Towards Flexible K-Anonymity

Rima Kilany[1(✉)], Maria Sokhn[2], Hussein Hellani[3], and Shaban Shabani[2]

[1] Université Saint Joseph, B.P. 11-0514 Riad El Solh, Lebanon
rima.kilany@usj.edu.lb
[2] HES-SO Valais Wallis, Technopole 3, 3960 Sierre, Switzerland
maria.sokhn@hes-so.ch, shaban.shabani@hevs.ch
[3] Université Saint Joseph, Bir Hassna, B.P. 13-6007 Beirut, Lebanon
hussein.hellani@hotmail.com

Abstract. Data published online nowadays needs a high level of privacy to gain confidentiality as well as to maintain the privacy laws. The focus on k-anonymity enhancements along the last decade, allows this method to be elected as the starting point of any research. In this paper we focus on the external anonymization through a new method: the « Flexible k-anonymity » . It aims to anonymize external published data, by defining a semantic ontology that distinguishes between sparse and abundant quasi-identifiers, and describes aggregation levels relations, in order to achieve adequate k-blocks. For the validation of our proposal, we apply the aforementioned anonymization method to the Comiqual dataset. Comiqual (Collaborative measurement of internet quality), is a large-scale measurement platform for assessing the internet quality access of mobile and ADSL users by collecting mobility traces and private data related to internet metric values.

1 Introduction

In order to have an efficient and useful anonymization process, we should consider data sanitization and refinement at two levels: First, the internal level, where the threat is mainly linked to employees or intruders. Indeed, they have been entrusted with authorized access to the network and can easily reach data repositories and violate individuals' privacies. Second, the external level, which mainly addresses published data, and attacks from people outside the organization.

In this paper we focus on the external level challenges by proposing an approach based on the k-anonymity principle. With the k-anonymity principle, the records are made indistinguishable from at least k-1 other records [1]. For this purpose, quasi-identifiers are examined for each record and a k-block is constructed in order to release them to the public. Quasi-identifiers are fields which, when combined, make a record unique and identifiable. Two methods are used to achieve the k-anonymity: *generalization* which substitutes the values of a given attribute with more general values and *suppression* which is used to mask the given information totally by an asterisk "*" and to moderate the generalization process when tuples with less than k-blocks occur [3].

A.-C. Ngonga Ngomo and P. Křemen (Eds.): KESW 2016, CCIS 649, pp. 288–297, 2016.
DOI: 10.1007/978-3-319-45880-9_22

For the validation of our proposal we used the Comiqual dataset. Comiqual (Collaborative measurement of internet quality)[1], is a large-scale measurement platform for assessing the internet quality access of mobile and ADSL users by collecting mobility traces and private data related to internet metric values.

The main contribution of this paper may be summarized as follows:

- The definition of a semantic ontology which distinguishes between scanty and abundant quasi-identifiers, by defining different classes for those identifiers as well as aggregation level relations between them.
- Flexible k-anonymity: a new anonymization method to be applied at the external level, by inferring aggregation levels from the ontology in order to be able to use different k-anonymity values and build appropriate k-blocks.
- The proposition of a complete anonymization process from the receipt of the data until its publishing.

The rest of this paper is structured as follows: Sect. 2 presents the related works, Sect. 3 gives an overview of Comiqual, Sect. 4 details the external anonymization approach, Sect. 5 presents the sanitization process, and finally, Sect. 6 concludes the paper.

2 Related Work

Most existing work on privacy, considered only location privacy of published data or what is called "second use of data" and employed k-anonymity based methods. In this section we will briefly explain the k-anonymity method [1] and some of its variants.

2.1 K-Anonymity Enhancements

The k-anonymity privacy protection achieved by L. Sweeney since 2002 [1] using generalization and suppression tools, opened the door to many researchers to add more enhancements or propose new methods based on the k-condition. In fact, most researchers [1, 8, 9] assume a uniform case study such as the medical dataset which focuses on grouping the attributes into personal identifiers (e.g. name, email address, etc.), quasi-identifiers (e.g. age, zip code, etc.) and sensitive attributes (e.g. disease) to apply the k-anonymity where each record in the same quasi-identifier block is indistinguishable from at least (k-1) other records within the same block [1]. The larger the value of k, the better the privacy is protected [2].

