Bounded Privacy: Formalising the Trade-Off Between Privacy and Quality of Service

Lukas Hartmann[1]

Abstract: Many services and applications require users to provide a certain amount of information about themselves in order to receive an acceptable quality of service (QoS). Exemplary areas of use are location based services like route planning or the reporting of security incidents for critical infrastructure. Users putting emphasis on their privacy, for example through anonymization, therefore usually suffer from a loss of QoS. Some services however, may not even be feasible above a certain threshold of anonymization, resulting in unacceptable service quality. Hence, there need to be restrictions on the applied level of anonymization. To prevent the QoS from dropping below an unacceptable threshold, we introduce the concept of *Bounded Privacy*, a generic model to describe situations in which the achievable level of privacy is bounded by its relation to the service quality. We furthermore propose an approach to derive the optimal level of privacy for both discrete and continuous data.

Keywords: Privacy; Quality of Service; Modelling Anonymity

1 Introduction

In the last years, methods to collect and process personal data have been dramatically improved and are widely used nowadays. Some of these systems could be used for tracking individuals or to obtain more personal information by aggregating existing data. As a consequence, some users wish to reveal as little personal data as possible and to stay private. One widely used approach is the anonymization of sensitive data so that the reported data cannot be easily mapped to a specific subject. For some services however, exactly this sensitive data may be actually necessary for operation.

Location information can be highly sensitive, revealing not only specific locations like home or work, but also for example religious or political views. By only knowing the home and work location of a subject, one can reveal the identity of an anonymous individual with a very high probability [GP09]. If location information is tracked over time, one can infer even more personal data of this user. Therefore, *Location Privacy* is very important for privacy-conscious users. Location Privacy is defined as a state when the location of a subject is not revealed to other subjects. A widely-used approach is the obfuscation of the exact geographical location and consequent reporting of the obfuscated information.

[1] Universität Regensburg, Lehrstuhl für Wirtschaftsinformatik IV, 93040 Regensburg, lukas.hartmann@ur.de

doi:10.18420/sicherheit2018_23

Nevertheless, location based services (LBS) require location data to work properly and they require a certain precision in this data to provide their service with an acceptable level of quality. For mobile route planning with a feasible quality, a precise geographical information for start and target location has to be submitted by the user to the LBS. Reporting highly anonymized geo-coordinates would not be sufficent and would result in a unacceptable service quality. The routing app would be useless. On the contrary, there are apps for which a rough location information would be sufficient. A weather app only needs the city or at most the postal code, but no precise address information. When using LBS, there is always a trade-off between the quality of service (QoS) and the amount of location privacy a user can obtain. Depending on the individual service and use case, the user has to accept less privacy, if appropriate quality should not fall below a certain threshold.

Stipulated by German law, operators of critical infrastructure in Germany have to report security incidents to the Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik, BSI). Critical infrastructure in this context is divided in different industrial sectors covering "classicalïnfrastructure like transportation, energy and water, but also IT infrastructure, cultural sites and media. The law permits however that incidents can be reported with related data partially anonymized, so that the operator does not have to reveal too much information about the affected business. This anonymization could be done by for example by obfuscation of the geographical location, the time when the issue arose or by a generalization of the infrastructure category. In any case, the reporting should still contain enough information so that the incident can be treated by the BSI in an appropriate way. If a coal power plant might suffer from a coal shortage due to issues with the internal enterprise resource planning system, it is possibly sufficient to report the affected region with a certain precision. However, if an IT incident leads to a critical damage to a power plant, more precise information of the actual power plant is needed. In the scenario of incident reporting, there is always a trade-off between the quality of reporting and the amount of anonymization/privacy one can obtain. Depending on the actual incident, one has to use less anonymization in order to provide enough information in the incident report.

In both scenarios we have a trade-off between anonymization and the quality of service, where the achievable level of privacy is bounded by its relation to the service quality. To prevent the QoS from dropping below an unacceptable threshold, we introduce the concept of *Bounded Privacy* describing situations when the quality of anonymization is bounded by an upper bound. We present a generic model to this problem and introduce methods how to obtain Bounded Privacy depending on the given information. To the best of our knowledge, this is the first paper which investigates privacy when the corresponding service shall have a decent, pre-defined minimum quality.

2 Related Work

Several methods have been proposed in the literature to anonymize location information. The probably best discussed concept is the notation of *k-Anonymity* introduced

by Sweeney [Sw02]. An extension of this model called *l-Diversity* was introduced by Machanavajjhala et al. [Ma07]. In a similar vein, Ağır et al. [Ağ16] try to mask semantic information about visited locations by generalizing it along a hierarchical semantic tree. Andrés et al. [An13] propose the idea of *ε-Geo-Indistinguishability* which attempts to restrict the information leakage to an observing adversary by employing a perturbation mechanism that obfuscates real location information in a probabilistic way. This approach is based on the concept of *Differential Privacy* introduced by Dwork [Dw11], originally proposed for statistical databases.

