Eurographics Conference on Visualization (EuroVis) 2016
K.-L. Ma, G. Santucci, and J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

Using Visualization to Explore Original and Anonymized LBSN Data

E. Tarameshloo[1] M. Hosseinkhani Loorak[1] P. W.L. Fong[1] and S. Carpendale[1]

[1]Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

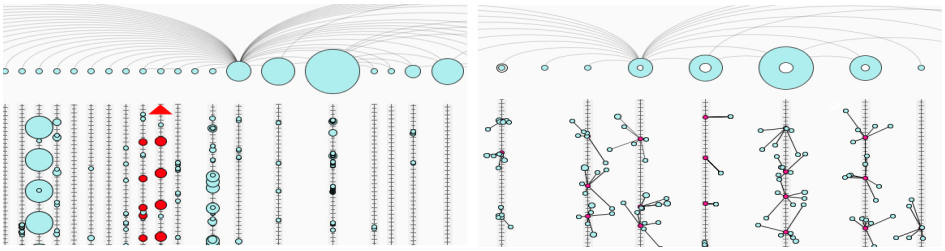

Figure 1: *The visualization of original (left) and anonymized (right) location-based social network (LBSN) data using GSUVis*

Abstract

We present GSUVis, a visualization tool designed to provide better understanding of location-based social network (LBSN) data. LBSN data is one of the most important sources of information for transportation, marketing, health, and public safety. LBSN data consumers are interested in accessing and analysing data that is as complete and as accurate as possible. However, LBSN data contains sensitive information about individuals. Consequently, data anonymization is of critical importance if this data is to be made available to consumers. However, anonymization commonly reduces the utility of information available. Working with privacy experts, we designed GSUVis a visual analytic tool to help experts better understand the effects of anonymization techniques on LBSN data utility. One of GSUVis's primary goals is to make it possible for people to use LBSN data, without requiring them to gain deep knowledge about data anonymization. To inform the design of GSUVis, we interviewed privacy experts, and collected their tasks and system requirements. Based on this understanding, we designed and implemented GSUVis. It applies two anonymization algorithms for social and location trajectory data to a real-world LBSN dataset and visualizes the data both before and after anonymization. Through feedback from domain experts, we reflect on the effectiveness of GSUVis and the impact of anonymization using visualization.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces D.4.6 [Security and Privacy]: —Data Anonymization

1 Introduction

We present GSUVis, a new visual analytic tool designed to provide better support for privacy experts and data consumers, when they work with location based social network (LBSN) data. GSUVis contains two visualizations: 1) SocialArcs, which represents the individuals' friendship graph, and 2) TravelLines, which visualizes the people's location trajectories. These two visualizations are combined with two powerful anonymization algorithms [LT08, GG03] to support the exploration and impact of anonymizing LBSN data. GSUVis visualizes LBSN data in both the original and anonymized form to support experts in exploring LBSN data and the effects of anonymization algorithms on data utility.

This research is motivated by the proliferation of Internet and GPS enabled devices, resulting in an increased demand for effective methods of exploring, analyzing, and understanding LBSN data. Improving the ability to analyze the massive amounts of data about the location and social connections of individuals that is already being collected by LBSNs such as Twitter and Yelp, could provide benefits in many contexts including transportation, marketing, health, and public safety. However, preserving the privacy of peo-

ple when publishing LBSN data is a challenge for data collectors and a concern for data contributors. Privacy experts commonly employ anonymization to preserve people's privacy before publishing the data for use by third-parties or the public. It is difficult to preserve the privacy of individuals while maintaining the utility of the information [ASNB12].

Since little effort has been invested in defining utility metrics for LBSN data [MJ11], privacy experts who want to design LBSN anonymization techniques and data consumers who want to examine the utility of the acquired data currently lack analytic support. Such support could help them understand the effects of the anonymization techniques on data utility. Working closely with privacy experts, we designed, implemented and assessed, GSUVis, with the goal of assisting privacy experts to explore and analyze LBSN data both in its original and its anonymized form (Figure 1).

We started our design with six initial interviews with privacy experts to better understand the requirements and challenges of our target domain. From these interviews, we extracted and classified a set of analytical tasks for LBSN data, and derived a set of design goals that informed the design of our proposed visualization. Using an iterative design process, we developed GSUVis to support the exploratory analysis of original and anonymized LBSN data. GSUVis is a hybrid representation that combines arc diagrams and timelines, with a new application of glyphs for displaying spatial data in an abstract form. To assess the benefits of GSUVis, we conducted an insight-based evaluation [LBI*12] with four domain experts. We discuss our evaluation results in terms of how GSUVis can help experts in LBSN data exploration (§6). We observed that working with our visualization generated excitement among the experts as it enabled them to visually observe the effect of anonymization algorithms. It also triggered the domain experts ask new questions and resulted in new insights for defining utility metrics for LBSN data. Specifically, our contributions are the following:

1. GSUVis is a new hybrid visual analytic tool designed and developed in close collaboration with the privacy experts to assist them in developing a better understanding of utility in anonymized data (§5).
2. A list of data exploration and analysis tasks in the domain of data privacy for LBSN datasets (§4). This list may be useful for subsequent visualizations of LBSN data, and might apply to data with similar features such as human activity trajectories, telephone call sequences, or individuals' shopping items.
3. A set of new insights that can be derived from LBSN data, and its anonymized counterpart gathered by evaluating the system with privacy experts (§6).

2 Related Work

We group related literature under: a) information visualization, and b) data privacy.

2.1 Information visualization

Information visualization has been used to represent both social networks [VLKS*11, CM11b], and location data trajectories [AAB*13]. For visual representation of networks three techniques are commonly used: node-link diagrams, adjacency matrices, and hybrids of both [VLKS*11, BEW95]. Heer *et al.* [HB05] designed a visualization system for navigation and exploration of large-scale online social networks using node-link diagrams. Dork *et al.* [DCW12] presented EdgeMap with a timeline layout that uses arc diagrams [Wat02] to visualize the influence relations. NodeTrix [HFM07], MatLink [HF07], and Papilio [LFC14] are hybrid visualizations designed for exploring relationships using combinations of node-link diagrams and adjacency matrices.

The famous visualization of Napoleon's march on Moscow [Min65] by Charles Minard may be the first to use location trajectory data on maps in an aggregated form. Andrienko *et al.* [AAB*13] surveyed visual analytic tools for movement data offering a taxonomy of three types of movement trajectories: 1) a single trajectory of an object, 2) multiple trajectories of a single object, and 3) simultaneous movement of many objects. Most of the work in the third category represents aggregated movements on the geographical maps, mainly in 3D. For example, Tominski *et al.* [TSAA12] proposed a 3D stacking approach for representing trajectory attributes on a map. Scheepens *et al.* [SWW11] presented a density map to interactively explore multiple attributes in trajectory data in an aggregated form. Although our visualization of LBSN data falls under the third category from a visual movement perspective, our objective, derived for an LBSN utility focus presents a detailed and non-aggregated 2D visualization of location trajectories.

