Location Data and Privacy: A Framework for Analysis

Aristea M. Zafeiropoulou, Kieron O'Hara, David E. Millard & Craig Webber*

Web and Internet Science Group
Electronics and Computer Science
University of Southampton
Highfield
Southampton SO17 1BJ
United Kingdom
{az4g09,kmo,dem}@ecs.soton.ac.uk

**School of Social Sciences*
University of Southampton
Highfield
Southampton SO17 1BJ
United Kingdom

c.webber@soton.ac.uk

Abstract: Innovative services have exploited data about users' physical location, sometimes but not always explicitly with their consent. As new applications that reveal users' location data appear on the Web it essential to focus on the privacy implications, in particular with respect to inferences about context. This paper focuses on the understanding of location and contextual privacy by developing a framework for analysis, which is applied to existing systems that exploit location data. The analysis highlights the primal role of location in linking and inferring contextual data, but also how these inferences can extend to non-contextual data.

Introduction

The exposure of location data in Web applications is an emerging issue, as innovative services appear which take users' location from sensors on increasingly popular devices as input. Social networking sites such as Facebook or Foursquare have been particularly important in hosting services that add value to their social networks.

However, to provide these services service providers have to harvest location data, which raises the question of privacy. Location is a powerful dimension for understanding a person's life, both in terms of patterns of behaviour, and in terms of what extra contextual or even non-contextual data may be inferred by these applications. Location privacy is a matter not only for location data, but also other types of data such as temporal and activity data that may be inferred from the location data.

Location privacy has a number of aspects that mark it out from other types of digital privacy issues:

- Contextual data inferred through location can support surveillance techniques such as tracking the traces of individuals, their activities and so on. This is very powerful for building psychological and behavioural profiles.
- Location data deals with an individual's real-time location (or, rather, the real-time location of an individual's device). The individual is very expressly targeted.
- Location data are not valuable for the data subject to build a persona. In general, location is published for purpose-driven reasons, and its contribution to informational self-determination is relatively small.

Given these issues, it is essential to shed light on the various effects on privacy which are caused by location data. To that end, this paper provides a framework for analysis

for location data, to provide a route to understanding the potential privacy implications that may arise from their exposure. The framework is tested on a sample of recent systems taken from the research literature.

Data properties

We need to take into account those properties that can highlight the implications of exposing location and contextual information, determining which information is particularly rich in identification and profiling possibilities. An initial set of properties was defined based on background literature and later refined with a small set of research papers retrieved from the Proceedings of the Mobile HCI conferences. Broadly speaking, it is relevant to determine not only what data are being harvested, but also how they are used, who has access, and so on.

Data degree

Location and contextual data can be classified into different degrees of data based on the complexity of the inference that produced them.

- 1st Degree. Data that are explicitly provided. For instance, a smartphone will explicitly declare its user's geographical location.
- 2nd Degree. Data that are inferred directly, such as co-location between two users.
- 3rd Degree. Data that require inference with complex heuristics. This may require retrieval of data from a range of users.

Of course, there will be borderline cases, and a different characterisation may be useful. Our aim is to provide a simple measure of the intuitiveness of inferred data.

The different degrees can be illustrated via the following scenario.

Alice is a regular smartphone user and allows her phone to update her location information through a location-based application. Mary, a friend of Alice, has the same functionality set in her own smartphone.

A third party collects and stores the tracks of users of this specific app, and is therefore aware of the movements of Alice and Mary. The app also identifies and calculates the number of co-locations between the users. If the number of co-locations between any two users is significant, it infers that they are socially related. Alice and Mary are often in the same location, and a connection is inferred. Furthermore, via analysis of the locations of a large number of people, the app can determine certain geographical 'hotspots' where many congregate, and which might be of interest to Alice and Mary.

These contextual elements can be classified into different degrees based on their inferential complexity. Location information is explicitly declared and is consequently of first degree. Data inferred from location data (of one or a small number of users), such as activity and co-location, are second degree. The third degree makes use of even more complex heuristics, such as combining Alice's data with the data from thousands of other users, in this case to identify geographical hotspots.

Personally identifiable data

Personally identifiable information (PII) includes any piece of data that identifies uniquely a particular person (given the caveat that the location data we discuss here relates strictly to a device rather than a person). Location data can potentially be PII, especially in combination with other information. A simple example could be the identification that a specific GPS coordinate is a person's current location.

- Directly Identifiable Data. An individual is explicitly related to a piece of information.
- Indirectly Identifiable Data. It can be inferred that an individual is related to a piece of information.
- Heuristically Identifiable Data. It can heuristically be inferred with some probability that an individual is related to a piece of information.
- Non Identifiable Data. A piece of information is not related to any individual.

