On the Privacy Offered by (k, δ)-Anonymity

Rolando Trujillo-Rasua and Josep Domingo-Ferrer

Universitat Rovira i Virgili
Department of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
{*rolando.trujillo, josep.domingo*}*@urv.cat*

Abstract

The widespread deployment of technologies with tracking capabilities, like GPS, GSM, RFID and on-line social networks, allows mass collection of spatio-temporal data about their users. As a consequence, several methods aimed at anonymizing spatio-temporal data before their publication have been proposed in recent years. Such methods are based on a number of underlying privacy models. Among these models, (k, δ)-anonymity claims to extend the widely used k-anonymity concept by exploiting the spatial uncertainty $\delta \geq 0$ in the trajectory recording process. In this article, we prove that, for any $\delta > 0$ (that is, whenever there is actual uncertainty), (k, δ)-anonymity does not offer trajectory k-anonymity, that is, it does not hide an original trajectory in a set of k indistinguishable anonymized trajectories. Hence, the methods based on (k, δ)-anonymity, like Never Walk Alone (NWA) and Wait For Me (W4M) can offer trajectory k-anonymity only when $\delta = 0$ (no uncertainty). Thus, the idea of exploiting the recording uncertainty δ to achieve trajectory k-anonymity with information loss inversely proportional to δ turns out to be flawed.

Key words: Spatio-temporal data, Trajectory, Data privacy, Anonymity, Uncertainty.

1 Introduction

The exponential growth of computational power, storage capabilities and telecommunication and wireless technologies expedites the collection of user-specific data. The true value of these data lies in their analytical usefulness, which means they should be eventually released to researchers and/or analysts. Therefore, data holders face the challenge of releasing information without compromising the privacy of their users.

In 1998, Samarati and Sweeney [1] proposed a novel formal model named k-anonymity for measuring the privacy of a released microdata set, that is, a collection of records corresponding to individual respondents. The idea is to focus on the set of attributes that can potentially appear also in other publicly available datasets that contain identifiers (*e.g.* electoral rolls, phonebooks, etc.). This set of attributes are called *quasi-identifiers*. If each combination of values of quasi-identifier attributes is shared by at least k records, k-anonymity holds. In this case, the probability of re-identifying a respondent by linking with external identified data sets is at most $1/k$.

The popularity of k-anonymity has led to extensions for specific types of data, like spatio-temporal data. One of these extensions is (k, δ)-anonymity [2,3], which is specifically designed for uncertain trajectories defined as the movement of an object on the surface of the Earth. In this privacy notion, parameter k has the same meaning as in k-anonymity, while δ represents a lower bound of the uncertainty radius when recording the locations of trajectories. To the best of our knowledge, two anonymization methods named *Never Walk Alone* (NWA, [2]) and *Wait for Me* (W4M, [3]), aimed at achieving (k, δ)-anonymity, have been proposed up to date.

1.1 Contribution and plan of this paper

In this article, we analyze the privacy offered by (k, δ)-anonymity and we prove that it does not offer trajectory k-anonymity when $\delta > 0$, that is, when there is actual uncertainty. Our proof is based on a formal definition of trajectory k-anonymity as indistinguishability within a set of k anonymized trajectories. A direct implication of this result is that the two methods NWA and W4M can offer trajectory k-anonymity only when $\delta = 0$ (when there is no uncertainty). Hence, the recording uncertainty δ cannot be exploited to reach trajectory k-anonymity with information loss inversely proportional to δ (which was precisely the aim of (k, δ)-anonymity).

Section 2 recalls (k, δ)-anonymity. Section 3 analyzes the privacy provided by (k, δ)-anonymity and shows that it does not offer trajectory k-anonymity for $\delta > 0$. Section 4 is a conclusion.

2 (k, δ)-Anonymity

The (k, δ)-anonymity privacy notion is based on the assumption that trajectories are imprecise by nature. Unlike records in traditional databases, trajectory data do not remain constant over time, because a moving object should report its position in real-time. However, this is impractical due to performance and wireless-bandwidth overhead. For this reason, Trajcevski *et al.* [4] suggest that a moving object and the server should reach an agreement consisting on an uncertainty threshold δ, meaning that a position is reported only when it deviates from its expected location by δ or more. Considering so, a moving object does not draw a trajectory anymore, but an uncertain trajectory defined by a trajectory τ and an uncertainty threshold δ.

Definition 1 (Trajectory) *A* trajectory *is an ordered set of time-stamped locations*

$$\tau = \{(t_1, x_1, y_1), \ldots, (t_n, x_n, y_n)\} \ ,$$

where $t_i < t_{i+1}$ *for all* $1 \leq i < n$.