Luo *et al.* [LM14] survey visual analytic tools that integrate social networks with geography, suggesting that the main objective of the surveyed geo-social analytic tools was to understand the social processes. Using this perspective, they discussed three groups, 1) data exploration, 2) decision-making, and 3) predictive analysis. Though there are some exceptions such as MacEachren *et al.*'s [MJR*11] system for visualizing geo tagged twitter data, Luo *et al.*, noted that integration of geography and social networks has not received the attention needed and they identified developing theory, methods, and tools capable of simultaneous consideration of spatial and social factors as open challenges in visual analytic tools. While GSUVis is designed for data exploration and decision-making, we used two linked visualizations instead of the more common multiple coordinated views with limited projected location points on an actual map.

Based on our analysis of the literature, our work differs from current LBSN visual analytics in the following aspects. First, GSUVis was designed with and for privacy experts. Second, the previous work considers the representation of one dataset at a time. However, in our collaboration with privacy experts, we observed that it is important to be able to compare original and anonymized data values in a single analytical view. Third, a large number of location points can be represented for each entity within our visualization.

2.2 Data privacy

Although there is a vast body of research on privacy preserving data publishing [FWCY10, WYLC10, MYYR13], the literature is relatively silent on the privacy of LBSN data in a sharing and publishing context. While Masoumzadeh *et al.* [MJ11] have considered anonymization of both location and social data together, their provided utility metric has focused only on location distortion.

Both Wu *et al.* [WYLC10] and Aggrawal *et al.* [ALY15] have surveyed of recent techniques and developments on privacy preserving graph and social network data. For simple graphs, they classified the anonymization techniques into three main categories:

1) k-anonymity based approaches via edge modification; 2) probabilistic methods via edge randomization; and 3) privacy preservation via generalization. Location privacy attacks and their current state-of-the-art countermeasures are extensively discussed in [WSDR14, Kru09]. Privacy preservation in location trajectory publishing is the most closely related point to LBSN data anonymization and *Clustering-based* [ABN08] and *Generalization-based* [NASG09] are the two main approaches in privacy of location trajectories [CM11a].

3 Background

To provide context, we outline location based social network systems, data privacy, and the dataset we used.

Location-based Social Network (LBSN) systems are largely used for data analysis of social networks in a geographical context. An LBSN is an extension of a social network that contains two key pieces of information: 1) a social graph of members, and 2) a set of location trajectories.

LBSNs allow people to declare their current locations through a mechanism commonly known as "check-in", that is in the form of time-stamped physical location coordinates (e.g., GPS coordinates). This spatial knowledge-base can then be employed to provide a variety of services to the members. Besides well-known LBSNs like Twitter, Instagram, and Facebook there are also LBSNs where knowing members' location is necessary to provide the service. For instance, in PulsePoint (pulsepoint.org), members who are trained in cardiopulmonary resuscitation (CPR) will be notified if someone nearby is having a cardiac emergency; in Banjo (www.ban.jo), people can explore new members in their vicinity while looking for nearby breaking news and live events; in Foursquare and Yelp, in addition to nearby friends, one can find restaurants and stores that have good reviews from friends and other active members.

Publishing LBSN data is possible through a set of application programming interfaces (APIs) that query the LBSN knowledge-base for data collection and analysis. One of the most well-known consumers of LBSN data are Social Media Monitoring Systems (SMMS). They actively collect information from different social media channels to analyze volume, trend, and opinion about a topic or brand in different geographical areas. WeLink, Snaptrends, and BlueJay are a few instances of SMMS. For example, BlueJay collects and scans tweets to monitor public safety at large events.

Data Privacy: in a typical scenario of data collection and publishing, there are three primary parties: *data owners (creator)*, *data publisher (collector)*, and *data recipient (consumer)*. Assuming a trusted data publisher, a major challenge in data publishing is to preserve the privacy of data owners while maintaining information usefulness for potential data recipients [ASNB12]. While there is vast body of literature on privacy preserving data publishing (PPDP) [FWCY10], in practice, every PPDP mechanism has its own assumptions as well as the requirements of the data publisher, the recipients, and the purpose of data publishing. If the publisher knows in advance the data analysis tasks required by the recipient, she could release a customized dataset that preserves specific properties of the data for such an analysis. However, in many cases, the data publisher does not know who the recipient is and how the published data will be analyzed. In this case, a general purpose metric based on the *principle of minimal distortion* [FWCY10] is defined to measure data quality in anonymized published data.

Table 1: *Dataset statistics before and after anonymization.*

LBSN	Nodes	Edges	Locations
Gowalla (G)	196,591	1,900,654	6,442,890
Anonymized G	196,591	1,820,692	4,387,478

3.1 Dataset

We use *Gowalla* [CML11], which is a published dataset of real-life LBSNs. The dataset is composed of a friendship graph and a set of location trajectories for each node of the graph. Location trajectories contain all of the public check-in data between Feb. 2009 and Oct. 2010. Table 1 shows some statistical information about our dataset before and after we apply the anonymization techniques. The published dataset in its current form contains personal ids that are replaced by a unique random number for each individual and a set of GPS coordinates with temporal information for some individual's ids. To evaluate our visualization tool, we applied two anonymization algorithms (described in §5.3) to the graph and location information in the Gowalla dataset, and call the resulting anonymized dataset *Anonymized G*.

4 Design Process and Goals

We started our design process by conducting six semi-structured interviews with experts involved in different aspects of data security and privacy research to gain a wide range of perspectives. Three of the experts are security consultants from a big enterprise responsible for verifying anonymization of large datasets before publishing; two experts are chief data officers working on data governance and secure data handling; and one is a privacy researcher who works on improving anonymization algorithms.

The interviews lasted between 30 to 60 minutes. Five interviews took place at the privacy specialists' offices, and one via Skype. All the interviews were audio recorded and transcribed. During the interviews, we asked general questions regarding 1) the different types of datasets that require anonymization before publishing, 2) the factors that are commonly considered in analyzing LBSN data and its anonymized form, 3) the role of privacy in LBSN data, 4) the regulatory requirement for data privacy, 5) the process of data anonymization, 6) common anonymization algorithms, and 7) the challenges in the process of anonymizing and publishing the results. Through these interviews, we gained a better understanding of the significant impact of LBSN data and its anonymization from the perspective of both data owners and data recipients.