User consent

This property concerns whether the individual is asked to provide consent before location data is retrieved or published. User consent may be given not only explicitly but also implicitly, in cases where the user is not directly asked to give out their data, but the data are published with their full knowledge and the user does not take any action against it. User consent is only legally required for data that are PII.

Data quality

The better the quality of data, the more accurate the inferences made upon it will be, all things being equal, and therefore the greater the threat to privacy. The quality of the data can be expressed based on a set of three different characteristics; accuracy, completeness and timeliness:

- Accurate Data. The data is precise.
- Complete Data. The data is complete in the sense that are no values missing from it or there is nothing to be added to it.
- Timely Data. The data is current and not out-of-date.

Data access

It is important to ask not only what data are being gathered, but who can use it? As shown in Figure 1 there are a number of different entities that may have access to the data. In addition to this, they might have different types of access (read/edit/disseminate). We assume that the system has always access to the data. The sample is adequately described by a hierarchy, as shown in the figure, but of course it may be that a more complex structure is appropriate for a wider sample (e.g. a system might give access to the data to itself and third party systems, but not to the user or his or her contacts).

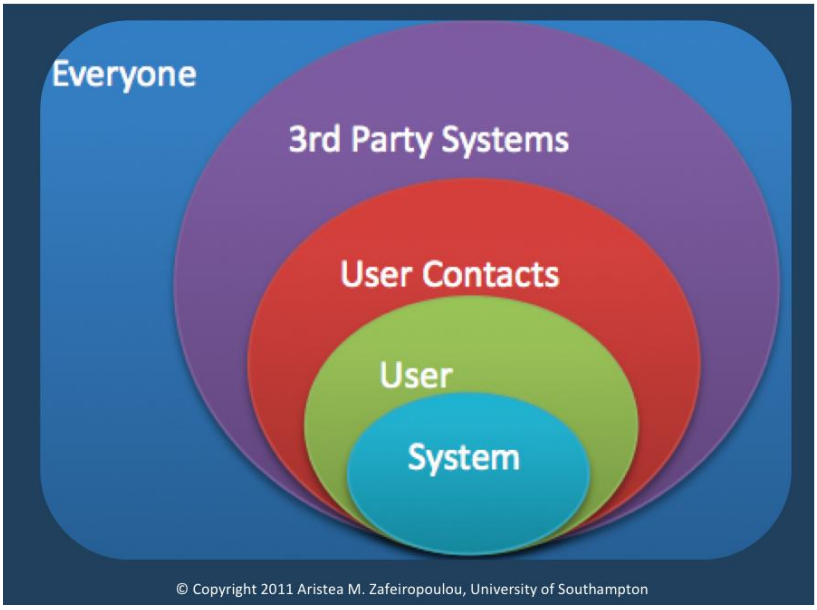

Figure 1: Who has access to data?

Data source

Data can come from different sources, not only the user. Inferences can be made from data provided by the system, or by contacts of the user and 3rd parties.

A sample of location-based systems

In order to assess the trends in location privacy, the systems presented in three years (2008, 2009, 2010) of the Ubiquitous Computing and Mobile Human Computer Interaction conferences were analysed; these conferences are the premier gatherings for ubiquitous computing, and so it was assumed that they would be 'ahead of the curve' and good predictors of trends.

Initially, any systems that made use of location data were selected. The selection was narrowed according to the following criteria.

- The paper contains a commercial or research system.
- The paper focuses on location and contextual data.
- The data are retrieved from real usage of the system (either in the context of an experiment or actual usage).
- The retrieved data refer to people.
- Only full and short papers are included in the analysis.

The result was a sample of systems described in 30 papers. Space precludes a full listing, but see the Appendix for an abbreviated list.

After selecting the systems to be included, the exposed location and contextual data in each system were identified. A category of data was selected for further analysis if it had either of the following characteristics:

- Explicitly discussed location information
- Explicitly discussed contextual information

Only data that were explicitly discussed in a paper were included in the analysis. Data that were not explicitly discussed, but might be inferred from the discussed, were not included.

Analysis of the sample

After the selection of data categories in all the selected papers, each category was analysed in terms the properties set out above. The analysis of the sample gives some indication of how the data used by the services described in the papers might impinge upon privacy.

The first question deals with the types of systems that make use of location data. As shown in Figure 2, the range is quite wide.