K-anonymity alone does not ensure privacy when sensitive values in an equivalence class lack diversity, which is known as the homogeneity attack. L-diversity [13] method is found to bridge this gap and is composed of three progressive levels: (1) distinct l-diversity, where each equivalent class has at least L values for each sensitive attribute but this doesn't prevent the probabilistic inference attacks. (2) in entropy l-diversity, each equivalence class must not only have enough different sensitive values, but the

[1] http://comiqual.usj.edu.lb/.

different sensitive values must also be distributed evenly enough. (3) the recursive l-diversity, makes sure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely. Obviously these methods are too restrictive and require a specific distribution of the data values. The main issue with l-diversity is that it does not consider semantic meanings of sensitive values. This leads to a less conservative notion of l-diversity. T-closeness [14] is a refinement of l-diversity and it aims to create equivalent classes that resemble the initial distribution of attributes in the table. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. The case studies that are elected to apply the anonymization methods, fit the researchers' purpose and prove high privacy protection. But what if we change the case study, do we still obtain the same results?

Three main issues are in common with the aforementioned methods: (1) they do not consider semantic meanings of sensitive values. (2) they are applied on sensitive attributes only, hence in our case these methods cannot be useful as there are no sensitive attributes in the dataset. (3) it is very difficult to build appropriate blocks in a sparse environment with scanty attribute values.

2.2 External Protection

Prior to the release of the "second use of data" version, the rule of thumb is to group the records into k similar blocks in order to satisfy the k-anonymity. Generalization and suppression are used for this purpose. According to Sweeny, in some cases, the price of removing an isolated record would be less than the price to pay in terms of information precision loss when generalizing all its possibly related data records [2]. This is not the case of the Comiqual dataset where suppression is forbidden and users should be able to view every and each measurement record sent to the server on the public website, using any device and from any location. According to the aforementioned analysis concerned with anonymization methods applied on sensitive attributes, some kind of threats such as homogeneity attacks become impractical to be applied on trajectory datasets. This is because this type of data distribution has no sensitive attributes and the sensitivity is embedded within the quasi-identifiers themselves such as location attributes. On another hand, traceability attacks with some background knowledge are very potential with any LBS system, where individual movements are disclosed using time-referenced location information as quasi-identifiers. The adversary may already know some portion of the trajectory of an individual in the dataset and may be interested in the rest (e.g. adversary knows that a particular person lives in a particular house. He also knows that she leaves the house and comes back home at specified times, so he may be interested in finding the locations she visited.). Standard generalization could not achieve good enough result in collecting the best k tuples and deceive the attackers to single out an individual, in other words, generalizing data in a non-smart manner leads to traceability attacks and sometimes cannot succeed in building an appropriate k-block in a sparse data environment (e.g. five different device models).

Due to the weakness of the above k-anonymity's enhancements in fixing all the shortcomings, introducing a semantic ontology system becomes necessary to let artificial

intelligence control the whole anonymization process. Ontologies allow us to model concepts, their relationships and properties as well as other more subtle aspects of a domain. The idea is to add an ontology layer on top of the k-anonymity method in order to create a more robust privacy-enforcing system [3]. Vocabulary k-anonymity method [6] perceived the extremely sparse data of web query logs, and proposed an algorithm to cluster vocabularies by semantic similarities. Such methods do not apply well on measurement applications like Comiqual, because these measurement platforms do not store any sensitive attributes.

3 Comiqual Overview

In Comiqual, users send measurement details periodically. These details include username, email address, machine type, installed operating system, battery status, cell info, cell id, GPS location, IP address, ISP provider, and network type. Comiqual mobile agent (MA) manages and controls the measurement process between the Comiqual server and the peer server that examines the measurement speed (e.g. michigan.mlab2.lca01.measurement-lab.org). MA sends the results back to the internal Comiqual database to be published on the website. Each participant's mobile has its own id represented by a combination of user's email address and mobile IMEI (International Mobile Station Equipment Identity). This unique id called Measurement Agent ID (MAID) is associated with each single measurement record generated by this user. Changing any of the mobile device or the email address by the user, leads to the generation of a new MAID.

