Towards geoprivacy guidelines for spatial data

 http://doi.org/10.3929/ethz-b-000225618

Towards geoprivacy guidelines for spatial data

Ourania Kounadi* and Bernd Resch*

* Ourania Kounadi, Schillerstraße 30 5020 Salzburg - Austria, University of Salzburg - Department of Geoinformatics - Z_GIS, e-mail: ourania.kounadi@sbg.ac.at
* Bernd Resch, Schillerstraße 30 5020 Salzburg - Austria, University of Salzburg - Department of Geoinformatics - Z_GIS, e-mail: bernd.resch@sbg.ac.at

Abstract. This paper proposes an approach towards practical privacy guidelines for the different stages of a research effort that collects and/or uses "sensitive" spatial data. Specifically, we focus on: a) initial tasks as prior to starting a survey, b) storing, anonymization, and assessment of datasets, and c) actions to eliminate disclosure from published data and deliverables or when datasets are shared with third parties.

Keywords. location privacy, spatiotemporal data, spatial error

1. Introduction

Geoprivacy, location privacy, and spatial confidentiality are interrelated terms that explain privacy or confidentiality implications that are associated with "sensitive" spatial data (SSD). We use the abbreviation SSD for data types that contain a location attribute which is connected to private or confidential information (e.g. location of an individual at a specific time, heart rate of an individual at a specific location and time, residencies of breast cancer patients). Thus, SSD can be distinguished from other data because they can lead to a variety of disclosures (identity, inferential, or attribute disclosure). The means to prevent such disclosures comprise protection methods and measures, and other privacy-preserving tasks during a research effort. However, literature review studies in spatial and a-spatial disciplines have shown that over the last 20 years several practitioners have not employed appropriate privacy-preserving measures when publishing their findings (Haley et al. 2016, Kounadi & Leitner 2014). These studies reveal the need to educate the practitioners and to offer them a complete

privacy-preserving guidance for research efforts that use SSD. In fact, while previous research has mainly focused on methods to preserve privacy and measures to examine information disclosure, limited efforts exist for practical privacy-preserving guidelines regarding the collection, storage, analysis, and dissemination of SSD (Kounadi and Resch, in press).

Thus in this paper, we present concrete geoprivacy guidelines that can be used for the four types of SSD. The first one is c*onfidential discrete location data* (CDL) that have been discussed for health and crime geocoded datasets, but may also refer to confidential information that is not about individuals. CDL are microdata that contain one measurement per data subject and several attributes. The attributes are the location, quasi-identifiers, and at least one confidential or sensitive attribute. The second type is s*patiotemporal trajectories of individuals (STI)*, which refers to data obtained from location based services and mobile phone data. STI data contain multiple measurements per data subject and at least the spatial and temporal attributes, which alone may lead to inferential disclosure. The third type is l*ocation based social network data* (LBSN) from which spatiotemporal trajectories of individuals can be inferred along with additional attributes that are inherent in such applications. For example, the text attribute on Twitter may lead to inferential disclosure of personal preferences, opinions, and other private matters. The fourth type is p*articipatory sensing data (PSD)* that are collected from mobile sensing applications (Resch 2013). PSD may be perceived as a combination of CDL and STI. PSD contain multiple measurements per data subject, the spatiotemporal attributes, at least one confidential or sensitive attribute, and possibly other quasi-identifiers.

2. Geoprivacy guidelines

We focus on three categories of guidelines; namely the a) initial tasks as prior to starting a survey - "*pre-survey activities*", b) storing, anonymization, and assessment of datasets - "*processing and analysis of data*", and c) actions to eliminate disclosure from published data and deliverables or when datasets are shared with third parties - "*disclosure prevention*". The categories are analysed in the next subsections and summarised at the end of each subsection in the form of a table (Tables 1, 2,3). We deliberately exclude aspects of data-related IT security as this is a more generic issue, which is also illustrated by Kounadi and Resch (in press).

