I Know Where You Live: Inferring Details of People's Lives by Visualizing Publicly Shared Location Data

Ilaria Liccardi[*][†]
ilaria@csail.mit.edu

Alfie Abdul-Rahman[*]
alfie.abdulrahman@oerc.ox.ac.uk

Min Chen[*]
min.chen@oerc.ox.ac.uk

[*]Oxford e-Research Centre
University of Oxford, UK

[†]MIT CSAIL
Cambridge, MA, USA

ABSTRACT

This research measures human performance in inferring the functional types (i.e., *home*, *work*, *leisure* and *transport*) of locations in geo-location data using different visual representations of the data (textual, static and animated visualizations) along with different amounts of data (1, 3 or 5 day(s)).

We first collected real life geo-location data from tweets. We then asked the data owners to tag their location points, resulting in ground truth data. Using this dataset we conducted an empirical study involving 45 participants to analyze how accurately they could infer the functional location of the original data owners under different conditions, i.e., three data representations, three data densities and four location types.

The study results indicate that while visual techniques perform better than textual ones, the functional locations of human activities can be inferred with a relatively high accuracy even using only textual representations and a low density of location points. *Workplace* was more easily inferred than *home* while *transport* was the functional location with the highest accuracy. Our results also showed that it was easier to infer functional locations from data exhibiting more stable and consistent mobility patterns, which are thus more vulnerable to privacy disclosures.

We discuss the implications of our findings in the context of privacy preservation and provide guidelines to users and companies to help preserve and safeguard people's privacy.

Author Keywords

Location data; data representations; empirical study; privacy.

ACM Classification Keywords

K.4.1 Computers and Society: Privacy

INTRODUCTION

People's location data is collected seamlessly [28] every day on a large scale, often without users' knowledge (i.e., silently through background processes) [20], [34].

Many people use social networking sites to share thoughts (tweets, status updates, etc.), pictures, videos, or interesting articles with friends, family and/or the public. Often, location data (geo-tags) is shared along with the timestamp – either as part of the information meant to be shared (secondary) or as the information itself (primary). Often emotions and feelings are attached to the information (Figure 1). Secondary sharing can sometimes be unintentional, since location sharing can be turned on as part of the tool and users can be unaware of it. Social networks such as Twitter, Facebook and Instagram allow location information to be shared either as a primary (Figure 1 (a), (c), (d)) or as a secondary (Figure 1 (b), (e), (f)) piece of information[1].

Figure 1. Examples of posts from social media sites with location data displayed as *primary* ((a), (c), (d)) and *secondary* ((b), (e), (f)) forms.

It is feasible to discover someone's identity by using only three location points [11]. Websites like PleaseRobMe[2] have in the past alerted Twitter users of the dangers of sharing geo-location data publicly. When Twitter users broadcast where they live, any subsequent tweets with a different geo-location publicly reveal that they are not at home, making their home

---

address more vulnerable to crimes. Foursquare[3] is another application that presents similar risks [29], since it allows users to publicly share (to review) their current location when reaching a place (restaurant, bar, museum etc.) [22].

Even though online service providers and their users might be aware of the risks [12] of sharing location data, the practice of capturing and broadcasting this information has not stopped or decreased, but rather has increased. Companies such as data brokers, social networking providers and advertisers use such data to profile their users to provide better targeted advertizing [40]. *GPS location* is the most requested permission in Android apps [21] and it is almost always associated with targeting advertizing due to its commercial value [23].

How easy is it to discover locational information that is private to people? Does one need a large dataset to do so? Can anyone infer these locations just by looking at data? How much data is required to get the right answer? In this paper we want to address these questions by conducting an empirical study in which we examine different ways of presenting location data, using different techniques (visual or textual), examining different people's routines, and different densities of location points (1 to 5 days). We will look at how these different factors might affect the ability to infer someone's location type, by anyone without any specialized tools, technical expertise and/or detailed knowledge of the area.

RELATED RESEARCH

With the widespread use of mobile devices, highly accurate location data is being collected [21] and often shared without users' knowledge [5]. The accurate and realistic nature of location data makes it one of the most valuable and personal types of information [35], [32]. Tracking users' locations has been shown to enable the inference of their behaviors [25], activity patterns [24], [37], the structure of their friendship networks [13], [32], semantic information about places [4] and personal associations [41] and even people's own identities [11]. It was also shown that only four spatio-temporal data points are needed to uniquely identify individuals in a set of de-identified data [11].

People generally know about these privacy issues, but nonetheless many still share their locations for their own rewards and benefits [38]. Individuals often share their locations to connect and coordinate with their social friends [26], and to recommend (and be recommended) nearby social and interesting events [30]. Location sharing has not only proven to benefit individuals but also society in general. In fact large datasets of people's locations have provided invaluable insights into the quality of urban services [18], [19] and socio-dynamics of neighborhoods [31]. These urban insights can lead to improvements in current and public structures and ultimately improve the quality of the geographical area itself. Location sharing can also benefit health services in preventing sickness [33] and tracking the spread of a disease [14].

Visualizing geo-referenced information has become increasingly popular. Numerous tools [2], [3], [8], [9], [10], [16], [17], [18], techniques [7], [27], [39], and apps are available

that provide simple and intuitive interfaces to view and plot large amounts of location data [4,5][1].

These analytic tools have used social location data and have been created to discover significant and common patterns, to understand the significance of locations [3], [17], and to identify people with common and related interests [2], [16]. Other tools have been created to help disaster responders [10] and/or police to focus and efficiently navigate and coordinate their efforts in emergency situations by identifying common [9] and/or anomalous [8] movements, and crowded places [3].

In order to improve the visualization of large location datasets, various new techniques have been designed and implemented to represent the directionality and routes of people's movements. Examples are flow maps [27] (origin to destination), as well as heat maps [39] in order to avoid occlusions. Boyandin *et al.* [7] describe *Flowstrates* that use heatmaps to represent changes from origin to destination taking into account time and space.

However, it is commonly assumed that extrapolating and inferring the above mentioned information requires specialized knowledge, technical expertise and sizable location datasets of large numbers of individuals. The aim of this research was to design a study to evaluate these assumptions.

AIMS & CHALLENGES

The aim of this research is to investigate how easy it is for a person (casual observer without specialized skills) to infer the type and/or relative function (*home*, *work*, *leisure* or *transport*) of a given geo-location point (shared as part of a tweet[6]) by visualizing location data in a simple and easily replicable manner (either on paper or using tools that require little to no technical abilities). This will highlight possible privacy risks related to the leak of this data (e.g., robbery).

We want to understand how accurately the type of a given geo-location point can be inferred based on different ways the data is visualized – visual (map-based) or textual (table-based) – and the number of location points (density) presented. We are also interested in understanding if certain location types can be more easily discovered than others and if the accuracy has any relation to the user's mobility patterns.

