and potential benefits and risks.

4.2 Kinds and amount of data recorded

Users (and researchers) of ubiquitous technology discuss data privacy in various ways depending on context, e.g. voice/audio, video, text [8] or location, acceleration, steps [3, 4], or entered, measured, and calculated [14]. Some data types are generally considered to be more intrusive than others, e.g. "they simply found any audio recording to be too intrusive because someone could determine if you were with someone, or the number of people in a conversation, or even the emotional tone" [11]. The context in which data is recorded and inferences are made reflects the kinds of reactions expected from users.

For instance, users of accelerometers and barometers found the use of GPS-enabled devices to be privacy intrusive and surveillance-related [4], while runners who want to have maps of their running routes to plan future workouts are more willing to allow collection of raw GPS data.

We argue that inferences made from recorded data (or data pulled from other sources) that cannot be easily or directly linked to visual output of a system can provoke speculation around *why and how a sensor-based tracker is able to do this, and what other things it tracks about the user.* Such inferences can range from within a user to across data sources. For instance, location and time data can be combined and correlated with camera footage from other sources to infer a location where someone was at a particular time, what s/he looks like, clothes worn at the time, or the type of car s/he drives. Informing a user about such inferences which are far from a normal and expected service can be surprising and likely provoke speculation [4].

4.3 Data management

Who has access to the data? Most companies generally share information about users with *third parties*, without re-

	Combining data within a user	*Comparisons across people*	*Integration across sources*
Frequency of recording	Places that a user visited at various times of a day, number of times that the user visited each place	Speeding behaviour against others, number of visits a user made to a particular location against others	Name of and directions to a place where a user currently is.
Retention period	User has moved to a new home, user's work schedule	Time spent driving weekly/monthly/annually against others	User has kids if she visits any K-12 schools or day-cares.
Kinds and amount of data	Number of smart devices a user has, number of links clicked in email	Number of other Automatic users who visited similar websites as a user	User was near or at a crime scene at a specific time.

Table 1: Example provocative inferences for Automatic

vealing who these are. However, users consider sharing their data (confidential or not) without seeking their permission to be *misuse of recorded data* [2]. They expect their service providers to inform them about who wants access to their data, and await their approval before sharing it. It is commonly accepted knowledge that people have different perceptions about various companies, and can choose to allow their data to be shared with some, but not others. Revealing to users the names of other companies that have access to raw or derived data can provoke speculation because it violates their mental models and expectations. For instance, in the example about combining location and time data with camera footage, disclosing who has access to such kinds of information, e.g. the police, can raise concerns about "being watched" [4]. Similarly, users can find it surprising that their TV sends data to a streaming service they did not subscribe to; just as their rice cooker sends data to a software company [15].

Where is the data stored? In some sensor-based services (e.g., activity tracking), people use devices for which interaction and/or feedback are provided via a mobile app. Some users perceive and prefer their data (raw or derived) to be stored on their mobile phone or on the tracker, and not on a remote server [4]. Such considerations depend on the perceived sensitivity of the data and the associated risk, e.g. requiring that location data should not to be stored (indefinitely) on a remote server, presumably for fears of an unauthorized user accessing the data and knowing where someone is [4].

5 Case Study: Provoking Speculation

We apply the model discussed in the previous section (see Figure 1) to generate a range of possible provocative inferences that can be made from data that is collected by the Automatic adapter (https://automatic.com/) . The adapter is a sensor device which is plugged into a car and tracks location and vehicle statistics (e.g. hard braking, acceleration, speed, ignition status). The tracker has a mobile and web interface where users can view their trip statistics and status of their vehicle. In its privacy policy (https://automatic.com/legal), Automatic states that it can share user data with third parties, but can also aggregate multiple datasets to make further inferences.

We make inferences about an individual, across users of Automatic, and from integrating the data from Automatic with data from other freely available data sources. We consider, among others, Automatic's ability to collect different data types at various frequencies and the platform's ability to retain data over a long period.

We aim for *surprising* inferences (i.e. which deviate more from Automatic's services) in order to provoke speculation. Table 1 shows examples of the inferences. We describe how we applied the model to generate two inferences from the table.

User has moved to a new home: longitudinal data (retention period) from Automatic can be used to infer a user's home by correlating a location where a trip routinely starts and ends with time, morning/evening (data within a user). If a user relocates, such patterns would be reflected in the data. Informing the user about where s/he had lived, when she moved, and where she is living can be associated with "being too watched" and can provoke speculation about how much the application knows about the user.