2.1. Pre-survey activities

Activities prior to the start of a survey involving SSD have not been widely discussed in geoprivacy literature. However, good organisation and planning at this stage will ease the activities to follow and, minimise the disclosure risk, and ensure that the study is in line with ethical standards. First, the study has to be designed in a way that minimises disclosure risk and reported in a research plan. The plan should contain criteria for access to restricted-access data (i.e. data subjects can be distinguished among each other), if the controller considers to share them. Also, a participation agreement must be prepared that has additional insertions with respect to the location privacy disclosure risk. It is important to communicate the risk in an understandable way since participants may not be familiar with spatial data. Last, the research plan and the participation agreement should be approved by a disclosure review board (DRB) from objective and experienced stuff of the institution or by an external DRB, if necessary.

Pre-survey activities
1. Design study with the least privacy invasive manner
2. Develop a privacy-preserving research plan
3. Define criteria for access to restricted-access data sets
3. Prepare a participation agreement
4. Ensure inform consent on location privacy disclosure risks
5. Obtain institutional approval preferably reviewed from a DRB

Table 1: Guidelines prior to the starting of the survey.

2.2. Processing and analysing data

The first essential step to protect SSD is to remove direct identifiers from the dataset such as names, email addresses, telephone numbers or others. Then, the anonymization should take place. At this stage, each candidate attribute of the dataset has to be examined regarding the increment of disclosure risk it involves. In particular, pseudonyms and quasi identifiers should be carefully analysed as they may lead to inferential and attribute disclosure. Pseudonyms can be used to divide the datasets into subsets of individuals. The subsets can be spatially analysed to infer home or other private addresses and then linked to other sources to infer the identities of the subjects (Krumm 2007). Quasi identifiers or data collection metadata can be used in a similar way to identify a subset that can be linked to an individual. Furthermore, SSD that contain multiple values in their confidential attributes (e.g. in PSD) should ensure that the values are "well-represented" within the anonymised groups according to the l-diversity notion (Machanavajjhala et al. 2007). Additional tasks are to classify datasets as "restricted-access" (i.e. identifiable data) or "anonymised", to es-

timate the disclosure risk and to assess the analytical accuracy of the anonymised dataset for future spatial analysis.

Processing and analysing data
1 Remove identifiers from the dataset
2. Ensure that the inclusion of pseudonyms and quasi-identifiers does not lead to disclosure
3. Ensure a sufficient l-diversity of the sensitive attributes
4. Classify each data set as a restricted-access or anonymised data set
5. Assess disclosure of anonymised datasets
6. Assess anonymization effect on spatial analysis

Table 1: Guidelines to store, anonymise, and asses derived datasets.

2.3. Disclosure prevention

When research findings derived from SSD are published or shared within the scientific community they are typically presented in the form of a map or a table. A convenient and easy way to minimise the disclosure risk of data subjects is to reduce the spatial and/or the temporal precision of the presented data. Rather frequently researchers opt for a point distribution map although it entails privacy implications. Research has shown that points on a map can be re-engineered with considerable accuracy regardless of the map's resolution (Brownstein et al. 2006). Hence, point distribution maps of original data should be avoided and replaced by heat maps, thematic maps, or point maps from masked data. The latter approach requires to anonymise ("mask") the data before plotting them on a map. In this case, information regarding the anonymization method as well as an assessment of the disclosure risk should accompany the "sensitive" deliverable (e.g. a message included in the legend). Last, the use of disclaimers may reduce liability of research outputs and eliminate ambiguous interpretations.

Furthermore, SSD may be released by the institutions to the scientific community as processed anonymised subsets (e.g. on a webpage of a research project, between research groups, or in a scientific journal). However, for techniques such as geomasking and aggregation the disclosure risk increases with multiple releases of anonymised copies of the original data and with metadata information (Cassa et al. 2008, Zimmerman & Pavlik 2008). Thus, we suggest avoiding such practices because they may provide hints to a potential attacker to re-identify the original data. Additional tasks are to provide information on the protection method, disclosure risk, and analytical accuracy of the anonymized dataset.

The last group of guidelines refers to data sharing practices with other institutions or members of the same institution that are not part of the research group that collected the SSD. The first step is to prepare a licensing agreement for the requestor or a disclosure sharing agreement for restricted-access data. Especially for the case of restricted-access data, the controller or a designated privacy manager should ensure the requestor's

appropriateness to handle sensitive personal data. This means that a secure environment is established and the personnel that will use the data is aware of possible geoprivacy threats and protection measures. Also, the modalities of storage, analysis, and publication of data should be pre-decided by the controller and the requestor. Last, there should be an inventory of all disclosed activities such as publications and released datasets.