In the use case of the Comiqual dataset, MAID, IMEI and email address can be considered as Personal Identifiable Information (PII). The quasi-identifiers are: GPS location, cell id, device model, ISP provider and network type. While username and email could be simply hashed when being stored on the internal servers, majority of the mentioned quasi-identifiers combined together along with the location attributes, could lead to single out an individual by detecting a certain mobility pattern or presence pattern and therefore should be wisely anonymized, on the internal and the external level.

Comiqual main constraints are: (1) the dataset does not include sensitive attributes. The sensitivity is embedded within the quasi-identifiers themselves as shown in Table 1 where for example the device model and the GPS location are the most two sensitive attributes in comparison to others. (2) Suppression is not allowed in any of the internal and external anonymization levels in order not to lose any collected measurement detail. (3) Generalization can be applied at the external level only in order not to lose accuracy of the collected data. It is important to mention that Comiqual dataset belongs to a family of broadband measurement applications as well as many similar applications like weather detection, prayer timings, etc. that are very trendy nowadays. A huge number of online participants are continuously sharing personal information on public sites (GPS locations, device model, etc.). Therefore the main purpose of this research is to foil the traceability of those users and protect them against many privacy attacks.

Table 1. Comiqual data representation

Quasi identifiers				Non-sensitive
Device model	GPS location	ISP	Net type	Result (Mbps)
Samsung-Galaxy	L1	ISP1	DSL	R1
IPhone 6	L2	ISP2	4G	R2
HTC one	L3	ISP3	3G	R3

Comiqual has sparse data represented by the "mobile model" which is very difficult to be grouped into k-block of similarity. Indeed there are different mobile series that fall under each brand. The rapid development and competency between mobile vendors is introducing more sparse data in the model since we have new brand series often. This added more challenges to the grouping of similar items during the anonymization process because of the existence of sparse data that does not satisfy the k-condition. Standard generalization could not achieve very good result in collecting the best k tuples and deceive the attackers to single out an individual. In other words, generalizing data via non-smart manner leads to traceability attacks and sometimes cannot succeed the k-block in a sparse data environment (e.g. five different device models). Due to the weakness of the k-anonymity's enhancements in fixing the sparse data issue, introducing a semantic ontology system becomes necessary to let the artificial intelligence control the whole anonymization process.

4 Process Overview

Simply removing the identifiers of individuals or replacing them by a pseudonym does not protect their privacy from inference attacks. Indeed several works, among which the one of Gambs et al. [15], demonstrate that a reverse analysis over the data may allow the identification of a person. Hence, it is important to combine the hashing aspect with the injection of some noise. The process is divided into two phases: the internal named the sanitization process, and the external process described in details in Sect. 5.

At each stage of the sanitization process (Fig. 1), we propose and apply an appropriate anonymization method in order to alleviate the specific attack risks associated to the data state at each level. The steps of the process are applied in the following order:

1. Hash the PII of each entry in order to avoid direct identification.
2. Encrypt the hashed PII used to filter the true data
3. Encrypted PII will be stored in a secure text file
4. → 7 Add fake records with fake PII - that are hashed and encrypted - until satisfying the k-anonymity (we are currently working on the enhancement of this series of steps that will be applied at the internal level)
8. The data that is composed from fake and real records will be published to the internal database and the temporary location will be cleared after a t time.
9. Filter the real records by means of true PII and preserve them in a temporary database