4.1 Task Analysis

We analyzed the transcripts of the interviews by a process similar to affinity diagramming, grouping the common ideas and concepts together. Through this analysis, we extracted a set of 20 analytical tasks that a visualization should support in order to help the experts better analyze LBSN data, and its anonymization. We then classified the analysis tasks into two main categories: 1) tasks for exploring the original LBSN data, and 2) tasks for exploring the anonymized data. The list of tasks is presented in Figure 2. The "data" column indicates whether the data needed for performing the task is available in the dataset. The columns P1-P6 show which the experts mentioned each task. This task classification informed the design of our visualization and can be used to guide future research in similar application areas.

	Tasks for exploring the original LBSN dataset	Data	P1	P2	P3	P4	P5	P6
T1	Identify the popularity of members based on the number of social connections.	■	■			■		
T2	Identify the co-located check-in points.	■	■		■	■	■	■
T3	Identify the co-located check-in points occuring at about the same time.	■	■		■	■		■
T4	Identify a person's favorite locations based on the number of check-ins within her location trajectory.	■	■	■		■	■	
T5	What is the most/least popular location among members?	■	■	■		■		■
T6	Is there any correlation between social connections and check-in points?	■		■			■	
T7	Is there any pattern in terms of the check-in behaviour of an individual?	■	■		■	■		
T8	Compare the check-in behaviour of two (or more) friends.	■			■	■		
T9	What is the trajectory of an individual based on location type? (e.g., home, hospital, home, shopping mall)			■				■
T10	Order check-in points according to their occurance time.	■	■			■		
T11	What is the temporal frequency of location check-ins for an individual?	■			■		■	■
T12	Are there any outliers in the sequence of declared locations of a person?	■		■				
	Tasks for exploring the anonymized LBSN dataset							
T13	Identify the data loss in the social graph (e.g., amount of adeleted edges).	■		■		■		
T14	Identify the data loss in the location trajectory of each individual (e.g., movement of anonymized point vs original ones).	■	■		■	■		■
T15	For which cases the data utility was/wasn't well preserved?	■			■		■	
T16	Compare different (two or more) anonymization algorithms in terms of data utility.		■	■	■		■	
T17	Is there a correlation between more/less social people and their data utility?	■				■		■
T18	Is there a correlation between people who have more/less check-in points and preserving their data utility?	■		■	■	■		
T19	Which location types (ie.g., hospital, shopping center) preserved after anonymization?			■				■
T20	Is there any outliers for co-located data after anonymization? (e.g., co-located points mapped to different locations)	■				■		

Figure 2: *The list of analytic tasks are presented in rows. The column "Data" shows whether the needed data dimension is available for performing the task. The columns P1 to P6 show the experts who mentioned the task.*

4.2 Design Goals

We derived our design goals from our observations, our interviews, and discussions with the privacy experts. These design goals informed the design of GSUVis.

DG1 Compact view: In LBSN data, social relations might affect the location trajectory of individuals and vice versa. Therefore, our goal is to visually integrate social relations and location data in a single compact view to better discover their effects on each other.

DG2 Readability through the locality: Providing readability is a useful factor in analyzing both social connections and the location trajectory data. Visualization of large LBSN datasets may cause readability issues. Thus, in showing a readable subset of data, our goal is to display the data entities that are locally relevant with each other to increase the chance of discovering data patterns.

DG3 Adjustable information representation: It should be possible for the analyst to adjust the amount of information that is represented based on her preferences. This will let the analyst to tune the level of detail at which she wants to perform her analysis tasks.

DG4 Holistic views: The experts were clear about wanting to see relationships between factors, in particular between the original and anonymized data, and amongst as many location details as possible.

DG5 Visible changes: The visualization components should allow the analyst to observe her expected changes on data after anonymization process.

5 The GSUVis Walk-through

We designed GSUVis to address our design goals (§4.2) and analytical tasks (§4.1) as extracted from interviews with experts as well as our observations. GSUVis has five principle parts: a) a social network visualization (SocialArcs); b) a location trajectory visualization (TravelLines); c) a social graph anonymization algorithm; d) a location anonymization algorithm; and e) the interactive interface that holds these parts together. The visualization components show comparative representations of original and anonymized data. Figure 3 shows GSUVis, visualizing LBSN data before anonymization (**DG1**).

We used an iterative paper prototyping design process. Since we could call upon the expertise that included security, privacy, and visualization, we could iteratively assess our prototypes, comparing the various design possibilities. To clarify our description of navigation and interaction with GSUVis, we include a use case walk-through example, annotated with the relevant design goals and the numbers of supported tasks where applicable. Emma starts working with GSUVis to examine a data subset, because from her experience of working with data, subsets can provide insight into data, provide a good understanding of how the applied anonymization technique works, and whether it sacrifices the data utility [YaKSJ07]. To select a subset of the Gowalla data, Emma uses Gephi [BHJ*09], filters our large social network dataset (**DG2**) and exports her selected data subset. Then, she imports her selected nodes into GSUVis.

5.1 SocialArcs

Our visualization, SocialArcs, of the social graph of an LBSN dataset, leverages arc diagrams [Wat02]. To start her analysis, Emma loads her selected sample dataset of Gowalla into SocialArcs (Figure 3). In this layout, she sees two types of information: individuals, and the friendship relations among them. Each individual is depicted by a circle with a size that is relative to the number of their friends (T1).

We demonstrate the friendship connections between entities with undirected arcs. An arc diagram is a Node-Link diagram [BETT98], in which arcs allow node adjacencies to be shown along a 1-dimensional layout (**DG1**). This leaves space on one side of the nodes, making it possible to show additional data dimensions in the space opposite to the arcs. Thus, we use arc diagrams to indicate friendships and use the remaining empty space for visualizing the location trajectory of the individuals (T6). To minimize the lengths