Literature shows a great variation in exchange formats used for the reporting of IT security incidents. The most recent overview on this topic was presented by Menges and Pernul [MP18] focusing on the on applicability of exchange formats for IT security incidents.

3 Bounded Privacy

As stated in the introduction, there is often a trade-off between privacy and the quality of service. This leads to scenarios where a decent service quality has to be guaranteed and thus the anonymization level has to be bounded. In this chapter, we introduce the approach of *Bounded Privacy* leading to a feasible level of anonymization, that allows for the given threshold on the service quality.

In general, an information consists of geographical coordinates and semantic information, for example location tags, infrastructure categories, etc. Therefore we model the information I as an element of the *domain*

$$\mathcal{D} = \mathbb{R}^2 \times \mathcal{S},$$

where the geographical information is given as a pair $(x, y) \in \mathbb{R}^2$ of coordinates and the semantic information $s \in \mathcal{S}$ comes from a generic set. In most scenarios, $\mathcal{S}$ would be a multi-dimensional space where each dimension would represent one semantic attribute. A semantic information $s \in \mathcal{S}$ would be of the form $s = (s_1, s_2, \ldots)$ where s_1 could for example be a category (“energy“), s_2 could be an impact category (“high impact“), etc.

When we anonymize an information to obtain privacy, we apply an *anonymization function*

$$AN\colon \mathcal{D} \to \mathcal{P}(\mathcal{D}),\ I \mapsto AN(I)$$

to the original information I, consisting of the functions $AN_{geo}\colon \mathcal{D} \to \mathcal{P}(\mathbb{R}^2)$ and $AN_{sem}\colon \mathcal{D} \to \mathcal{P}(\mathcal{S})$ for obfuscating geographical and semantic information, respectively. The anonymized or obfuscated information $AN(I)$ is an element of the power set $\mathcal{P}(\mathcal{D})$ and hence a subset of the information domain $\mathcal{D}$, since common anonymization techniques use cloaking mechanisms to obtain privacy. The most popular example for this is k-Anonymity (see section 2).

For measuring how much two (possibly anonymized) pieces of information differ from each other, we introduce a generic *information distance* which is given by

$$d\colon \mathcal{P}(\mathcal{D}) \times \mathcal{P}(\mathcal{D}) \to \mathbb{R}^2_{\geq 0}.$$

It calculates both the geographical and semantic difference as non-negative numerical values. Analogous to the anonymization function AN, the information distance d consists of two functions d_{geo} and d_{sem} to measure the geographical and semantic distance, respectively.

The threshold below which the quality of service shall not fall - and as a consequence the bound on the level of privacy - is modelled via a *restriction function* r consisting of $r_{geo}, r_{sem}\colon \mathcal{D} \to \mathbb{R}$. With this function, we can define when an anonymization is valid with respect to the given threshold restrictions: An anonymization of a given information I is valid if

$$d(\{I\}, AN(I)) \leq r(I)$$

i.e. if both restrictions $d_{geo}(\{I\}, AN_{geo}(I)) \leq r_{geo}(I)$ and $d_{sem}(\{I\}, AN_{sem}(I)) \leq r_{sem}(I)$ on the information distance between original information I and anonymized information $AN(I)$ are fulfilled.

Depending on the use case and the individual information, one could have different restrictions which level of anonymization can be applied. As the level of allowed restrictions is dependent on the information quality necessary for the lower bound on the quality of service, its definition should be determined at least in part by the service provider. The restriction function r is also dependent on certain parts of the information I. Following the previously used example from section 1 comparing between a coal shortage and a nuclear melt-down, this means, that the restriction function $r(I)$ would have a smaller value in the latter case although both examples come from the same domain. One can see that the function r can have a significantly different figure within even a single scenario.

We call the concept *Bounded Privacy*, when the level of anonymization is bounded by a minimum level of service quality and therefore valid anonymization with respect to the given restriction function is necessary.

4 Methods for Obtaining Bounded Privacy

In the generic model from section 3, an information I consists of data from discrete and continuous domains. Geographic information is normally given by continuous geo-coordinates $(x, y) \in \mathbb{R}^2$, whereas semantic data is normally given as a tuple of discrete values, like categories, sectors or location tags.