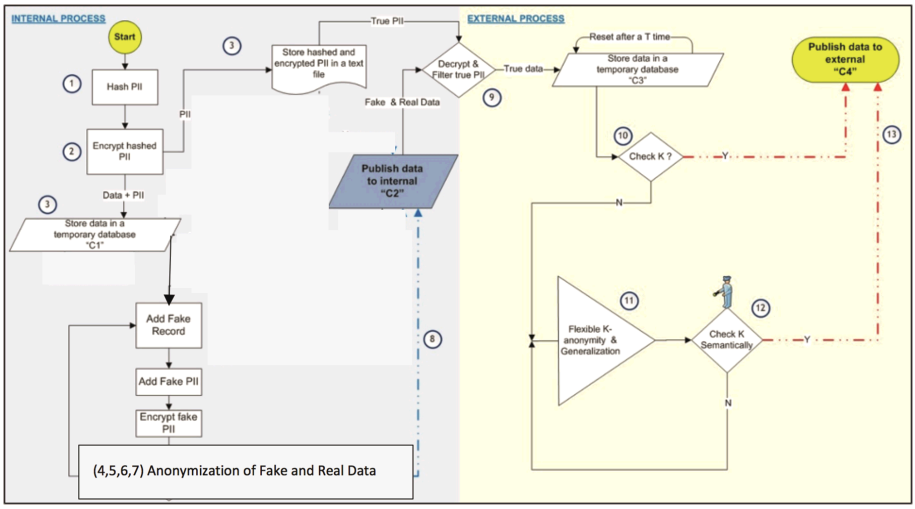

Fig. 1. Sanitization process

10. Check k satisfaction, publish immediately to external database if records' number satisfy k-condition, and the temporary location will be cleared after a t time
11. If the records' number is less than k, flexible k-anonymity will be applied to differentiate sparse and non-sparse attribute and enforce K_S and K simultaneously based on probability of sparse P(s). First generalize the normal and sparse attributes to satisfy k and k_S then use the ontology to fill the remaining records ($k\text{-}k_S$) based on their most common criterion
12. Ensure K and K_S satisfaction (using semantic ontology)
13. Publish the anonymized data to the external website

5 External Anonymization

Experiments show that constructing a k-block of multiple quasi-identifiers fails to succeed due to the nature of data. For example, our case study includes mobile model attributes such as: Samsung, iPhone, HTC, Huawei, etc. In addition, there are different mobile series that fall under each brand. The rapid development and competency between mobile vendors is introducing more sparse data in the model since we have new brand series every short time. This adds more challenges to the grouping of similar items during the anonymization process because of the existence of sparse data that does not satisfy the k-condition.

5.1 Semantic Ontology Model

To enhance the anonymization methods, many researchers such as the p-sensitivity [4], vocabulary k-anonymity [6], ontology k-anonymity [3] and ontological semantics technology [8] are based on semantic ontology. The hardness of applying k-anonymity in

sparsely data environment and the eventual benefit of using semantics encouraged us to use a semantic ontology and change the core of the k-anonymity process. Contrary to the above cited methods, we propose a semantic ontology to describe the domain of quasi-identifiers in order to distinguish between sparse and non-sparse attributes and to maximize k-block in a sparse data environment.

Resolving sparsity of the Comiqual dataset, lead us to create new ontology system that has the role of providing best common criteria of the extremely mobile brands. We focus on mobile device model that is a sub-class of sparse attribute class. "Mobile" is encountered by a many relationships and properties like operating system, country etc. The idea is to find always a common name for a list of different mobile brands, e.g. the following devices are very sparse: Samsung Note, HTC One, Huawei, LG. Ontology system will infer that all these devices have the same operating system, thus instead of suppressing them all, ontology will alter their names to "android mobile" therefore the k-anonymity is achieved without too much losses. Our semantic model (Fig. 2) consists of two main hierarchies: the sparse and the normal attributes branches. The mobile concept model is inspired from MOKM, a mobile ontology knowledge model that supports information intercommunication among mobile applications and improves the cooperative ability of mobile users [7]. We extended this model by adding the distinction between sparse and non-sparse attributes.