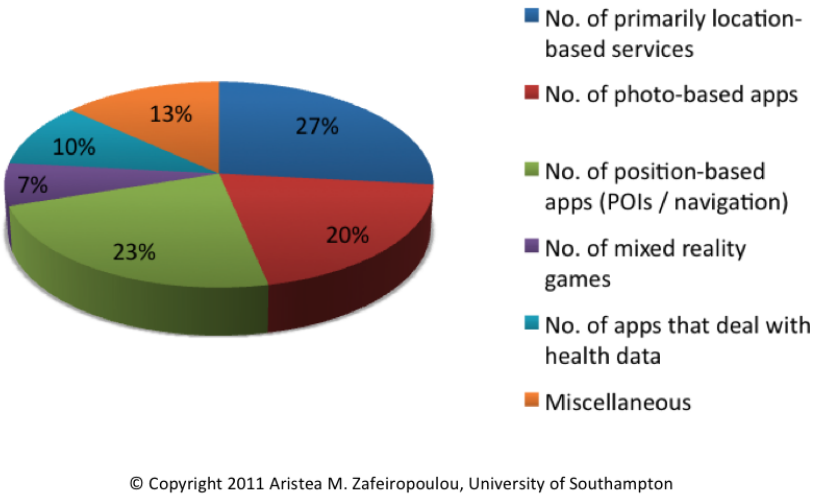

Figure 2: Types of system identified in the analysis

Half of the systems do make complex heuristic inferences of the 3rd degree (15 out of 30), as shown in Figure 3. We can go further and identify systems that make 3rd degree inferences where the inferred data go beyond the context of the person's location, i.e. where the inferred data have no semantic relation with the 1st degree location data. Out of 15 systems that make 3rd degree inferences, 5 use location data to make inferences beyond location.

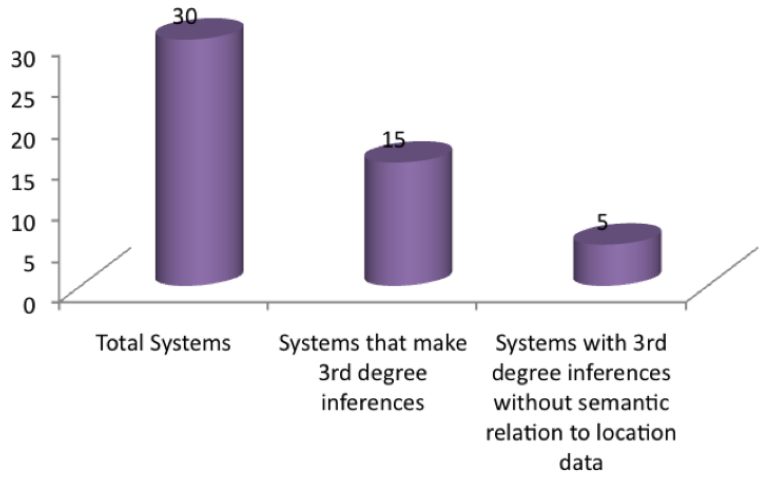

Figure 3: Systems that make 3rd degree inferences

The different properties can be profitably investigated along the dimension of the degree of inference. For each of the relevant properties, the sample was analysed to see whether the systems treated inferred data differently from data presented to the system.

Personal Identifiable Data. Regardless of the level of inference the majority of the data was directly personal identifiable (Table 1). That means that most of the information could easily be associated with a specific individual.

	1st Degree	2nd Degree	3rd Degree
Directly	**67%**	**56%**	**43%**
Indirectly	11%	12%	7%
Heuristically	3%	6%	4%
Non Identifiable	18%	24%	39%

Table 1: Degree-based analysis of personally identifiable data

User Consent. With regards to 1st degree data most systems expected users themselves to expose their data (e.g. user location), so it was taken for granted that the user consent was given. However, when it came to 2nd and 3rd degree data there was not sufficient information to suggest that the consent of the user was requested (Table 2).

	1st Degree	2nd Degree	3rd Degree
Explicit	**69%**	21%	18%
Implicit	15%	21%	18%
No Information	16%	**44%**	**39%**

Table 2: Degree-based analysis of user consent

Data Quality. Most of the systems were provided with high quality 1st degree data in terms of completeness, timeliness and accuracy. However, there was relatively little information with regards to the quality of 2nd and 3rd degree data (Table 3).

	1st Degree	**2nd Degree**	**3rd Degree**
Good Quality(Accurate/ Complete/ Timely)	**55%**	18%	7%
Low Quality	9%	24%	43%
No Information	36%	**59%**	**50%**

Table 3: Degree-based analysis of data quality

Data Access. As Table 4 shows, regardless the degree of the data the majority of the data in these systems were available to the user who they refer to. Nevertheless, in most cases the access rights of the user were not clear in the papers. It is worth pointing out that in most of these systems there was little clear indication about 3rd party systems involved.