Notation. For any time-stamp $t_1 \leq t \leq t_n$, the function $\tau(t)$ outputs the location of τ at time t. If $t = t_i$ for some $i \in \{1, \cdots, n\}$ then $\tau(t) = (x_i, y_i)$, otherwise $\tau(t)$ is the linear interpolation of the poly-line τ at time t. Similarly, $\tau(t)[x]$ and $\tau(t)[y]$ denote the spatial coordinates of the location $\tau(t)$.

Definition 2 (Uncertain trajectory) *An uncertain trajectory is a pair* (τ, δ) *where* τ *is a trajectory and* δ *is an uncertainty threshold. Geometrically, the uncertain trajectory is defined as the locus*

$$UT(\tau, \delta) = \{(t, x, y) | d((x, y), (\tau(t)[x], \tau(t)[y])) \leq \delta\} \ ,$$

where $d((x_1, y_1), (x_2, y_2))$ *represents the Euclidean distance between the locations* (x_1, y_1) *and* (x_2, y_2).

As shown in Figure 1, an uncertain trajectory $UT(\tau, \delta)$ is the union of all the cylinders of radius δ centered in the lines formed by (x_i, y_i) and (x_{i+1}, y_{i+1}) for every $1 \leq i < n$. Then, any continuous function $PMC^\tau : [t_1, t_n] \to \mathbb{R}^2$ such that $PMC^\tau([t_1, t_n]) \subset UT(\tau, \delta)$ is said to be a *possible motion curve* of the uncertain trajectory $UT(\tau, \delta)$.

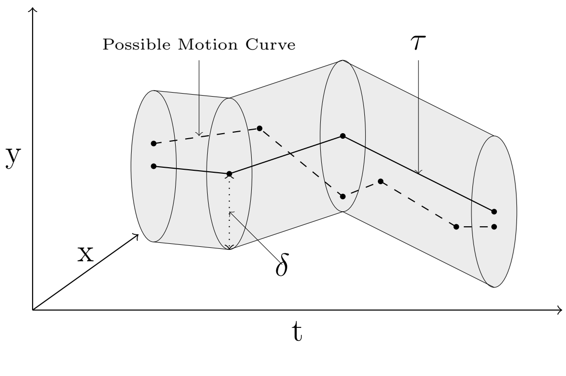

Fig. 1. A trajectory τ and its uncertain trajectory $UT(\tau, \delta)$. A possible motion curve within $UT(\tau, \delta)$ is also shown.

If a trajectory τ_1 is a possible motion curve of the uncertain version (τ_2, δ) of another trajectory τ_2 and viceversa (τ_2 is a possible motion curve of (τ_1, δ)), then τ_1 and τ_2 are said to be *co-localized* with respect to δ [2,3]. This relation is denoted as $Coloc_\delta(\tau_1, \tau_2)$ and provides the rationale behind (k, δ)-anonymity.

Definition 3 ((k, δ)-anonymity set) *Given an uncertainty threshold δ, a set of trajectories S is considered an anonymity set if and only if $Coloc_\delta(\tau_i, \tau_j)$ $\forall \tau_i, \tau_j \in S$.*

Then, (k, δ)-anonymity is defined as follows in [2,3]:

Definition 4 ((k, δ)-anonymity) *Given a database of trajectories $\mathcal{D}$, an uncertainty threshold δ, and an anonymity threshold k, (k, δ)-anonymity is satisfied if, for every trajectory $\tau \in \mathcal{D}$, there exists a (k, δ)-anonymity set $S \subseteq \mathcal{D}$ such that $\tau \in S$ and $|S| \geq k$.*

3 Privacy analysis of (k, δ)-anonymity

The concept of k-anonymity [1] is built upon the definition of quasi-identifier. However, there is no agreement yet about how quasi-identifiers can be defined in spatio-temporal data. Potentially, every location could be regarded as a quasi-identifier [5]. For this reason, some extensions of k-anonymity to spatio-temporal data [6–8] do not consider quasi-identifiers at all and are aimed at releasing groups of k indistinguishable trajectories independently of the adversary's knowledge. (k, δ)-Anonymity [2,3] is also based on this worst case.

Let us use the formalization of this notion of trajectory k-anonymity given in [6].