This research presented several challenges:

1. **Real-life tagged location dataset**: A dataset of people's real life locations needed to be created. This dataset needed to include accurate and realistic tags for each location point describing its type and/or function, classified into *home*, *work*, *leisure* or *transport*. This dataset also needed to include **realistic mobility patterns** based on people's routines. Routines could prove to have an effect when identifying the type and/or function of the location type. **Different densities** of location points to be presented needed to

---

[3]https://foursquare.com/

[4]http://www.hamstermap.com/quickmap.php
[5]https://cartodb.com
[6]Twitter was selected based on the availability of the social network itself. In Twitter we were able to collect users' tweets and geo-locations in addition to being able to ask permission and availability to be part of this research.

be included in this dataset. However, since we wanted to test different densities we needed to ensure that repetition of the same dataset would not affect responses.

2. **Data representations**: Data representations need to be designed to be easily replicated with online/offline tools without demanding advanced technical abilities, ideally in a way that could even be presented with only pen and paper.

3. **Control variables vs. study length**: There are many factors that may influence location inference. Ideally, a study should examine all these factors. However the number of stimuli required increases exponentially in relation to the number of control variables. In addition, repeated measures are also desirable, but lengthy studies would suffer from tiredness and fatigue. We thus have to design a balance between the number of control variables, the number of stimuli, the number of repeated measures and the overall length of study. It was important to test visual representations alone, hence we needed to **remove familiarity** since this could have had an effect on the responses.

4. **Learning effect vs. confounding effect**: On the one hand, the same real-life dataset is ideally tested under different conditions in order to minimize confounding effects. On the other hand, all real-life datasets are semantically rich and thereby sensitive to learning effects. After participants saw a dataset in one condition (e.g., visual), they would be able to reason about the same dataset in another condition (e.g., textual) with some ease. This is a common design challenge with identification tasks.

CREATING A REAL-LIFE LOCATION DATA SET

This part of the research was designed to capture people's real location patterns and relative location types, classifying the locations of people into: where they live (*home*), where they work (*work*), their movements (*transport*) and where they relax (*leisure*)[7] resulting in ground truth data.

Procedure

To solicit participation in our study, we identified and direct-messaged people using Twitter, posted advertisements on (Boston/Cambridge) Craigslist, and also emailed various university mailing lists in the Boston area. In all solicitations we attached a link with explanations of requirements for this study. We looked for people located in the Boston or Cambridge, Massachusetts (USA) area. We asked permission[8] to collect their tweets and the corresponding geo-location data attached to the tweets (latitude and longitude) over a period of three weeks. We informed them that we were interested in capturing their locations and that if they had enough location points at the end of the three week period we would contact them to ask them to participate in a study with the sole purpose of tagging their location as *home*, *work*, *leisure* or *transport*.

---

[7]An option was also given for uncertainty (*don't know*) or for not compliance with any of the other tags (*others*).

[8]Permission was not required or needed to collect tweets via the Twitter API; however, we used this opportunity to enquire about their intentions in participating in the tagging part of the experiment. It would have been counter-productive to collect tweets from people who were unwilling to participate in the tagging part of study.
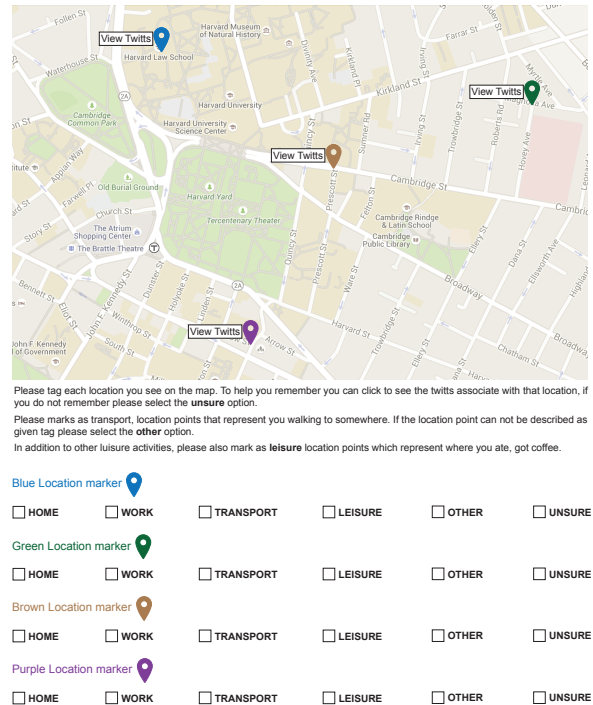

Figure 2. Example of Location Tagging Interface.

We recruited 230 people who were using Twitter and had location sharing turned on, and collected all of their tweets over a period of three weeks. At the end of the three week period we analyzed the location data and identified 87 users who had location points spanning three weeks. We contacted these users and asked them to participate in the study to disclose their location type. We explained that if they successfully tagged their locations we would add them to a random draw to receive a $20 Amazon gift voucher. Each user answered 12-25 questions depending on the variety of their location points. In each question we displayed four data points (Figure 2) and users had to tag each of the points as either *home*, *work*, *leisure*, *transport*, *other* or *unsure*. To help users remember we also showed the tweets associated with the location for the day. Tweets could be viewed by clicking the link next to the location marker (we limited the number of tweets shown to 15). We created a survey link for each user tailored to their location patterns and emailed it to them.

Each user also answered four demographic questions (gender, age, occupation, ethnicity) and four additional questions targeted at understanding their daily routines. In particular we were interested in understanding if their patterns could be categorized as regular/repetitive, irregular/non-repetitive or somewhere in-between these types.

Trustworthiness of the data

In order to ensure that participants tagged each location point with the appropriate description tag, we created five repeated questions in each survey designed to check their behavior, i.e., whether they were randomly clicking or not paying attention to the tagging task. We used location points either extremely close to one another (based on latitude and longi-

tude) or alternatively reused the same location point (in the absence of a close one). We presented these repeated questions at a distance from one another.

Results
43 people tagged their locations; of these, 37 participants (22 females, avg. age 29; 15 male, avg. age 27) consistently tagged their locations. The remaining 6 gave different types of location tags between repeated questions (designed either using location points very close to one another or the same point).

Table 1. The numbers of location points for each of the 27 data subjects grouped by the type of location for each category of density of location data: (L) low (1 day); (M) medium (3 days); (H) high (5 days).