	System	**System + User**	**System + User + User Contacts**	**Everyone**	**Unknown**
1st Degree	4%	**63%**	22%	5%	6%
2nd Degree	15%	**67%**			18%
3rd Degree	36%	**46%**			18%

Table 4: Degree-based analysis of data access

Data Source. As expected the 1st degree data were mostly user-generated, whereas the 2nd and 3rd degree data were generated by the system (Table 5).

	System	**User**	**Unknown**
1st Degree	47%	**53%**	
2nd Degree	**79%**	21%	
3rd Degree	**93%**	3%	4%

Table 5: Degree-based analysis of data source

The analysis also focused on the relation between location and the type of data that systems publish. As shown in Figure 4, in more than half of the systems location plays a primary role on the type of data that are published about a user. 27% of the systems are affected by location data but not primarily, whereas 17% of the systems use location data only as metadata.

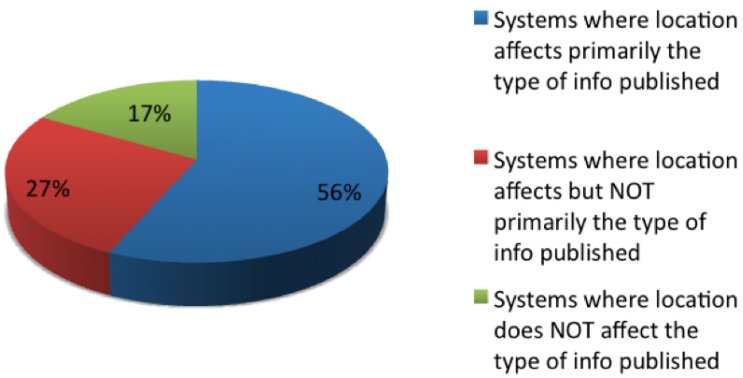

Figure 4: The role of location in the sample

Discussion

These results highlight the power of location data as a starting point for aggregating and inferring data. For instance, one third of the inferred data have no semantic relation to location data – they are inferences about some totally new aspect of the users' behaviour. Figure 4 confirms the role of location data as a catalyst for linking data across the Web.

As the results revealed, the majority of the data in the analysed systems could be characterised as PII. Although in the majority of the systems the 1st degree data were exposed with the individual's explicit consent, there was not sufficient information to suggest that the consent of the users was requested before making 2nd and 3rd degree inferences on their data. This may of course not be sinister in many cases, but only a minority of the systems are explicitly concerned with privacy, and contain privacy-protecting mechanisms. As we move out from 1st degree data, the user may be losing control in at least some cases.

It has long been understood that exposing location data online can present a number of privacy-related risks, but the framework for analysis given above allows a more targeted exploration of the relations between the complex issues of consent, inference and access. Who can see the data? How is it used? Where is it from? Answers to these questions will strongly affect the assessment and management of risk.

It is worth pointing out that the majority of these systems were research systems and not commercial. The data were collected in many cases in the context of an experiment instead of real usage of the systems. In addition, in many cases there was not sufficient indication of data quality, user consent or even whether the system took any actions to anonymise the collected data. Nevertheless, silence on these topics implies that the imperative to build a functioning system outweighs the methodology of privacy by design.

The work reported in this brief paper is intended to show that location privacy is not a homogeneous phenomenon. Location data can be used in more or less sophisticated ways, and affects privacy (and the awareness of privacy breaches) differently depending on how it is acquired, who has access to it, and so on. The small sample shows that many systems treat data differently depending on how it was acquired. The growing popularity of location-based services shows the need for comprehensive

analysis and description of data usage patterns. We do not claim that the framework presented in this paper is definitive, or provides all the relevant categories; however, it does indicate that a simple set of categories, or crude mechanisms such as consent tickboxes, are unlikely to allow users to manage their privacy and consent in a nuanced and sensitive manner.

Appendix: the papers used in the sample

From MobileHCI '08:

Bamford et al

Clawson et al

Froelich et al

Hang et al

Herbst et al

Hutter et al

Melto et al

Preuveneers et al

Robinson et al

Yoon et al

You et al

From MobileHCI '09:

Ankolekar et al

Cherubini et al

Harper & Taylor

Robinson et al

Von Watzdorf & Michahelles

From MobileHCI '10:

Brush et al

Cui et al

Sohn et al

Wagner et al

From UbiComp '08

Stewart et al

Zheng et al

From UbiComp '09

Lim & Dey

From UbiComp '10

Cranshaw et al

Dearman et al

Lin et al

Lovett et al

Madan et al

Tang et al

Toch et al