Definition 5 (Trajectory k-anonymity) *Let T^* be an anonymized set of trajectories corresponding to an original set of trajectories T. Let $\Pr_{\tau^*}[\tau|\sigma]$ denote the probability of the adversary's correctly linking the anonymized trajectory $\tau^* \in T^*$ with its corresponding original trajectory $\tau \in T$ given that the adversary's knows a strict subset σ of the locations of τ. Then T^* satisfies trajectory k-anonymity if $\Pr_{\tau^*}[\tau|\sigma] \leq 1/k$ for every $\tau \in T$ and σ subset of the locations of τ.*

In Definition 5 above, the adversary's knowledge is represented as a *sub-trajectory* of an original trajectory, that is, as a subset of the set of time-stamped locations of the original trajectory. This background knowledge representation is appropriate for the trajectory anonymization schemes [6–8]. However, the uncertainty on the data under (k, δ)-anonymity does not permit to assume that the adversary knows a sub-trajectory in the above sense, except when $\delta = 0$ (no uncertainty). For $\delta > 0$, the adversary at best could know a possible motion curve PMC_τ of a trajectory τ contained in the original database $\mathcal{D}$. In other words, the adversary cannot be sure that her knowledge PMC_τ is exactly what was recorded in $\mathcal{D}$. It should be remarked that the adversary's knowledge was not explicitly defined in [2] or [3]. However, it is required in this article in order to provide formal privacy proofs.

Definition 6 *The adversary's knowledge in a database $\mathcal{D}$ of uncertain trajectories is defined as a random possible motion curve PMC_τ of some trajectory $\tau \in \mathcal{D}$.*

Definition 6 can be seen the other way round: the adversary is assumed to have the ability to acquire true actual locations about a user, such as home address or visited places, but the locations recorded in the database form a random possible motion curve of the adversary's knowledge due to the location uncertainty δ. Note that *not* considering the recorded trajectory as a random possible motion curve of the true original trajectory contradicts the (k, δ)-anonymity concept.

Theorem 1 *Let $\mathcal{D}$ be a database satisfying (k, δ)-anonymity. In general, $\mathcal{D}$ does not satisfy trajectory k-anonymity for any $\delta > 0$.*

Proof: We first give a counterexample which satisfies $(2, \delta)$-anonymity for any $\delta > 0$ but does not satisfy trajectory 2-anonymity; we will then generalize the argument for any k. Let τ_1 and τ_2 be two different but co-localized trajectories w.r.t. δ such that each of them consists of a single location. By the co-localization condition, the time stamp of both locations is the same and the distance d between the spatial coordinates of both locations satisfies $0 < d \leq \delta$.

Let $\mathcal{D}$ be the original dataset containing τ_1 and τ_2 only. Let us provide the adversary with a random possible motion curve PMC_{τ_i} where $i \in_R \{1, 2\}$ is randomly chosen. According to Definition 5, trajectory 2-anonymity is achieved if the adversary cannot guess with probability greater than $\frac{1}{2}$ whether $i = 1$ or $i = 2$.

However, let us consider the following adversarial strategy:

(1) The adversary computes $d(PMC_{\tau_i}, \tau_1)$ and $d(PMC_{\tau_i}, \tau_2)$.
(2) If $d(PMC_{\tau_i}, \tau_1) < d(PMC_{\tau_i}, \tau_2)$, the adversary's guess $i = 1$; otherwise, the adversary's guess is $i = 2$.

Now we will show that the previous strategy achieves a probability of success greater than $\frac{1}{2}$. To that end, let us compute the probability that $d(PMC_{\tau_1}, \tau_1) \geq d(PMC_{\tau_1}, \tau_2)$ for a random PMC_{τ_1}.

Let A and B the two points of intersection of the uncertainty circles of τ_1 and τ_2 (see Figure 2). Then, $d(PMC_{\tau_1}, \tau_1) \geq d(PMC_{\tau_1}, \tau_2)$ only holds when PMC_{τ_1} lies in the arc segment area formed by the points A, B, and the uncertainty circle of τ_1 (shaded area in Figure 2). Since the line $\overline{AB}$ intersects the line formed by τ_1 and τ_2 in its middle point, it can be concluded that $0 \leq d(A, B) < 2\delta$. As $d(A, B)$ grows towards 2δ, the aforementioned arc segment area becomes asymptotically close to its maximum value $\pi\delta^2/2$. This means that:

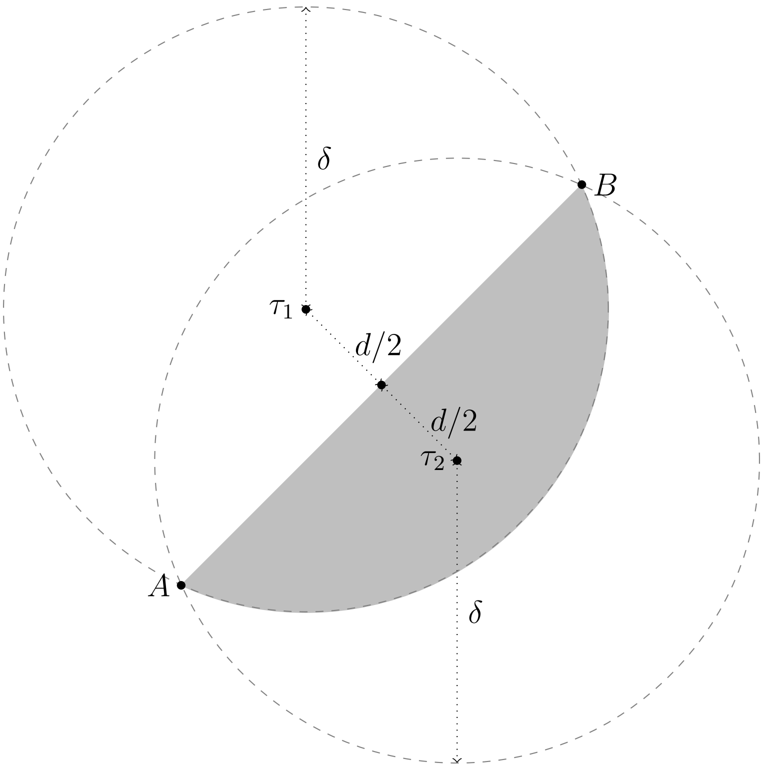

Fig. 2. Two trajectories τ_1 and τ_2 of size 1 such that $d(\tau_1, \tau_2) = d \leq \delta$. The two circles that intersect at A and B represent the uncertainty areas of both trajectories according to Definition 2.

$$\Pr(d(PMC_{\tau_1}, \tau_1) \geq d(PMC_{\tau_1}, \tau_2)) < \frac{1}{2}. \tag{1}$$

From Expression (1), it can be concluded that the adversary's success probability is always greater than $\frac{1}{2}$ for any $\delta > 0$, which contradicts 2-anonymity.

The above reasoning can be generalized to any number k of trajectories. The generalized adversarial strategy is:

(1) The adversary computes $d(PMC_{\tau_i}, \tau_j)$ for all $j \in \{1, \cdots, k\}$.
(2) The adversary's guess is trajectory τ_g such that

$$g = \arg\ \min_{1 \leq j \leq k} d(PMC_{\tau_i}, \tau_j)$$

By generalizing the geometric argument of Figure 2, it can be seen that the adversary's success probability with the above strategy is greater than $\frac{1}{k}$. This contradicts trajectory k-anonymity for any k and δ. $\square$

Corollary 1 *The methods NWA [2] and W4M [3] can only offer trajectory k-anonymity for $\delta = 0$, that is, when all k trajectories in any (k, δ)-anonymity set are identical. In other words, trajectory k-anonymity is offered only when the set of anonymized trajectories consists of clusters containing k or more identical trajectories each.*

4 Conclusions

We have shown that, in general, (k, δ)-anonymity does not offer trajectory k-anonymity for any $\delta > 0$. It only offers this property for $\delta = 0$, that is, when the set of anonymized trajectories consists of clusters containing k or more identical trajectories each. In this situation, the uncertainty of trajectory recording is no longer exploited and a high information loss in incurred: a cluster of k original trajectories are replaced by k identical anonymized trajectories.

We conclude that the idea of exploiting the recording uncertainty δ to achieve trajectory k-anonymity with information loss inversely proportional to δ turns out to be flawed.

Acknowledgments

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 projects "DwB" and "Inter-Trust". The second author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

[1] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Tech. Rep. SRI-CSL-98-04, SRI Computer Science Laboratory (1998).

[2] O. Abul, F. Bonchi, M. Nanni, Never walk alone: uncertainty for anonymity in moving objects databases, in: Proceedings of the 24th International Conference on Data Engineering, ICDE 2008,Cancun, Mexico, 7-12 April 2008, IEEE, 2008, pp. 376–385.

[3] O. Abul, F. Bonchi, M. Nanni, Anonymization of moving objects databases by clustering and perturbation, Inf. Syst. 35 (8) (2010) 884–910.

[4] G. Trajcevski, O. Wolfson, K. Hinrichs, S. Chamberlain, Managing uncertainty in moving objects databases, ACM Trans. Database Syst. 29 (2004) 463–507. doi:http://doi.acm.org/10.1145/1016028.1016030.
URL `http://doi.acm.org/10.1145/1016028.1016030`

[5] C. Bettini, X. S. Wang, S. Jajodia, Protecting privacy against location-based personal identification, in: Secure Data Management, LNCS 3674, Springer, 2005, pp. 185–199.

[6] J. Domingo-Ferrer, R. Trujillo-Rasua, Microaggregation- and permutation-based anonymization of movement data, Inf. Sci. 208 (2012) 55–80.

[7] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, S. Wrobel, Movement data anonymity through generalization, Trans. Data Privacy 3 (2) (2010) 91–121.

[8] M. E. Nergiz, M. Atzori, Y. Saygin, B. Guc, Towards trajectory anonymization: a generalization-based approach, Trans. Data Privacy 2 (1) (2009) 47–75.