D.S. ID	HOME			WORK			LEISURE			TRANSPORT		
	LOCATION POINT DENSITIES											
	L	M	H	L	M	H	L	M	H	L	M	H
1	7	16	24	5	18	7	-	3	14	2	4	8
2	3	10	20	5	25	31	-	-	-	3	1	2
3	3	5	7	5	14	19	-	2	-	1	1	4
4	8	34	47	3	3	-	3	5	7	-	-	-
5	6	7	14	6	27	42	-	-	-	-	3	4
6	3	4	17	8	25	31	-	1	4	-	4	2
7	12	15	36	-	9	5	-	-	-	2	3	4
8	1	13	20	8	18	31	2	2	2	1	2	2
9	4	16	8	4	20	24	1	2	15	-	1	3
10	9	30	41	-	-	-	-	-	-	-	2	1
11	2	1	16	8	16	27	-	11	4	2	8	12
12	4	21	41	1	-	-	-	-	-	1	-	1
13	3	13	12	-	1	-	1	-	11	3	6	12
14	5	22	22	-	-	12	-	2	10	3	4	5
15	8	3	12	-	23	32	-	3	4	-	2	2
16	3	23	24	2	4	4	3	-	-	3	5	12
17	2	3	8	7	22	35	3	4	10	2	2	8
18	4	12	15	4	13	32	-	-	5	2	5	4
19	2	11	17	4	10	27	3	4	5	3	6	4
20	3	11	21	-	5	-	2	8	17	1	6	9
21	6	25	41	-	-	1	2	4	6	3	-	4
22	3	9	15	6	22	32	1	1	4	1	1	-
23	5	16	21	-	-	18	-	1	3	-	-	2
14	5	18	25	8	18	33	-	1	2	1	1	3
25	9	7	16	3	21	41	-	2	-	1	4	4
26	10	32	42	1	-	5	-	1	1	-	2	6
27	7	34	47	-	-	-	3	3	4	2	3	4
TRAINING DATA												
28	5	21	19	-	1	20	6	6	1	1	2	5
29	8	20	28	1	-	3	1	-	4	-	3	11
30	4	6	6	7	13	25	-	-	2	-	4	4

We selected data belonging to 30 of these participants (Table 1). 27 participants' data was used in the study and 3 participants' data was used as part of the training. Table 1 shows the number of location points[9] for each participant (data subject). The number of points for each type of location is also shown. Three different location point densities are also shown. They represent the number of location points within 1 day (low);

---

[9]Location points marked as *others* or *unsure* were removed from the dataset and were not used in the study.

3 days (medium); 5 days (high). For each density the same day of the week was chosen. The low density represented Monday of week 1, the medium option represented Tuesday, Wednesday and Thursday of week 2, and the high option represented Monday to Friday of week 3. The data in different categories of density is from different weeks to prevent learning effects.

Participants' occupations and routines/mobility profiles
Selected participants have various levels of occupation and interests. We covered occupations such as undergraduate, masters and graduate students, part-time worker, salesman, housewives with and without children, engineers working from home, office or a colocation space, and self-employed people working from home or a colocation space. We selected these participants with various occupations to introduce necessary stochasticity in the real-life data.

Interesting observations
From the data we collected we can see that participants tend to mostly tweet when at home or work (Table 1). They sometimes tweet when moving (transport, car) or when they are out (leisure time). These results show that people tend to publicly give away their most commonly visited locations (which are often the most sensitive ones) when tweeting. Using public data repositories, the location information of someone's work or home can be used to determine the average income of one's neighborhood, average housing cost, debt, number and length of car ownership, demographics, likely political views, etc.

Removing data points
We removed location points where the data owners selected the tag to be *others* or *unsure*, but these only accounted at most for 10% of their individual data sets.

DATA REPRESENTATIONS: SIMPLE AND REPLICABLE
We designed two simple ways of visualizing the gathered location data, a visual and a textual representation. Location data has been commonly represented using a map, hence for for the visual techniques we showed the location data as points on a map. For the textual representation we used a table format. These two representations [36] have been shown to raise awareness on a user's perception of privacy.

These two data representations can be easily reproduced either by hand or by available tools online. By hand a person could print a map of the area and annotate location points (visual) or could annotate sequential location points on paper (table). Several (free) tools exist that allow people to insert the desired location data (either as geo-tags or addresses) and automatically visualize the results on a map (e.g., mapsdata[10] or CartoDB[11]).

Visual Representation
The visual technique displays each location point (marker) on a map (Google Maps was used). At the center of each marker we showed a number (ID) representing the order of appearance (captured time was used to order each point). In order to avoid occlusion as much as possible, each marker
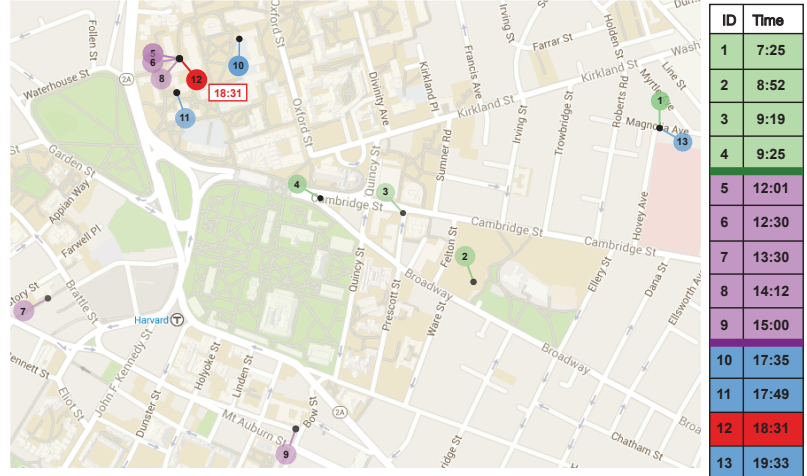
---

[10]http://www.mapsdata.co.uk
[11]https://cartodb.com

(a) Textual (table-based)

ID	Latutude	Longitude	Address	Time
1	42.3769	-71.1073	Magnolia Ave.,Cambridge	7:25
2	42.3743	-71.1111	Broadway, Cambridge	8:52
3	42.3755	-71.1155	Cambridge St., Cambridge	9:19
4	42.3753	-71.1136	Cambridge St., Cambridge	9:25
5	42.3783	-71.1192	Massachusetts Av., Cambridge	12:01
6	42.3783	-71.1193	Massachusetts Av., Cambridge	12:30
7	42.3737	-71.1220	Brattle St., Cambridge	13:30
8	42.3783	-71.1193	Massachusetts Av., Cambridge	14:12
9	42.3717	-71.1154	Bow St., Cambridge	15:00
10	42.3784	-71.1171	Oxford St., Cambridge	17:35
11	42.3776	-71.1190	Massachusetts Av., Cambridge	17:49
12	42.3782	-71.1192	Massachusetts Av., Cambridge	18:31
13	42.3769	-71.1073	Magnolia Av., Cambridge	19:33

Legend: Location in question — Morning — Afternoon — Evening — End of morning — End of afternoon

(b) Visual (map-based)

ID	Time
1	7:25
2	8:52
3	9:19
4	9:25
5	12:01
6	12:30
7	13:30
8	14:12
9	15:00
10	17:35
11	17:49
12	18:31
13	19:33

Legend: Location in question — Morning — Afternoon — Evening — End of morning — End of afternoon

Figure 3. Textual (a) and visual (b) representations used in the study. The location density displayed is *Low* (1 day). The detailed captured time is shown in addition to the color depending on the period of the day.

was assigned a transparency of 80% and tilted by 20°. The line from the location point to the marker was alternated in size (from long to short) to reduce occlusion problems. The ID number of each location point was also reported on the side of the map as a table (Figure 3 (b)). The marker of the location to be inferred was set to 3px larger than the other ones. The visual representation was presented in two forms, static and animated visualization:

- **Animated**: In the animated visualization, numbered location points (markers) were shown at one second intervals following the order of the ID table positioned next to the map (Figure 3 (b)). The ID table was also animated. Markers and corresponding ID appeared at the same time. Unlike a traditional memoryless animation [15], once the marker appeared, it remained on the map and within the side table. The table next to the map showed the location point ID and captured time. The table was colored according to the time of day (morning, afternoon or evening). Similarly to the text-only technique, colored lines were used to delimit the day periods and different days.