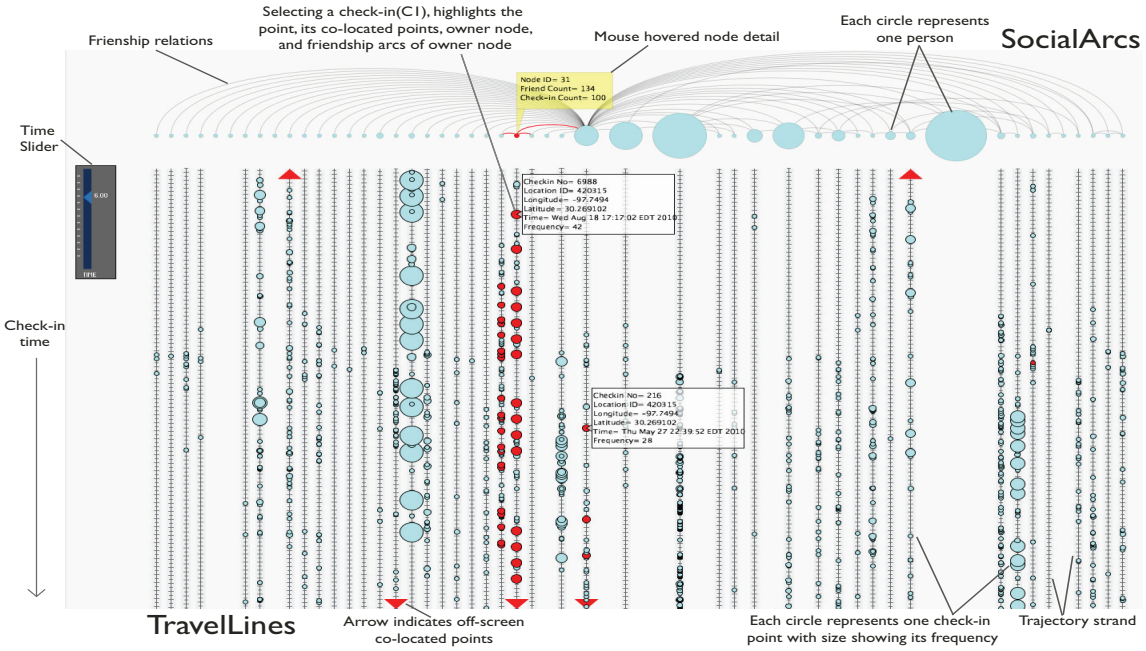

Figure 3: *The GSUVis system showing part of the Gowalla dataset before anonymization. It is composed of (a) Social Network Visualization (SocialArcs), and (b) Location Trajectory Visualization (TravelLines).*

of the arcs and to reduce the link crossings, we adapted the barycenter heuristic algorithm [MS05] to order the nodes in a way that the socially connected ones are positioned in close proximity (**DG2**).

By default, all the arcs are light gray to allow Emma to readily see her selected, highlighted nodes during her targeted exploration. Emma starts her interaction with SocialArcs to explore the LBSN social data by selecting an individual's node which highlights the node and its connecting arcs to other nodes (Figure 3). In this way, Emma can find friends and friends-of-friends of entities. For each individual, she also explores more detailed information namely ID, number of friends, and number of check-in points through hovering over its corresponding circle (see Figure 3). Through her analysis of the social graph in SocialArcs, she develops a few hypotheses. For example, she wonders whether social individuals (nodes with bigger sizes) also have social friends. She can test this hypothesis by checking the node sizes of friends of individuals with large node size. In an another hypothesis, she checks if friends or friends-of-friends of a selected node have similar features to that node (example features include, number of friends and number of check-ins).

5.2 TravelLines

Commonly, each person within an LBSN dataset has their own trajectory of locations (see §3.1). A typical approach to trajectory data visualization is to represent the data on spatially constrained layouts such as maps. However, maps pose challenges such as: visual occlusion for comparing the location trajectories across multiple individuals (T8) [LM14]; connections indicated as lines on maps can imply movement along a path when actual movement is uncertain; and some individuals have consecutive long-distance gaps in geographical check-in points, which introduces another scalability issue.

Switching her focus of exploration to location data, Emma interacts with the visualization of location trajectories of individuals. Each individual has a unique vertical line, which we call a *trajectory strand*, connected directly beneath their node that contains a number of circles each representing one check-in point of that individual (T7, T12) (see Figure 3). In order to understand the check-in behaviour of individuals (T7, T8), Emma needs to access the time interval between check-in points and the frequency of their occurrence. Conveniently, each trajectory strand is timeline visualization that shows this type of data. Furthermore, combining the social and location trajectory data and presenting them in a single view can assist Emma as she explores and studies the social connections in the context of location trajectories and vice versa (**DG1**).

After interacting with TravelLines, Emma observes that individuals performing check-ins more frequently have more check-in points almost adjacent to each other on their trajectory strand (Figure 3). This is because the space between points on each trajectory strand are defined based on a function of check-in time (T10, T11). On the other hand, the location circles of people who tend not to check-in very frequently, are shown far from each other. Emma also notices that the size of each circle located on the trajectory strand is different. After examining information of a few check-in circles she discovers check-in circle size is in accordance with the number of check-ins that the person has in that specific location over their whole location history (T4).

Based on the statistics provided with the dataset used (see Table 1), each individual has around 30,000 check-ins on average. As often the screen size is limited, not all check-ins can be presented on a single screen without causing occlusion. Thus, Emma uses the *time slider* (see Figure 3) to interactively alter the visible time frame of the trajectory strands depending on her preferences (**DG3**). She can see more check-in points with higher occlusion, or

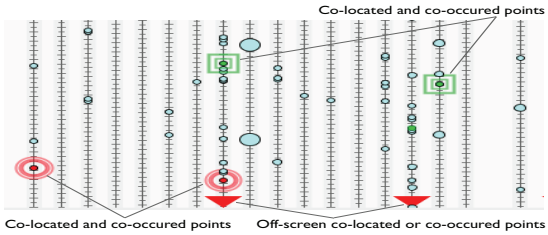

Figure 4: *Co-located and co-occurred locations for two selected check-in points (represented in green and red).*

less check-in points (with lower occlusion) on the screen (T7, T8). She can also use a smooth sliding interaction feature on the trajectory strands to explore the off-screen check-in points. Sliding up or down over a trajectory strand alters its visible check-in points. This change is animated smoothly based on the dragging speed the analyst applies. Thus, to reach to the far check-in points Emma drags quickly down the trajectory strands. For accessing relatively close check-in points Emma simply slows her dragging speed.

Studies have shown that, finding co-located individuals can lead to friendship link prediction [CTH*10]. This indicates that discovering co-located individuals within an LBSN dataset could potentially be insightful for an analyst. This is in accordance with our collected set of tasks in §4. Emma uses our implemented *co-located filtering* technique to visually explore co-located points and discover the patterns among them (**DG2**). She clicks on a check-in point and then selects the circle, showing its co-located check-in points (T2). The base node shows the owner, and the connecting arcs are highlighted with a unique color and show the owner's friends (see Figure 3). We limited the use of unique colors to the twelve perceptually distinguishable colors as recommended by Ware [War13]. Exploring further, Emma right-clicks on a check-in point to see detailed information about the chosen location. This information includes location ID, longitude, latitude, and timestamp (Figure 3).