Disclosure prevention
Dissemination of findings
1. Reduce spatial & temporal precision
2. Consider alternatives to point distribution maps
3. Assess disclosure on a point distribution map
4. Provide protection vs disclosure information
5. Use disclaimers
Anonymised Datasets
6.Avoid the release of multiple versions of anonymised datasets and metadata information
7. Provide information on protection, disclosure risk, and effect
8. Maintain log of anonymised disclosed datasets
Data sharing with third parties
9. Plan a mandatory licensing agreement
10. Check on requestor's safe settings (personnel and environment)
11. Have an active role regarding the privacy-preservation practices of the requestor
12. Maintain log of restricted-access disclosed datasets

Table 3: Guidelines to prevent disclosure.

3. Future work

Previous research in geoprivacy focused on the development of methods to preserve privacy and measures to assess information disclosure. In contrast, this paper presents a practical set of privacy-preserving guidelines for the collection, storage, analysis, and dissemination of individual measurements from mobile participatory sensing applications. The next step of this work is to develop **a comprehensive set of guidelines**, taking legal and ethical regulations into account, that can be applied to different research efforts. The similarities and differences between SSD types will be examined so as to distinguish and specify the guidelines by the type of dataset. For example, the concept of L-diversity is important for CDL and PSD but not for STI. Also, CDL are typically captured by official authorities and institutions and then may, or not, be given to other institutions for research purposes (e.g. a police department shares crime data to a researcher for a single study purpose). This means that there might be additional restrictions specific to the body that shares the data to be taken into account.

Possible **disclosure risk scenarios** will be examined and the guidelines are intended to serve as a mean to prohibit such disclosures. Ad-

ditionally, two more sources of information will be used. First, the limitations and benefits of available **anonymization approaches** and second the **experts' suggestions** to minimize or eliminate the disclosure risk. Also, it is important to **identify the stakeholders** that are involved in each research effort and assign them privacy-preserving responsibilities.

Research efforts entail privacy implications throughout their different stages although some of these stages are yet to be addressed in the literature of location privacy. For instance, an issue that has not been discussed is how to adapt and improve existing templates of participation agreements when conducting surveys so as to ensure inform consent on location privacy disclosure risks. This requires the description of **geoprivacy insertions within the confidentiality statement** that are easily communicated to non-experts. Another topic that lack discussion and precise guidelines are the requirements that ensure **a secure environment** for collecting, storing, and sharing data (includes personnel, IT system, and devices).

References

Brownstein, J S, Cassa, C A, Kohane, I S, & Mandl, K D (2006) An unsupervised classification method for inferring original case locations from low-resolution disease maps. International Journal of Health Geographics, 5(1), 56

Cassa C A, Wieland, S C, & Mandl K D (2008) Re-identification of home addresses from spatial locations anonymized by Gaussian skew. International Journal of Health Geographics, 7 (45)

Haley D F, Matthews S A, Cooper H L, Haardörfer R, Adimora A A, Wingood G M, & Kramer M R (2016) Confidentiality considerations for use of social-spatial data on the social determinants of health: Sexual and reproductive health case study. Social Science & Medicine, 166, 49-56

Kounadi O, Resch B (2017) A Geoprivacy by Design Guideline for Research Campaigns that use Participatory Sensing Data. Journal of Empirical Research on Human Research Ethics, (in press)

Kounadi O, Leitner M (2014). Why Does Geoprivacy Matter? The Scientific Publication of Confidential Data Presented on Maps. Journal of Empirical Research on Human Research Ethics, 9(4), pp.34-45

Krumm, J (2007) Inference attacks on location tracks. In A. LaMarca, M. Langheinrich, & K. Truong (Eds.), Pervasive Computing (Vol. 4480, pp. 127-143) Berlin Heidelberg: Springer

Machanavajjhala A, Kifer D, Gehrke J, & Venkitasubramaniam M (2007) l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 1(1), 3

Resch, B (2013) People as sensors and collective sensing-contextual observations complementing geo-sensor network measurements Progress in location-based services (pp. 391-406): Springer

Zimmerman, D L, & Pavlik, C (2008) Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. Geographical analysis, 40(1): 52-76