- **Static**: The static visualization showed all the numbered location points (markers) on a map with corresponding ID table next to it. The static techniques are equivalent to the final state of the animated one.

Textual Representation
The text-only technique displayed data in a table (Figure 3 (a)). The table was composed of an ID row (as shown in the visual techniques) ordered by capture time, with corresponding geo-location in the form of latitude and longitude. The street address associated with each location was also displayed (Figure 3 (a)) followed by the captured time. Each row (representing a location) was also colored according to the time of day (morning, afternoon or evening). At the end

of each day a colored line depicting the end of that period was added. This was included to avoid confusion when the location to be inferred (in red) appeared at the end and the beginning of a period. When multiple days were presented (data densities medium (3 days) and high (5 days)) a yellow line was inserted to show the end and the start of the new day. All this information was documented in the legend below each table and map.

PRIVACY & GEO-TAG: INFERRING TYPES OF LOCATION
We are interested in measuring the feasibility and accuracy of uncovering the functional types of people's places (*home*, *work*, *transport* and *leisure*) by visualizing different amounts of data about real people's locations using different data representations. In particular we are interested in:

1. How do different data representations impact participants' ability to infer functional location?

2. Does the accuracy of inference depend on different types of location?

3. Does increasing data density improve or impede the accuracy of inference?

Apparatus
The study was developed as a web application using Google Maps and D3 [6]. The map was 780px × 585px and any interactions, such as zooming, panning, scaling were disabled, because they were not the focus of the study but could introduce significant confounding effects. All landmarks were also removed to ensure that there were no advantages between the visual or textual techniques. To avoid any cross browser compatibility issues, the study was performed using only the Chrome browser. A progress bar was shown on top of the study to indicate the participant's progress in the study. Four

optional answers – *home*, *work*, *transport* and *leisure* – were displayed below the visual or textual data representation. A *don't know* option was added in case people could not motivate an answer based what they saw.

Visual consistencies between data densities

We wanted to ensure that the visual representations (static and animated map) were consistent between density levels (for each data subject). To do this we removed data points that would create a different zooming level between the different data densities. These points are outliers representing one-off "leisure" activities. Table 1 reports the number of locations after this change. We showed zoomed-out versions of the map for the low or medium densities to ensure that the same zooming and visual clues were conveyed at each level.

Process for choosing target location

One target location was set to be identified for each data subject. This enabled a comparison between inferring the same location type across different data representations and data densities (Table 1). Only one type of location was targeted for each data subject to avoid skewing assessment. This was done because people might not have visited (hence tagged) all the locations types we are interested in within each collected density. For example a person working part-time could have gone to work on Tuesday but not on Monday. This kind of routine would not allow *work* to be suitable since it was not present as a location tag in the low density level (low data density uses location points captured on Monday of week 1).

Suitable target locations are the ones where the data subject has visited that location type at a close time within each collection period (density level). If the data owner has not visited a possible target type during a similar or close time in any of the days, the location target was not a suitable choice. This was done to avoid the confounding problems generated by different times throughout the day. This strategy also alleviates would-be confounding effects caused by using different weekdays for stimuli at the same density levels, while preventing showing the same data (e.g., Monday of week 1) in the stimuli for different levels. The day of the week was never shown to participants. Participants were informed that the location data displayed was collected from weekdays. Functional locations of *home*, *leisure* and *transport* were asked to be inferred for seven data subjects, and *workplace* was asked to be inferred for six data subjects[12] (a total of 27) (Table1).

Procedure

The study was conducted in a computer lab with identical machines in order to avoid any variation of screen size and computer speed and to ensure that participants were not distracted or disturbed during the study execution.

As each of the 27 data subjects yielded 3 non-overlapping datasets (low, medium and high density), there are 81 distinctive datasets. Each participant could encounter each dataset once in the study to preventing learning effects. Meanwhile,

---

[12]There is a good reason to fix the number of data subjects to 27 (see the section on Procedure), we accepted the compromise of having one fewer workplace. All results are reported as a percentage.

each dataset needed to be tested with three data representations (text, static and animated visualization). In order to achieve this, we adapted a between-group design by dividing participants into three groups. For each of the 81 datasets, one group saw text, another saw static visualizations and the other saw animated visualizations. With a careful planning, we ensured that each dataset was seen only once by one group, and each group saw exactly 9 stimuli in each of the 9 conditions (3 representations and 3 density levels). [Detailed stimuli info for each group is given in the supplementary material.]

Participants in the study answered a total of 90 questions: 9 training questions and 81 study questions. Participants were randomly assigned to one of the three groups.

Information session

The information session consisted of a presentation used to familiarize participants with the details about execution, requirements and respective remunerations of the study. Each participant was given a £10 Amazon gift voucher after successfully completing the study. We explained what participants should see, how they could answer, and how they could progress to the next question. We presented the three different representations and explained the day periods and the detailed captured time. We also informed participants that the data to be viewed was captured during the week and that they were going to see data representing 1, 3 and 5 days. Each day was separated by a yellow line (similarly to the the day periods separation shown in Figure 3).

When discussing remuneration, we explained that if random clicking was detected they would forfeit compensation. As an added incentive, an additional reward (£15 Amazon gift voucher) was also given to the top three participants with the most accurate answers. At the end of the information session, participants could ask questions and seek clarification. Before starting, they signed the consent form. At this point they could start the training session. Participants could leave the study at any time without any penalties.

Training

Nine training questions were created to familiarize and train participants for the study. These questions showed all combinations of visualization (static, animated and text only) and location density (low, medium, high).

Prior to each question, participants were shown a page explaining what they would be seeing and details on how to conduct the study. In this page we emphasized again that they should not randomly guess an answer but rather select the *don't know* option if they were unsure about the type of location. After the explanation, the question was shown as it would appear in the study. The location data points (corresponding data subjects) used in the training session were not reused in the actual study (Table 1). Questions were not randomized in the training part.

Study

After the training session, participants had to answer 81 questions. Questions within the study were randomized. Participants had to provide an answer (choosing from *home*, *work*, *leisure*, *transport* and *don't know*) before they could progress

to the next question. Participants were allowed to change their answer prior to pressing the "next" button to proceed to the following question.

Survey
At the end of the study we gave participants a paper survey (4 questions) designed to understand their motivations and reasoning behind their answers. We enquired about their preferences among data representations and data densities.

Trustworthiness of the data
Questions in the study were randomized, appearing in a different order for each participant. The order of the questions was saved to account for fatigue effects and random clicking. The accuracy of participants' answers did not decrease towards the end of the study. We did not detect any random clicking. This suggests that the incentive of the additional reward of £15 appeared to be effective.