Emma is also interested in identifying individuals with co-located check-in points at approximately the same time (T3). She double-clicks on a check-in point to highlight all the circles co-located and co-occurred within the chosen point. The points are displayed using animated-bullseye marks that all have the same geometric shape. However, it is possible that some target points fall off the current screen view. Therefore, an arrow on one of the pertinent end of the trajectory strand indicates the existence of other off-screen co-located or co-occurred check-in points on that strand (see Figure 4).

Currently, we consider a time window of three hours to determine whether two co-located check-ins happen at the same time. We chose three hours time window threshold based on the minimum time that seemed to be reasonable to an expert, e.g., *"social gatherings often go on for at least three hours" (P04)*. GSUVis also allows Emma to perform multiple check-in and co-occurrence selections in TravelLines. A new geometric bullseye shape will be assigned to each new selected point and its co-located, and co-occurring points. This assists Emma in differentiating various selected check-in points based on their shape, which makes their comparison easier (T3, T6). As an example, in Figure 4, two distinct geometric patterns of bullseye (i.e., circle and square) indicate that two co-located, and co-occurring filterings have been performed on data.

5.3 Data Utility Visualization

After exploring the original LBSN data, Emma wants to analyse the same data, however, now in its anonymized form (T13-T20). She is also interested in exploring the effect of anonymization on data utility (T13, T14). To this end, while she already has loaded her original LBSN data, she also loads anonymized data simultaneously. After she loads both original and anonymized data, GSUVis provides a visual representation of the modifications (**DG4**) that have been applied to the data through anonymization (see Figure 6). She observes that both SocialArcs and TravelLines are capable of reflecting changes, that is the amount of data loss after anonymization (**DG5**). In the following, we first describe the anonymization algorithms we implemented for anonymizing Emma's dataset. Then, we explain how GSUVis visualizes anonymized data in SocialArcs, and TravelLines.

5.3.1 Deployed Anonymization Algorithms

To support experts in evaluating anonymized data and comparing original and anonymized data, we implemented a variation of graph anonymizer approach proposed by Liu *et al.* [LT08] to add or delete edges for graph anonymization. We also implemented a location anonymizer based on k-anonymity, inspired from Gruteser *et al.* [GG03]. However, our proposed visualization is independent of the applied anonymization methods as it can represent the edge modification and location points displacement independently from the anonymization algorithm. Essentially, in the graph anonymizer with edge modification, the number of added or deleted edges is a metric to define data utility in the anonymized graph. In the location anonymizer, we translate each GPS coordinate of the trajectory to a GPS point that the anonymizer calculates as the representative point of generalized area. In this approach, a general metric based on displacement from original points is defined to measure data utility.

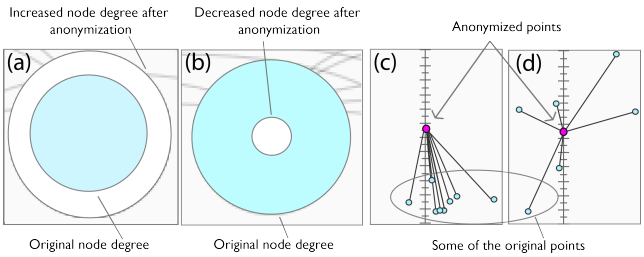

Figure 5: *Anonymization detail on node degree and location trajectories. a) Increase in node degree, b) decrease in node degree, c) a star glyph with dense original points d) a star glyph with divergent rays. In star glyphs, direction and length of rays is based on the angle and Euclidean distance of original points to the anonymized point.*

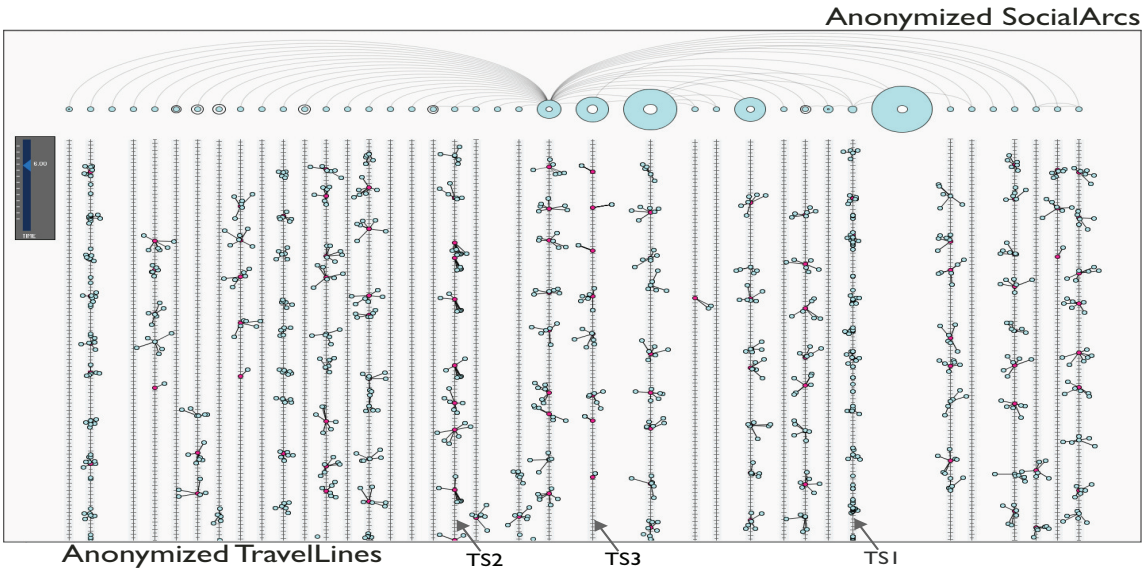

Figure 6: *The GSUVis system showing part of the Gowalla dataset after loading anonymized data. SocialArcs and TravelLines represent data before and after anonymization. Trajectory strands (T1, T2, and T3) are explained in §6.*

5.3.2 Data Utility in SocialArcs

In GSUVis, Emma observes that the effect of the graph anonymizer on LBSN data is visually represented by the node degree of the graph (T13, T15). Figure 5a and 5b show two nodes, one with some edges added and another with some edges deleted. For Emma to compare the node degree before and after anonymization she makes use of concentric rings which have two different meanings (**DG5**). First, if the degree of a node has increased, then the inner filled circle shows the original node degree and the external ring shows the anonymized degree (Figure 5a). Second, if the node degree has been reduced, then the outer ring demonstrates the original node degree and the inner unfilled ring shows the degree after anonymization (Figure 5b) (T16, T18). Always the colored part reflects the original data.