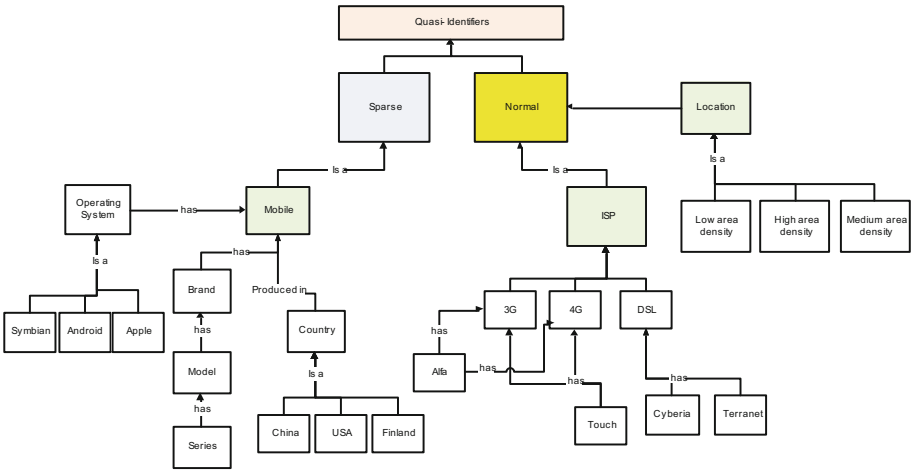

Fig. 2. Semantic ontology model for comiqual data

5.2 Flexible K-Anonymity

In order to select k similar mobiles, we scroll up the sparse attributes hierarchy to infer the aggregation level from the ontology that will enable us to build the appropriate k-block. The properties "is a", "produced in", "has OS", etc. with the joint/disjoint relations between subclasses, enables us to infer at least one common criterion between many devices. For instance, assume we encounter in our dataset the following device

models: HTC One, Samsung S3, Huawei P8 and LG G2. The system infers that those devices have the same android operating system, and that they are produced in the same country, therefore their diverse names will be replaced by either "android mobile" or "Chinese mobile" in order to satisfy the k-anonymity and overcome the sparse data problem.

Keeping sparse attributes (e.g. "mobile model") out of consideration during the construction of k-anonymity blocks can lead to eventual individual identification. While dealing with such attributes, in the same way as for non-sparse ones, it could end up with the anonymization process failure. Flexible k-anonymity is based on splitting the quasi-identifiers attributes into sparse and normal classes towards having sparse data partially contribute in the creation of k-blocks, by k of sparse value ks, where ks < k. This method allows us to use different k values for the same dataset rather than fix a static one. The level of contribution depends mainly on the percentage of sparse data available within the dataset that is called sparse probability P(s) that indicates how much "sparse data" should contribute in k-block construction. The number of the remaining tuples to be filled into the dataset will be inferred from the ontology through the detection of an appropriate aggregation level, by going up in the hierarchy of sparse attribute classes as shown in Fig. 1. We will begin by some definitions, then explain how to apply the flexible k-anonymity approach in order to choose the k values in an optimal way for sparse and non-sparse attributes.

A: Attribute; A_S: Sparse Attribute; *P(s):* sparse probability. K_S*:* k-anonymity of sparse; *K:* k-anonymity of non-sparse
Let $RT(A_1,...,A_n)$ be a table, $QI_{RT}(A_i,\ldots,A_j)$ the quasi-identifiers where $A_i,\ldots,A_j \subseteq A_1,...,A_n$ and $A_S \subseteq A_i,\ldots,A_j$. $K_S = P(s) \times K$ (with $P(s) > 0$) *RT* is said to satisfy *flexible k-anonymity* if each sequence of values in $RT[QI_S]$ appears at least K_S occurrences in $RT[QI_S]$ and each sequence of values in $RT[QI_{RT}]$ appears with at least *K* occurrences in $RT[QI_{RT}]$. The complementary of $K\ (K\text{-}K_S)$ is semantically selected.

5.3 Optimal K-Anonymity and K of Sparse

Flexible k-anonymity approach is based on: (1) determine the sparse and non-sparse quasi-identifiers, i.e. "network type" and "ISP name" are non-sparse attribute while "mobile model" is a sparse attribute. (2) Assign the static k value in a way that enables normal attributes to realize k-anonymity. (3) Evaluate the sparse probability P(s) by conducting a heuristical study on the sparse data values within the dataset. In the Comiqual use case, we found that this probability for the "device mobile" attribute represents no more than 30 % of the whole dataset. (4) Calculate k of sparse value K_s by multiplying P(s) times k. P(s) represents the direct relation between K and k_s:

$$a)\ K_s = P(s) \times K \qquad b)\ K_s := \begin{cases} KforP(s) \ll \\ otherwise; Ks < K \end{cases}$$