RESULTS & ANALYSIS
Participants
We solicited participation in our study using internal mailing lists. 47 people participated. 2 participants did not finish the study. 45 people successfully completed the study (15 in each of the 3 in-between groups).

Of these, 24 were male (avg. age = 33) and 21 female (avg. age 34), 1 gender not disclosed (age = 32). Level of education varied from having attended but not completed high school (3), completed high school (2), two-year college degree (5), four-year college degree (3), undergraduate student (6), completed four-year college degree (3), completed master degree (7), being a graduate student (4) to advanced graduate work or completed Ph.D (12). The study took an average of 54 minutes to complete (*min* = 45; *max* = 98 minutes). Participants did not live in or have any extensive knowledge of Cambridge/Boston MA[13]. All participants lived in England, UK. This was done in order to alleviate confounding effects due to significant biases towards those with local knowledge.

Responses
We collected 3,645 responses from 45 participants (Table 2). The responses cover all combinations of the three visualizations (animation, static and text-only) with the three different data densities (low, medium and high (Table 1)) and four location types (*home*, *leisure*, *transport*, *workplace*).

Table 2. Number of participants' responses grouped by visualization factor (animation, static and log-based), shown for each location type (*home, leisure, transport and workplace*) within each data size (LOW (one day), MEDIUM (three days), HIGH (five days)).

LOC.	ANIMATION			STATIC			LOG-BASED			
TYPE	L	M	H	L	M	H	L	M	H	Tot.
HOME	105	105	105	105	105	105	105	105	105	945
LEISURE	105	105	105	105	105	105	105	105	105	945
TRANSP.	105	105	105	105	105	105	105	105	105	945
WORK	90	90	90	90	90	90	90	90	90	810
TOTAL	405	405	405	405	405	405	405	405	405	**3,645**

[13]The Twitter data used for this study was collected from people living in the Cambridge/Boston MA, area.

We gathered 1,215 responses for each of the three visualization conditions: animated, static and table-based. Each condition was shown with the same location variable. For location type of *home*, *leisure* and *transport*, we gathered 945 responses each. For location type of *workplace* we gathered 810 responses[14]. Table 2 outlines responses for each visualization type and time variable. Questions were evenly distributed between each data representation (Table 2).

Location Types and Data Representations
The level of accuracy to which location can be inferred depends on both the type of location and how it is presented.

Are certain types of location more easily inferred than others?
Location types tagged as *leisure* were more difficult to deduce resulting in the lowest accuracy level overall ($\mu = 53\%$).

Table 3. Accuracy level for each type of location, showing the mean of the percentage of accuracy, the standard deviation, the odds ratio and confidence level and the corresponding p-values.

TYPE OF LOCATIONS.	(%) MEAN μ	(%) STD. DEV σ	ODDS RATIO	95% CONFIDENCE INTERVAL MIN	MAX	P-VALUES
Home	58	18	0.757	0.651	0.881	0.001*
Leisure	53	18	0.587	0.505	0.682	0.001*
Transport	72	14	1.712	1.46	2.01	0.001*
Workplace	69	17	1.431	1.21	1.69	0.001*

Home was found to present a lower mean accuracy of discovery than *work*, with *home* being discovered $\mu = 58\%$ while *work* $\mu = 69\%$. Discovering when a person was moving (*transport* type of location) was found to be the easiest to deduce with $\mu = 72\%$ (Figure 4, Table 3).

There are statistically significant differences between inferring the different types of locations determined by one-way ANOVA $F(3, 176) = 12.235$, $p < 0.001$. A Tukey post-hoc test showed that there were statistically significant differences between discovering *home* and *workplace* ($p < 0.010$), *home* and *transport* ($p < .001$), *workplace* and *leisure* ($p < 0.001$) and *leisure* and *transport* ($p < 0.001$).

The odds of identifying the type of location when it is *transport* or *workplace* are 1.7 and 1.4 respectively (Table 3) compared to the odds of identifying when the location is *home* or *leisure* with odds of 0.881 and 0.68 respectively (Table 3).

However in all cases participants were found to be able to identify the type of location two to three times higher than chance (25%), underlining serious privacy disclosures based on location data.

Do data representations impact on accuracy of inferencing?
There are statistically significant correlations between accuracy of inferencing of functional locations within each data representation. This is determined by one-way ANOVA $F(2, 132) = 8.841$, $p < 0.001$ for correct responses.

A Tukey post-hoc test showed that there were statistically significant differences between table-based and static techniques

[14]As explained in the procedure section, *workplace* was asked to be identified six times (instead of seven) to balance the stimuli (data representations and data densities) in the study.
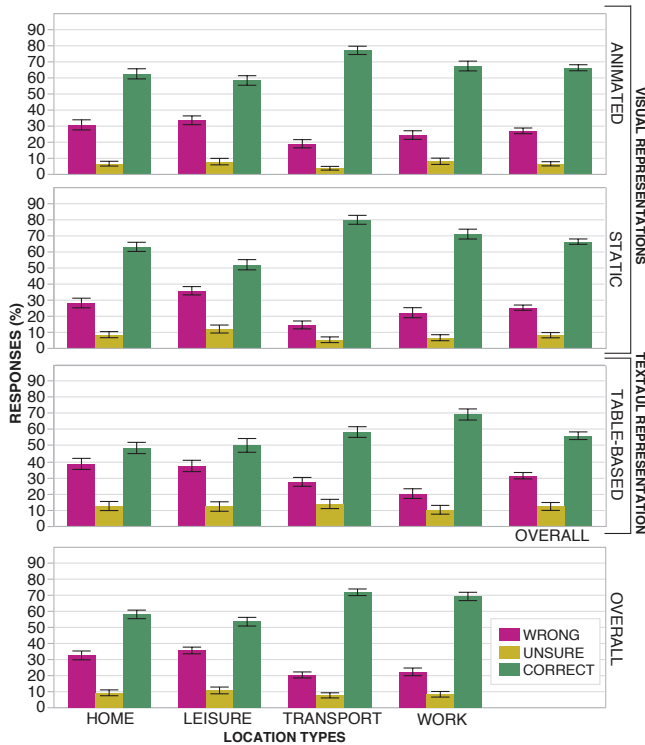
Figure 4. Percentage of correct, unsure and incorrect responses for each type of location grouped by different representations of the data. An overall count for location type is included. The standard error is shown, each error bar is constructed using 1 standard error from the mean.