5.3.3 Data Utility in TravelLines

In publishing anonymized data, a set of original points translate to one anonymized location. To illustrate the effects of anonymization algorithm on location data (T14, T15, T20), we take advantage of *star glyphs* [BKC*13] in the design of TravelLines. Each star glyph is a representation of an anonymized location shown in the middle and a set of original locations represented in the form of glyph rays. Using star glyphs enable us to visually demonstrate two important aspects of original data points compared to the anonymized one (**DG4**). First, the length of each glyph ray is in proportion to the actual Euclidean distance between the original location and its anonymized one (T15). Second, the angle of each ray represents the direction of its attached original location compared to the anonymized point (T15). Although glyphs are often used for representing multi-dimensional data, in GSUVis, we employed them in a novel way for visualization of homogeneous location data that differ from each other in terms of position and distance from a certain point (**DG5**). Figure 5c, 5d illustrate two star glyphs. The centre point in each glyph belongs to the anonymized dataset, while the scattered rays around the centre point are from the original dataset.

Emma starts assessing anonymized and original points in TravelLines. She observes different patterns in glyphs as a result of location anonymization. For example, in Figure 5c, the original points are all located in either south or southeast direction relative to the position of the anonymized point. However, in Figure 5d, the original points are scattered in various directions around the centre point. Based on this observation, Emma can imply that in a short period of time the individual of Figure 5c was interested in checking-in only in the southern part of a given area. Whereas individual of Figure 5d did not have any preferences. This also shows Emma that original location points for the first individual will be relocated to an anonymized point which is not an accurate representative of original points, thus a possible loss in data utility.

Figure 6 demonstrates how the star glyphs are integrated into the trajectory strands to represent the original as well as anonymized location data, in the context of anonymized SocialArcs visualization (T17, T18). In this visualization, the glyphs are presented in a sequence based on a function of time (T10). This means that if a set of consecutive check-ins map to one point, then one glyph stands for all of them. However, if a check-in (X) occurs in the middle of the sequence (Y) with a different anonymized location, then, the glyph of locations for the sequence Y divides into two: glyph of locations before X, and the one after X.

6 Expert Feedback

To assess the effectiveness of our proposed visualizations, we presented GSUVis to four privacy experts and researchers. One of them took part in the initial interviews and the three others were the experts we were meeting for the first time. Each interview lasted approximately 75 minutes. We first briefly explained how to use GSUVis with a small synthetic dataset (20 minutes), then the experts interacted with the system for approximately 15 minutes and then we conducted a semi-structured interview where the experts also asked

questions and generally discussed the visualization with us. Then, we loaded the real dataset (§3.1) to GSUVis and showed this to the participants. For approximately 40 minutes, they explored the data using GSUVis, simultaneously discussing their insights, as well as providing comments and feedback about the system.

Generally the feedback was extremely positive and the experts were excited about their insights and findings. They described several different ways in which they could take advantage of this tool, e.g., *"Very insightful system! We can turn this into an assessment tool for privacy auditors [...] If the auditor was asked where are the gaps, then he can tell look! Here!"*, *"If I am buying data I would definitely like to have this system in my back pocket"*, and *"...make a mobile application of your system. People can see how much data is collected about them and then the next check-in is smarter [laughs]."* Also, they all mentioned that they were not aware of any other analytic tool that can help them visualize and explore LBSN datasets before and after anonymization.

6.1 Data Insights

During our study, the experts' exploration of the data led them to gain more insight into the data and also raised new questions based on their increased understanding of the dataset. These new insights led quickly to ideas for further investigation. We categorized the experts' insights into three main groups, namely, **people's behaviour**, **hypothesis generation**, and **utility evaluation**. People's behavior refers to behavioral insights and findings about individuals' habits or patterns of action. Hypothesis generation refers to insights that assist an analyst to generate different hypothesis about data. The utility evaluation relates to any insight about data utility and about the anonymization. We show examples of data insights gained by the experts within the context of these three categories.

People's behavior: While our participants were exploring the data using GSUVis, they discovered several behavioral facts about some individuals within the dataset. For example, by adjusting the time slider, a participant visually recognized a person who has repeatedly checked into locations with similar frequency within his trajectory (see the highlighted points in the trajectory of the red node in Figure 3). She picked one of the points and further explored other co-located points. As a result, she found that the chosen person always checked-in one particular place during the week. "*[...] this guy is in the same place during the week but different places on weekends. Kind of a unique.*" As another example, an expert was interested in individuals who have rarely declared their locations. She picked a few of these individuals in the data and tried to examine if those few check-in points in their trajectories are within popular places or even exist in the trajectories of other people. She found that these members only check-in in special places where others normally do not check-in. "*[...] they are into places that are kind of unique in the data. No one else has those points*".

The experts were also interested in the level of detail that the glyphs can provide about individuals' behavior and how the anonymization affects this information (see TS1 in Figure 6). For instance, a participant mentioned: "*These small stars with divergent arms show one who has been in various nearby places, right?*"

Hypothesis generation: This series of observations relates to when participants found something unique and started hypothesizing about the possible explanations behind them. As the first example, participants noticed that a member's trajectory contains many location points with similar frequency of check-ins. Selecting one of these similar frequency locations, revealed that those places are actually the same location. Thus, they started hypothesizing that this location must be the person's favorite place. "*She has been in these locations very often. maybe these are her favorite places.*" "*[...] why all her places are the same? looks like she really loves this place.*" As another example, one member who has the similar check-ins during the week (U1), caught our participant's attention. After exploring and highlighting co-located points in the trajectory of the friends of this particular member, he realized that the same location points also appear in the trajectory of two friends of the target person. Thus, he hypothesized that they might have a social gatherings in that place. "*These friends have similar locations. Maybe it's a get together thing.*" Our own exploration of this specific point after the study showed that U1 has been going to that place since August. However, the friends just started checking-in to that location in October. A possible hypothesis is that U1 had an influence on his friends since they had never been in that location before. Soon after U1 visited the place, his friends started visiting too.

The individuals who are not very active in reporting their location caught an expert's attention. She found some outliers in the behavior of these inactive members. She found some relatively dense active parts in their trajectories, and hypothesized that the member must be in a special location during the time of being active in reporting locations. "*This guy doesn't check-in usually, but he checks a lot in the third week of May. What are these locations? Is he on vacation?*" Looking and exploring through the visualization, a privacy expert discovered that there are no obvious correlations between social people and their check-ins. "*Some with many friends don't have many check-ins and some do.*" "*I thought my friends affect my check-ins. Apparently they don't.*"

Utility evaluation: While exploring the anonymized data using GSUVis, participants started evaluating the data utility of the anonymization algorithm. Going back to the example of one member who has similar check-ins during the week (U1), a participant decided to verify if the pattern she found in the original data for the member U1 is still valid in anonymized data. By exploring the anonymized trajectory of her targeted member, she found that the anonymized trajectory still shows a regular pattern of check-ins for that member (see TS2 in Figure 6). However, she pointed out an interesting privacy concern. "*I kind of can see the same pattern here but the pattern for him is still unique within other members.*" "*[...] this can make him vulnerable to be identified even in anonymized data.*" This finding generated a hypothesis about whether the anonymization algorithm should create more similar location trajectories within members with unique habits, however, this might cause even moreloss in data utility.