To implement Bounded Privacy for discrete data, a bottom-up based tree approach could be used: The discrete information units represent the leafs of a tree. Related units can be grouped together and get a common parent node in the tree, representing the aggregated

information of its child nodes within a topical hierarchy. Instead of reporting the exact semantic information on the leaf level, one would report the aggregated information given by the parent nodes to anonymize the exact information. The further away from the leaves we move, the less detailed information is reported and therefore the anonymization is better. Following the example of critical infrastructure, categories like "Nuclear Power Plant" and "Coal Power Plant" could be leafs of the semantic tree and could be grouped together to the obfuscated category "Fossil Energy Power Plant". This parent category could be anonymized to the more general category "Power Plant" or even to "Energy". There are already efforts to create similarly structured semantic ontologies for different sectors that could be used as a basis for such topical hierarchies. This is an ongoing topic of research and could prove an important piece to bridge the gap between our theoretical model and practical applicability [SS10]. For measuring the information distance d, one could use the graph metric d_G which calculates the length of the shortest path between two given nodes. If a restriction $r_{sem}(I)$ is applicable for a given information I, it is only allowed to step that far in the semantic graph such that $d_G(I, AN_{sem}(I)) \leq r_{sem}(I)$.

For the anonymization of data derived from a continuous domain, an approach based on the concept of *ε-Geo-Indistinguishability* could be used. Originally, this mechanism was proposed for geographical data only, but one can adapt it for other domains of continuous data as well. The main idea of ε-Geo-Indistinguishability is the following: With a perturbation mechanism K, a geo-coordinate $(x, y) \in \mathbb{R}^2$ is mapped to another point (x^*, y^*) in a "probabilistic way". The obfuscated point will then be reported instead of the original location. Andrés et al. [An13] propose to use a two-dimensional Laplacian distribution which creates noise and obfuscates the geographical location. This mechanism can be used as anonymization function AN_{geo}. The distribution of the Laplacian noise highly depends on the parameter ε and influences the level of perturbation. Therefore, the restriction function r_{geo} has to restrict the distribution of the Laplacian noise so that the obfuscated points are not too far away from the original geographical location with a high probability. Since the anonymized geographical-information is given in probabilistic way, the information distance d_{geo} must also be measured probabilistically. A suitable approach for the restriction function r_{geo} would be to introduce a radius s around the original geo-coordinate (x, y) in which the obfuscated point (x^*, y^*) should be mapped with at least probability p. This probability would be dependent on the information I so that in some scenarios higher probabilities could be used as in others. In order to keep our model deterministic, we propose to redraw the obfuscated point (x^*, y^*), if the distance to the original point (x, y) is higher than allowed by our restriction function, to ensure that the QoS cannot drop below the specified bound.

5 Conclusion

We presented a generic model to the problem when privacy is bounded such that a minimum Quality of Service has to be guaranteed. We introduced generic anonymization functions and

tools for measuring and restricting the information distance between original and anonymized information. Furthermore, basic approaches on how to apply these concepts to discrete and continuous data were given. For the discrete domain, we used a tree-based approach in which a broader semantic tag/category shall be reported as long as this is valid with respect to the restriction function. A method which is based on ε-Geo-Indistinguishability was used for continuous geographical data. Noise was added to the original location following a planar Laplacian distribution.

In future work, it is envisioned to evaluate the generic model of *Bounded Privacy* on a real data example and define restriction functions for a concrete scenario like incident reporting. Furthermore, we want to extend the generic model so that upper bounds on the provided information are included as well. Another possible extension of the model would be the introduction of an evaluation mechanism for the reported information so that for example the monetary value of the anonymized information could be measured. With this approach, one could also investigate in which circumstances it might have a positive effect for a subject to report more information than needed, if this reporting has some benefits. Based on these two extensions, one could even quantitatively formalize the trade-off between privacy and QoS, thus enabling users to pick the optimal privacy level for their needs and ultimately realizing their informational self-determination.

References

[Ağ16] Ağır, Berker; Huguenin, Kévin; Hengartner, Urs; Hubaux, Jean-Pierre: On the Privacy Implications of Location Semantics. Proceedings on Privacy Enhancing Technologies, 2016(4):165–183, 2016.

[An13] Andrés, Miguel E; Bordenabe, Nicolás E; Chatzikokolakis, Konstantinos; Palamidessi, Catuscia: Geo-indistinguishability: Differential privacy for location-based systems. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. ACM, pp. 901–914, 2013.

[Dw11] Dwork, Cynthia: Differential privacy. In: Encyclopedia of Cryptography and Security, pp. 338–340. Springer, 2011.

[GP09] Golle, Philippe; Partridge, Kurt: On the anonymity of home/work location pairs. In: International Conference on Pervasive Computing. Springer, pp. 390–397, 2009.

[Ma07] Machanavajjhala, Ashwin; Kifer, Daniel; Gehrke, Johannes; Venkitasubramaniam, Muthuramakrishnan: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):3, 2007.

[MP18] Menges, Florian; Pernul, Günther: A comparative analysis of incident reporting formats. Computers & Security, 73(Supplement C):87 – 101, 2018.

[SS10] Staab, Steffen; Studer, Rudi: Handbook on ontologies. Springer Science & Business Media, 2010.

[Sw02] Sweeney, Latanya: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.