The animation response time is shown as **BEFORE** representing the duration from when the animation started to when the location to be inferred (red) was shown; **AFTER** representing the duration from when the location to be inferred was shown until an answer was selected;
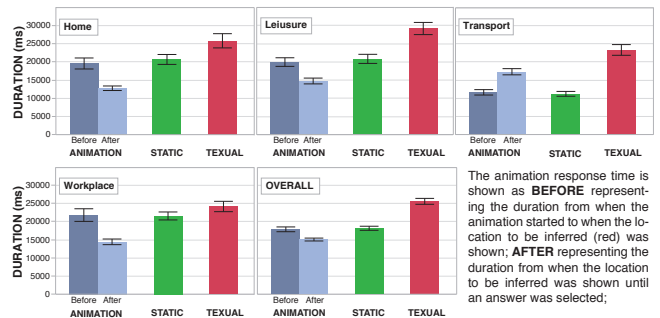
Figure 5. Average of response time (ms) of correct answers for each location type, grouped by visualization technique. An overall response time for each technique is included. The standard error is shown, each error bar is constructed using 1 standard error from the mean.
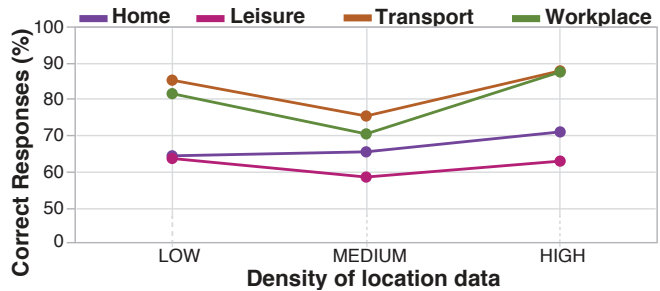

Figure 6. Percentage of correct responses for each type of location shown as increasing level of location density (low, medium, high).

($p < 0.001$), and between the animated and table-based techniques ($p < 0.001$), with visual techniques leading to more correct deductions of location type (animation $\mu = 66.3\%$; static $\mu = 66.4\%$) than the textual ($\mu = 56.4\%$) one. There was no statistical significant difference between visual techniques. Textual techniques were shown to present a higher accuracy of inferencing when location type was *work* and data densities were low and high (Figure 7). Visual representations presented a higher accuracy when location type was *transport*. This is due to the inherent type of locations, *transport* and the directionality that can be better visualized using a visual representation.

Figure 5 shows the mean of response time for each of the three visualization techniques grouped by type of location. We can see that visual techniques, which lead to more correct answers, do not take a longer time to reach a correct answer (with the exception of transport, where static visualizations give the quickest response). Textual techniques require more time to answer than the two visual techniques. *Transport* location types, which are the ones with the highest accuracy of inference, also show the lowest response time across all representations in comparison with the other location types.

The number of location points presented was found to have an effect on accuracy in both visual and textual representations of the data. In particular the accuracy of inferring the type of location was affected within each technique.

Does data density affect accuracy of inference?
Different location types - *home*, *work*, *leisure* and *transport* - were found to present different accuracy levels based on the number of location points presented (Figure 6). The highest level of accuracy across all location types was achieved when the highest number of location points was used (5 days).

While increasing the number of location points improved accuracy when it came to *home* locations, this was not always the case when it came to other types of location. In fact when the location type was either *transport*, *work* or *leisure*, the accuracy decreased when a medium density (3 days of location data) was presented. This shows that presenting more points of location data does not always improve accuracy, and, it may sometimes decrease accuracy.

Location types and data densities
The density of location points presented and the visual representations used affected the accuracy of inferring the location type. Textual representations performed worse than visual ones for all location types except *work* where when the low and high density of location points were presented, textual representation performed better (Figure 7).

Location types and mobility patterns
Are there routines that are more privacy-vulnerable?
We selected data subjects with different routines in order to be able to present a variety of patterns to participants. Based on data owners' self reporting and from analysis and motiva-
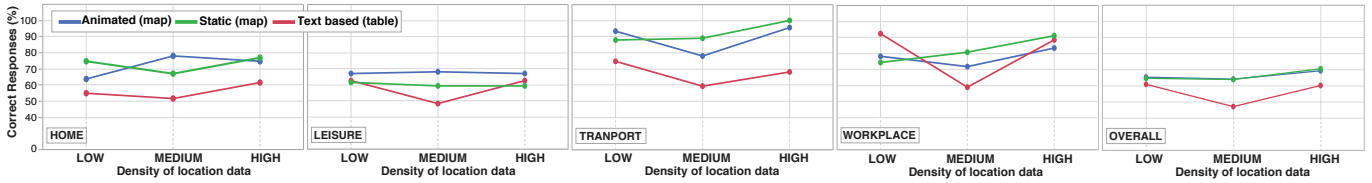
Figure 7. Percentage of correct responses for each type of location shown as increasing level of location density (low, medium, high).

tion behind this report, we categorized each routine as regular, irregular or semi regular[15].
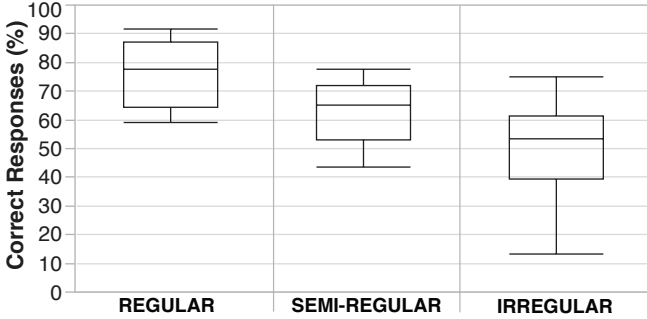

Figure 8. Box plot of percentages of correct responses grouped by the data subjects' routine patterns such as *regular*; *semi-regular*; *irregular*.

A *regular* routine depicts a stable working and home schedule; *semi-regular* routines encompass routines which can be somewhat irregular, for example a temporary worker who might have a stable routine only 2 or 3 days a week; *irregular* routines are routines with no stable schedule (i.e., going to work and/or coming home at different hours, working from different locations, taking transport at different times, etc.).

The types of location to be inferred were distributed between these different mobility patterns depicting users' routine types[16] (Table 4). Participants reported that when looking at the data, they made assumptions about people's routines based on the clustering and time of the day (Figure 9).

Table 4. Location types to be inferred within each routine/mobile pattern; R = Regular; SR = semi-regular; IR = irregular.

LOCATIONS	R	SR	IR	LOCATIONS	R	SR	IR
HOME	2	3	2	WORK	2	2	2
LEISURE	2	3	2	TRANSPORT	2	3	2

Figure 8 shows that the functional locations of people living structured and routine-based lives are highly likely to be inferred correctly compared to less structured or chaotic ones.

In fact, participants found it difficult to answer when several clusters were present in the data or when no clear clusters

were shown. This could be related to the fact that participants looked for repetitive patterns. This is also probably the reason why textual techniques performed worse, especially when it came to location types like *transport* where an animated representation outlines the mobility aspect. *Work* was the one type that was the easiest to infer using textual techniques, likely due to the fact that clear repetitive patterns can be seen even when in a table[17].
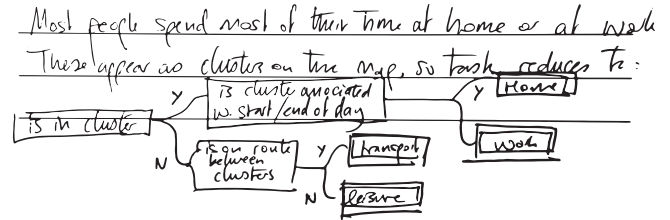

Figure 9. One participant graphical explanation behind their choice during the experiment. This motivation has been reported being behind all other participants' responses as well.