One expert had an interesting observation in the anonymized dataset that led to finding a visual cue for data utility. "*[...] why do some people have fewer glyphs? But their glyphs have more dots around them.*" We described how fewer glyphs for a trajectory is an indication of more data loss in anonymized data. In other words, fewer dense glyphs means that more original points are mapped to the same anonymized points. This visual cue would also assist a data consumer (with no expertise in anonymization) to select anonymized data with higher data utility (data with more sparse glyphs).

Through looking at glyphs, an expert became interested in the pattern of glyph sequences in each member's trajectory and how it could reveal the person's check-in characteristics. He found an individual who normally has very sparse glyphs. This means that, his consecutive check-ins are far from each other. However, during a short time period, the glyphs became dense, which means that his consecutive check-ins are close to each other and they are mapped to one center location. As a result, the expert found this as an outlier in the person's check-in behaviour (see TS3 in Figure 6). "*[. . .] this is unusual for her [. . .] she checks-in far places but [. . .] she stayed in a place for quite some time*" This insight brought us another finding for data utility. Although the sequence of glyphs can reveal members' behavior, it is also an indication of losing more information in anonymization for members who are declaring consecutive nearby locations. Thus, this could be another hypothesis for improvement of anonymization algorithms to preserve better utility in similar situations.

6.2 New Ideas for GSUVis

Our participants also discussed possible extensions for GSUVis to make this system more accessible to other privacy experts and data consumers. In addition, they discussed about scalability and applicability of our design for other data types in different context. They envisioned having the system running on a web server so that privacy experts or data consumers can upload and analyse their own data.

Two experts requested simultaneous comparison of two or more anonymizers on a single dataset. They pointed out that this feature could help them in situations when different anonymizers have the same privacy features yet different data utility outcomes. "*The most powerful feature of this system is the level of detail that I can get. I like to have options for choosing different anonymization algorithms, and see the effect of each.*"

Three participants also requested to link more external information for LBSN data in our system. For instance, they suggested that an embedded map for each glyph would offer more visual context when showing original and anonymized location data. As another example, linking a dataset of location and event types to LBSN data would be beneficial for the system users. "*. . . I like to see the location points for public transit, like the people who are in a train station. Does the anonymization still keep people in station or move them into the hotels and streets around it?"*

7 Future Directions

Scalability is a universal challenge for visualization and the rapidly increasing size of datasets continues to put more pressure of this. There are several things we have done in this regard and several more than will make interesting future work. From our close work with domain experts, we actually intensified this problem by agreeing to visualize both original and anonymized in the same view. This led to the choice of an arc diagram approach for the friendship relations and a timeline trajectory visualization for location information. These both take considerably less space than the more commonly used force-directed graph layout and map-based location approaches. In fact, this led to the successful combination of both visualizations in one view. We have also added interaction geared towards scalability issues. For instance, our visualization represents infinite check-in points in travel strands that are accessible with our sliding interaction feature. However, there are sometimes challenges with this interaction due to changing granularity of time distances between check-in points, which may cause occlusion or sparseness in some parts of travel lines at a given Time Slider scale. We are planning to apply space-folding techniques to resolve this limitation in representing travel lines. In addition, there are good scalability ideas such as aggregation, however, our experts specifically wanted the details to be directly available, so we have not at this point included aggregation. It is an interesting research challenge to discover how one might use aggregation to get a bit more scalability while still keeping with the experts' requests.

Current design is suitable only for the most fundamental structure of LBSN datasets, i.e. social relations and location points. However, LBSN systems can also carry more complex data including, members' comments, location reputations, discussion trends, etc. wherein current design should be extended to represent more complex data. Discovering ways to include this extra data richness could prove very beneficial. Also, in the future, we plan to extend GSUVis according to the new requirements that we collected from our experts. For instance, we currently rely on Gephi to assist the analyst in randomly selecting a set of nodes from a big social graph of LBSN data. We are planning to add this feature into our visualization tool.

Moreover, in GSUVis, the system design focuses particularly on LBSN data. It would be also an interesting future work to adapt GSUVis design for use in other scenarios such as the exploration of credit card transaction data, Netflix Prize dataset, phone call history, and other types of trajectory data with possible relationships between individuals.

8 Conclusions

This paper introduced GSUVis, a visual analysis tool designed and developed in collaboration with privacy experts to assist them in data exploration and analysis of original and anonymized LBSN datasets. To the best of our knowledge, GSUVis is the first visualization system that helps analysts to visually explore and compare original and anonymized LBSN data. We designed the system based on a set of interviews with domain experts. As a result of these interviews, we derived a set of tasks that an LBSN visualization system should support. We designed GSUVis as a combination of arc diagrams and linear temporal representation of location trajectories in a single view, to address as many collected tasks as possible.

We evaluated the effectiveness of GSUVis by conducting insight-based studies with privacy specialists. We received enthusiastic feedback from our participants and we reflect on the new insights that experts gained about the data. We categorized the insights into three main categories: people's behaviour, generating new hypothesis, and evaluating data utility. Furthermore, the experts suggestions for future improvement of GSUVis were based on their ideas that this could form the basis of a useful tool for them.

Acknowledgments

This research was supported in part by AITF, NSERC, GRAND, Surfnet, and SMART Technologies. We would like to thank our participants and reviewers for the expert knowledge they brought to this project.