Generally, optimal k-anonymity is NP-hard and not easy to be evaluated [12]. In practice for *k* to be a small constant around 5 or 6, gives positive results [1]. In *flexible k-anonymity*, defining the *k* value is done intuitively by estimating the occurrences of similar or approximate records of the non-sparse attribute, whereas k of sparse K_S represents portion of this predefined k value, defined by sparse data proportion of the whole dataset. Semantic ontology provides the complementary of k by unifying the sparse data under the most common criterion name. For instance, assume that $k = 6$ mobiles and sparse probability $P(s) = 50\,\%$, the K_S value will be $6 \times 0.5 = 3$. This means that we should have at least 3 similar mobiles of the 6-blocks to satisfy the flexible k-anonymity. The remaining three sparse records are semantically altered into the most common criterion such as "mobile model name", e.g. the three mobile names might be turned into as "android mobile" or "Chinese mobile". Flexible k-anonymity usage, can be extended to enhance some non-sparse attribute like location, e.g. it can be used to distinguish between low, medium and high area density, then assign different k value for each class. Accordingly, a user with Samsung S6 could be generalized to Samsung mobile in a low area density but when moving to high area density, generalization could not be applied on same model's group that satisfy the k-anonymity. As a result of introducing flexible k-anonymity, we are going to have more accurate anonymized data, with a high level of privacy.

6 Conclusion

In this paper we showed that in order to achieve data anonymization, applying k-anonymity enhancements such as l-diversity, t-closeness and others, could be impractical for datasets holding spatio-temporal information of individuals and which do not contain any sensitive attributes, such as the Comiqual dataset. For such datasets, simply removing the identifiers of individuals or replacing them by a pseudonym does not protect their privacy from inference attacks. In this paper we demonstrated the need for a sanitization process, which should introduce a level of protection at both the internal and the external levels. Future work will detail a new anonymization algorithm to be applied at the internal level when data is at rest, based on hashing the PID, as well as the addition of noise, because deletion of sensible data is not always permitted, as in the case of the Comiqual dataset. For the external level, we proposed the *flexible k-anonymity* for the second use of data, to fight counter sparse data when constructing k-block, by using semantic ontology system that infers the common criteria for the sparse data.

As for future work, we would also like to extend our study to investigate how the continuous flow of data affects the proposed sanitization process and how the arrival of new collected entries can be synchronized with the anonymized data without affecting the efficiency of the process.

References

1. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(05), 557–570 (2002)
2. Lv, P.: Utility-based anonymization for continuous data publishing. In: Computational Intelligence and Industrial Application. Pacific-Asia Workshop (2008)
3. Omran, E., Bokma, A., Abu-Almaati, S.: A k-anonymity based semantic model for protecting personal information and privacy. In: 2009 IEEE (IACC 2009), Patiala, India, 6–7 2009
4. Xiao, Z., Meng, X.: p-sensitivity: a semantic privacy-protection model for location-based services. In: Mobile Data Management Workshops (2008 MDMW) (2008)
5. Daubert, J., Grube, T., Muhlhauser, M., Fischer, M.: Internal attacks in anonymous publish-subscribe P2P overlays. In: 2015 International Conference on NetSys (2015)
6. Liu, J., Wang, K.: Enforcing vocabulary k-anonymity by semantic similarity based clustering, in data mining. In: 2010 IEEE 10th International Conference on ICDM (2010)
7. Junwu, Z., Bin, L., Fei, W., Sicheng, W.: Mobile ontology. Int. J. Digit. Content **4**(5), 46–54 (2010)
8. Ringenberg, T., Taylor, J.: Semantic Anonymization of Medical Records. IEEE, San Diego (2014)
9. Bertino, E., Ooi, B., Yang, Y., Deng, R.: Privacy and ownership preserving of outsourced medical data. In: 2005 ICDE (2005)
10. You, T.-H., Peng, W.-C., Lee, W.-C.: Protecting moving trajectories with dummies. In: Proceedings of the 2007 International Conference on Mobile Data, pp. 278–282 (2007)
11. Kido, H., Yanagisawa, Y., Satoh, T.: An anonymous communication technique using dummies for location-based services. In: ICPS, pp. 88–97 (2005)
12. Meyeson, A., Williams, R.: On the complexity of optimal K-anonymity. In: PODS 2004, New York (2004)
13. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: privacy beyond K-anonymity. In: ICDE (2006)
14. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: ICDE 2007, pp. 106–115 (2007)
15. Gambs, S., Killijian, M.-O., Núñez, M., del Cortez, P.: De-anonymization attack on geolocated data. J. Comput. Syst. Sci. **80**(8), 1597–1614 (2014). http://dx.doi.org/10.1016/j.jcss.2014.04.024