Inclinations towards particular data representations
Are people visual- or textual-inclined?
Our results highlight three distinct inclinations among participants: **visually**, **text** and **hybrid** inclined participants.

Visually inclined: These participants (36) were found to be able to deduce location type when the location data was represented using one (or both) of the *visual* visualization techniques (static or animated). The majority of these participants however were more inclined to answer correctly when using the animated (17) rather than the static technique (14), with five participants showing equal accuracy level with either technique. Participants might have been more inclined towards the animated technique because it delivered temporal information more intuitively and participants used more time to answer these questions[18] (Figure 10).

Text inclined: These participants (6) were found to be able to better recognize the functional type of location when the data was represented using the textual technique (Figure 10).

Hybrid: These participants (3) were found to able to infer the functional type of locations equally well using visual and textual representations (Figure 10).

---

[15]Data subjects maintained their routine patterns in between the different weeks i.e., data densities.

[16]Semi-regular had three more questions than the other two patterns. The is due to the fact that we were using real data and that we had a limited number of participants within the regular and irregular mobility patterns.

[17]*Home* presented a lower accuracy level due to the fact that *home* was often confused with *workplace* due to the participants' assumptions that a person would go to work in the morning.

[18]At the end of the animation, the data was shown as it appeared in the static technique.
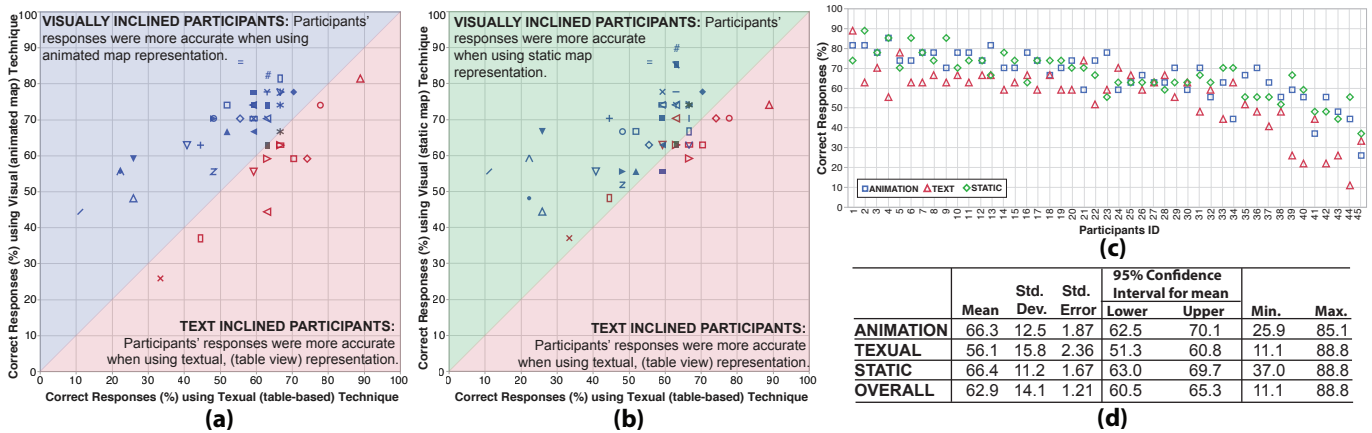
Figure 10. Inclinations of participants towards particular data representations, showing accuracy (%) for textual vs. visual representations: (a) animated vs. textual; (b) static vs. textual. The accuracy (%) for each participant for each data representation is also shown (c). Participants are ordered from most accurate to least accurate. The descriptive statistics of accuracy between the three data representations are also included (d).

CONCLUSION

When we think of collecting personal data, it is commonly framed in the form of big data collection and analysis of mobility patterns over several days. However we have seen that with a small number of data points, people's locations can be inferred. This kind of information can lead to several privacy disclosures. Using publicly available data, the type of locations can be used to estimate someone's average income based on one's neighborhood, average housing cost, debt, and other demographic information, such as political views etc.

We have shown that deducing people's most frequent and private locations such as *work* and *home* can be achieved using only a small sample of location points (1 day worth). Adding a larger sample of location points has the potential to increase accuracy or confusion. We saw that 3 days worth of data led to more confusion and decreased accuracy. *Transport* was found to be identified more accurately than the rest. *Work* was discovered more accurately than *home*, while deducing other types of location like *leisure* proved to be more complicated even when a larger location dataset was used.

The study showed that most participants benefited from visual techniques and that these have higher response time than than textual one. It is interesting to note that private locations can also be inferred without using visual techniques.

Our study showed that using the textual visualization could lead to correct identification over 50% of the time, with 6 participants correctly deducing functional location more accurately than with visual techniques. We showed that three distinct and different affinities for data visualization were present within our study participants. We found participants were either visually or textually inclined (only nine presented hybrid results, having correctly answered the same number of questions in both visual and text-only techniques).

Guidelines
This study has shown the sensitivity of location data and the need to adapt technologies to allow people to be able to specify which personal location should be (or not be) shared.

- **Enquiring about sensitive locations**: Tools could enquire about a functional location after a certain amount of sharing, and hence either stop sharing this location information or share nearby locations rather than specific ones which could present privacy risks to the users.

- **Creating confusion**: Adding confusion to the data, e.g., adding fake locations to make regular and semi-regular routines look irregular.

- **Adding meaningless clusters of data**: Participants in our study described looking for clusters to identify personal locations such as *home* and *work*, hence adding more clusters of location points might confuse analysis and help people preserve their privacy and minimize possible risks.

- **Tagging sensitive locations**: Users could tag their location and decide a priori which locations should be shared or not.

EXPERIMENT LIMITATIONS

We wanted to use real data, however this brings associated problems. Our understanding of users' locations and functional locations depends on self-tagged data from Twitter users. Due to the nature of real data, it is possible that we have not covered all possible routines. In addition, the privacy risks we have highlighted are associated with leakage of location data specific to Twitter users. People tend to use different social networks for different purposes and this can affect the locations where they share the information. This study is not representative of all social networks. For example data leaked with Instagram might reflect functional locations representing likes, dislikes or hobbies rather than personal locations like *home* or *work*. This location leakage could have other risks rather than disclosing personal locations.

ACKNOWLEDGMENTS

REFERENCES

1. S. Ahern, M. Naaman, R. Nair, and J. H. Yang. 2007. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-referenced Collections. In *Proc. of ACM/IEEE JCDL*. 1–10.

2. G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom. 2013. Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. *Computing in Science & Engineering* 15, 3 (2013), 72–82.

3. G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. 2011. From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *Proc. of IEEE VAST*. 161–170.

4. N. Andrienko, G. Andrienko, G. Fuchs, and P. Jankowski. 2015. Scalable and Privacy-respectful Interactive Discovery of Place Semantics from Human Mobility Traces. *Information Visualization* (2015).

5. R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. 2013. "Little Brothers Watching You": Raising Awareness of Data Leaks on Smartphones. In *Proc. of ACM SOUPS*. 1–11.