References

[AAB*13] ANDRIENKO G., ANDRIENKO N., BAK P., KEIM D., WROBEL S.: *Visual Analytics of Movement*. Springer Berlin Heidelberg, 2013. 2

[ABN08] ABUL O., BONCHI F., NANNI M.: Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In *Proceedings of the IEEE 24th ICDE* (Washington, DC, USA, 2008), ICDE'08, pp. 376–385. 3

[ALY15] AGGARWAL C., LI Y., YU P.: On the anonymizability of graphs. *KAIS 45*, 3 (2015), 571–588. 2

[ASNB12] ASKARI M., SAFAVI-NAINI R., BARKER K.: An Information Theoretic Privacy and Utility Measure for Data Sanitization Mechanisms. In *Proceedings of the Second ACM CODASPY* (San Antonio, Texas, USA, 2012), ACM, pp. 283–294. 2, 3

[BETT98] BATTISTA G. D., EADES P., TAMASSIA R., TOLLIS I. G.: *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998. 4

[BEW95] BECKER R., EICK S., WILKS A.: Visualizing network data. *IEEE TVCG 1*, 1 (Mar 1995), 16–28. 2

[BHJ*09] BASTIAN M., HEYMANN S., JACOMY M., ET AL.: Gephi: an open source software for exploring and manipulating networks. *ICWSM 8* (2009), 361–362. 4

[BKC*13] BORGO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. *Eurographics State-of-the-Art Reports* (May 2013), 39–63. 7

[CM11a] CHOW C.-Y., MOKBEL M. F.: Trajectory Privacy in Location-based Services and Data Publication. *SIGKDD Explorations 13*, 1 (Aug. 2011), 19–29. 3

[CM11b] CORREA C., MA K. L.: Visualizing Social Networks. In *Social Network Data Analytics*. Springer, 2011, pp. 307–326. 2

[CML11] CHO E., MYERS S. A., LESKOVEC J.: Friendship and Mobility: User Movement in Location-based Social Networks. In *Proceedings of the 17th ACM SIGKDD* (San Diego, California, USA, 2011), KDD'11, pp. 1082–1090. 3

[CTH*10] CRANSHAW J., TOCH E., HONG J., KITTUR A., SADEH N.: Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM UbiComp* (Copenhagen, Denmark, 2010), UbiComp '10, pp. 119–128. 6

[DCW12] DORK M., CARPENDALE S., WILLIAMSON C.: Visualizing explicit and implicit relations of complex information spaces. *Information Visualization 11*, 1 (2012), 5–21. 2

[FWCY10] FUNG B. C. M., WANG K., CHEN R., YU P. S.: Privacy-preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys (CSUR) 42*, 4 (June 2010), 14:1–14:53. 2, 3

[GG03] GRUTESER M., GRUNWALD D.: Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st ACM MobiSys* (San Francisco, California, 2003), MobiSys'03, pp. 31–42. 1, 6

[HB05] HEER J., BOYD D.: Vizster: Visualizing Online Social Networks. In *IEEE INFOVIS'05* (Oct 2005), pp. 32–39. 2

[HF07] HENRY N., FEKETE J.-D.: Matlink: Enhanced matrix visualization for analyzing social networks. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part II* (Berlin, Heidelberg, 2007), INTERACT'07, Springer-Verlag, pp. 288–302. 2

[HFM07] HENRY N., FEKETE J.-D., MCGUFFIN M. J.: Nodetrix: A hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov. 2007), 1302–1309. 2

[Kru09] KRUMM J.: A Survey of Computational Location Privacy. *Personal Ubiquitous Computing 13*, 6 (Aug. 2009), 391–399. 3

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *IEEE TVCG 18*, 9 (Sept 2012), 1520–1536. 2

[LFC14] LOORAK M. H., FONG P., CARPENDALE S.: Papilio: Visualizing android application permissions. *Computer Graphics Forum 33*, 3 (2014), 391–400. 2

[LM14] LUO W., MACEACHREN A. M.: Geo-social Visual Analytics. *JOSIS*, 8 (May 2014), 27–66. 2, 5

[LT08] LIU K., TERZI E.: Towards Identity Anonymization on Graphs. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (Vancouver, Canada, 2008), SIGMOD'08, pp. 93–106. 1, 6

[Min65] MINARD C. J.: Carte figurative relative au choix de l'emplacement d'un nouvel hŹtel des postes de paris. In *ENPC*. 1865. 2

[MJ11] MASOUMZADEH A., JOSHI J.: Anonymizing Geo-social Network Datasets. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS* (Chicago, Illinois, 2011), SPRINGL'11, pp. 25–32. 2

[MJR*11] MACEACHREN A. M., JAISWAL A., ROBINSON A. C., PEZANOWSKI S., SAVELYEV A., MITRA P., ZHANG X., BLANFORD J.: Senseplace2: Geotwitter analytics support for situational awareness. In *VAST* (Oct 2011), pp. 181–190. 2

[MS05] MŁKINEN E., SIIRTOLA H.: The barycenter heuristic and the reorderable matrix. *Informatica (Slovenia) 29*, 3 (2005), 357–364. 5

[MYYR13] MA C., YAU D., YIP N., RAO N.: Privacy Vulnerability of Published Anonymous Mobility Traces. *IEEE/ACM Transactions on Networking 21*, 3 (June 2013), 720–733. 2

[NASG09] NERGIZ M. E., ATZORI M., SAYGŌN Y., GUC B.: Towards Trajectory Anonymization: A Generalization-Based Approach. *Trans. Data Privacy 2*, 1 (Apr. 2009), 47–75. 3

[SWW11] SCHEEPENS R., WILLEMS N. WETERING H., WIJK J.: Interactive visualization of multivariate trajectory data with density maps. In *IEEE PacificVis* (March 2011), pp. 147–154. 2

[TSAA12] TOMINSKI C., SCHUMANN H., ANDRIENKO G., ANDRIENKO N.: Stacking-based visualization of trajectory attribute data. *IEEE VCG 18*, 12 (Dec. 2012), 2565–2574. 2

[VLKS*11] VON LANDESBERGER T., KUIJPER A., SCHRECK T., KOHLHAMMER J., VAN WIJK J., FEKETE J.-D., FELLNER D.: Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. *CGF 30*, 6 (2011), 1719–1749. 2

[War13] WARE C.: *Information visualization: perception for design*. Elsevier, 2013. 6

[Wat02] WATTENBERG M.: Arc Diagrams: Visualizing Structure in Strings. In *Information Visualization (INFOVIS 2002)* (2002), pp. 110–116. 2, 4

[WSDR14] WERNKE M., SKVORTSOV P., DÜRR F., ROTHERMEL K.: A Classification of Location Privacy Attacks and Approaches. *Personal Ubiquitous Computing 18*, 1 (Jan. 2014), 163–175. 3

[WYLC10] WU X., YING X., LIU K., CHEN L.: A Survey of Privacy-Preservation of Graphs and Social Networks. In *Managing and Mining Graph Data*, vol. 40 of *Advances in Database Systems*. Springer US, 2010, pp. 421–453. 2

[YaKSJ07] YI J. S., AH KANG Y., STASKO J. T., JACKO J. A.: Toward a deeper understanding of the role of interaction in information visualization. *IEEE TVCG 13*, 6 (2007), 1224–1231. 4