User was near or at a crime scene at a particular time: crime APIs, e.g. CrimeOmeter, provide detailed, near real-time crime data for various locations. Using a user's location coordinates from Automatic's adapter, we can query the crime API to retrieve data about when and where the user was close to a crime scene, including details of the crime (integration across sources). This can *instill fear* in the user and provoke speculation about how and why Automatic knows this.

The model discussed above aims to guide work to help users of ubiquitous technology make informed privacy decisions. This is a step forward in finding solutions to "users' inability to see a technology [which] makes it difficult for them to understand how it might affect their privacy" [1]. This will help in exploring issues around completeness and comprehensiveness of privacy policies and finding better ways to allow users have more control over their data and be able to negotiate allowable use of the data.

Building on this model, our future work will investigate ways to represent and measure uses of derived data so that users can monitor the resource, coordinate about uses of derived data, and maintain accountability.

Acknowledgments

We thank Joe Freedman and Steven Cauthen for their input on early prototypes, Chris Fennell and Megan Knittel for being early prototype test users, and the BITLab @ MSU research group for feedback on this project. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1524296.

References

[1] R. Beckwith. Designing for ubiquity: the perception of privacy. *IEEE Pervasive Computing*, 2(2):40–46, April 2003.

[2] Giovanni Iachello, Khai N. Truong, Gregory D. Abowd, Gillian R. Hayes, and Molly Stevens. Prototyping and sampling experience to evaluate ubiquitous computing privacy in the real world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 1009–1018, New York, NY, USA, 2006. ACM.

[3] I. A. Junglas and C. Spitzmuller. A research model for studying privacy concerns pertaining to location-based services. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 180b–180b, Jan 2005.

[4] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. Exploring privacy concerns about personal sensing. In Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. J. Bernheim Brush, and Yoshito Tobe, editors, *Pervasive Computing*, pages 176–183, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[5] Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, pages 279–284, New York, NY, USA, 2013. ACM.

[6] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

[7] Vivian Genaro Motti and Kelly Caine. Users' privacy concerns about wearables. In Michael Brenner, Nicolas Christin, Benjamin Johnson, and Kurt Rohloff, editors, *Financial Cryptography and Data Security*, pages 231–244, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[8] David H. Nguyen, Alfred Kobsa, and Gillian R. Hayes. An empirical investigation of concerns of everyday tracking and recording technologies. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 182–191, New York, NY, USA, 2008. ACM.

[9] Andrew Odlyzko. Privacy, economics, and price discrimination on the internet. In *Proceedings of the 5th International Conference on Electronic Commerce*, ICEC '03, pages 355–366, New York, NY, USA, 2003. ACM.

[10] Deger Ozkaramanli, Peter MA Desmet, Peter Lloyd, and Erik Bohemia. Provocative design for unprovocative designers: Strategies for triggering personal dilemmas. In *Proceedings of Design Research Society 50th Anniversary Conference*, pages 1–16, 2016.

[11] Scott R Peppet. Regulating the internet of things: first steps toward managing discrimination, privacy, security and consent. *Tex. L. Rev.*, 93:85, 2014.

[12] Stefanie Pötzsch. Privacy awareness: A means to solve the privacy paradox? In Vashek Matyáš, Simone Fischer-Hübner, Daniel Cvrček, and Petr Švenda, editors, *The Future of Identity in the Information Society*, pages 226–236, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[13] Blaine A. Price, Karim Adam, and Bashar Nuseibeh. Keeping ubiquitous computing to yourself: A practical model for user control of privacy. *International Journal of Human-Computer Studies*, 63(1):228 – 253, 2005.

[14] Emilee Rader and Janine Slaker. The importance of visibility for folk theories of sensor data. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 257–270, Santa Clara, CA, July 2017. USENIX Association.

[15] Jingjing Ren, Daniel J. Dubois, David Choffnes, Anna Maria Mandalari, Roman Kolcun, and Hamed Haddadi. Information exposure from consumer iot devices: A multidimensional, network-informed measurement approach. In *Proceedings of the Internet Measurement Conference*, IMC '19, pages 267–279, New York, NY, USA, 2019. Association for Computing Machinery.

[16] Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2347–2356, New York, NY, USA, 2014. Association for Computing Machinery.

[17] Nili Steinfeld. "i agree to the terms and conditions": (how) do users read privacy policies online? an eye-tracking experiment. *Computers in Human Behavior*, 55:992 – 1000, 2016.

[18] Liesbet van Zoonen. Privacy concerns in smart cities. *Government Information Quarterly*, 33(3):472 – 480, 2016.

[19] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the places you've been! user reactions to longitudinal transparency about third-party web tracking and inferencing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 149–166, New York, NY, USA, 2019. ACM.