6. M. Bostock, V. Ogievetsky, and J. Heer. 2011. D3: Data-Driven Documents. *IEEE Trans. Visualization & Comp. Graphics* 17, 12 (Dec 2011), 2301–2309.

7. I. Boyandin, E. Bertini, P. Bak, and D. Lalanne. 2011. Flowstrates: An Approach for Visual Exploration of Temporal Origin-destination Data. In *Proc. of EuroVis*. 971–980.

8. J. Chae, Y. Cui, Y. Jang, G. Wang, A. Malik, and D. S. Ebert. 2015. Trajectory-based Visual Analytics for Anomalous Human Movement Analysis using Social Media. In *IEEE EuroVis Workshop on Visual Analytics*, E. Bertini and J. C. Roberts (Eds.).

9. J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D.S. Ebert, and T. Ertl. 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Proc. IEEE of VAST*. 143–152.

10. J. Chae, D. Thom, Y. Jang, S.Y. Kim, T. Ertl, and D.S. Ebert. 2014. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics* 38 (2014), 51–60.

11. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 3, 1376 (2013).

12. S. Dredge. 2014. Tinder dating app was sharing more of users' location data than they realised. *The Guardian* (20th February 2014).

13. N. Eagle, A. Pentland, and D. Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proc. of National Academy of Sciences* 106, 36 (2009), 15274–15278.

14. S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429 (May 2004), 180–184.

15. D. Fisher. 2010. *Animation for Visualization: Opportunities and Drawbacks*. O'Reilly Media, Chapter 19, 329 – 352.

16. G. Fuchs, G. Andrienko, N. Andrienko, and P. Jankowski. 2013. Extracting Personal Behavioral Patterns from Geo-Referenced Tweets. In *AGILE Conf. on Geographic Information Science*.

17. T. Fujisaka, R. Lee, and K. Sumiya. Discovery of User Behavior Patterns from Geo-tagged Micro-blogs. In *Proc. of ICUIMC*. Article 36, 10 pages.

18. L. Gabrielli, S. Rinzivillo, F. Ronzano, and D. Villatoro. 2014. From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns. In *Citizen in Sensor Networks*, Jordi Nin and Daniel Villatoro (Eds.). Lecture Notes in Computer Science, Vol. 8313. 26–35.

19. S. Jiang, Jr. J. Ferreira, and M. C. Gonzalez. 2012. Discovering Urban Spatial-temporal Structure from Human Activity Patterns. In *Proc. of ACM SIGKDD Int. Workshop UrbComp*. 95–102.

20. B. Krishnamurthy and C. E. Wills. 2010. On the leakage of personally identifiable information via online social networks. *SIGCOMM Comput. Commun. Rev.* 40, 1 (Jan 2010), 112–117.

21. I. Liccardi, J. Pato, and D. J. Weitzner. 2014. Improving Mobile App Selection through Transparency and Better Permission Analysis. *J. of Privacy & Confidentiality: Vol. 5: Iss. 2, Article 1*. (2014), 1–55.

22. J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. 2011. I'm the Mayor of My House: Examining Why People Use Foursquare - A Social-driven Location Sharing Application. In *Proc. of ACM CHI*. 2409–2418.

23. M. Merisavo, J. Vesanen, A. Arponen, S. Kajalo, and Mika R. 2006. The effectiveness of targeted mobile advertising in selling mobile services: An empirical study. *Int. J. Mob. Commun.* 4, 2 (Jan 2006), 119–127.

24. A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. 2011. An Empirical Study of Geographic User Activity Patterns in Foursquare. *ICWSM* (2011), 570–573.

25. T. Parka, Rashmi S., and Gavriel S. 2008. Effective advertising on mobile phones: A literature review and presentation of results from 53 case studies. *Behav. Inf. Technol.* 27, 5 (Sep 2008), 355–373.

26. S. Patil, G. Norcie, A. Kapadia, and A. J. Lee. 2012. Reasons, Rewards, Regrets: Privacy Considerations in Location Sharing As an Interactive Practice. In *Proc. of SOUPS*. 5:1–5:15.

27. D. Phan, L. Xiao, R. Yeh, P. Hanrahan, and T. Winograd. 2005. Flow Map Layout. In *Proc. of IEEE INFOVIS*. 219–224.

28. D. Pierce. 2015. Location is your most critical data and everyone's watching. *Wired* (27th April 2015).

29. T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida. 2012. We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proc. of ACM UbiComp*. 898–905.

30. D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. 2010. Recommending Social Events from Mobile Phone Location Data. In *Proc. of IEEE Data Mining (ICDM)*. 971–976.

31. R. Rösler and T. Liebig. 2013. Using Data from Location Based Social Networks for Urban Activity Clustering. In *Geographic Information Science at the Heart of Europe*, Danny Vandenbroucke, Bénédicte Bucher, and Joep Crompvoets (Eds.). Springer Int. Publishing, 55–72.

32. A. Sadilek, H. Kautz, and J. P. Bigham. 2013. Modeling the Interplay of People's Location, Interactions, and Social Ties. In *Proc. of AAAI IJCAI*. 3067–3071.

33. A. Sadilek, H. Kautz, and V. Silenzio. 2012. Predicting disease transmission from geotagged micro-blog data.. In *Proc. of AAAI Conf. on Artificial Intelligence*. 136–142.

34. F. Shih and J. Boortz. 2013. Understanding People's Preferences for Disclosing Contextual Information to Smartphone Apps. In *Human Aspects of Information Security, Privacy, and Trust (Lecture Notes in Computer Science)*, L. Marinos and I. Askoxylakis (Eds.), Vol. 8030. Springer, 186–196.

35. J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe. 2014. Money Walks: A Human-centric Study on the Economics of Personal Mobile Data. In *Proc. of ACM UbiComp*. 583–594.

36. K. P. Tang, J. I. Hong, and D. P. Siewiorek. 2011. Understanding How Visual Representations of Location Feeds Affect End-user Privacy Concerns. In *Proc. of ACM UbiComp*. 207–216.

37. I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. 2009. Measuring serendipity: Connecting people, locations and interests in a mobile 3G network. In *Proc. of ACM IMC*. 267–279.

38. J. Tsai, P. G. Kelley, L. F. Cranor, and N. M. Sadeh. 2010. Location Sharing Technologies: Privacy Risks and Controls. *J. of Law & Policy for the Information Society* 6, 2 (2010), 119–151.

39. J. Wood, J. Dykes, and A. Slingsby. 2010. Visualisation of Origins, Destinations and Flows with OD Maps. *Cartographic Journal* 47, 2 (2010), 117–129.

40. H. Xu, H.-H. Teo, B. C. Y. Tan, and R. Agarwal. 2009. The Role of Push-Pull Technology in Privacy Calculus: The Case of Location-Based Services. *J. of Management Inf. Syst.* 26, 3 (Dec 2009), 135–174.

41. C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. 2007. Discovering Personally Meaningful Places: An Interactive Clustering Approach. *ACM Trans. Inf. Syst.* 25, 3 (July 2007), 31.