The Price of Privacy Control in Mobility Sharing

F. Martelli[a] and M. E. Renda[a,b] and Jinhua Zhao[b]

[a]Istituto di Informatica e Telematica – CNR, Pisa, Italy; [b]JTL Mobility Lab – MIT, Cambridge, USA

ABSTRACT

One of the main features in mobility sharing applications is the exposure of personal data provided to the system. Transportation and location data can reveal personal habits, preferences, and behaviors, and riders could be keen not to share the exact location of their origin and/or destination. But what is the price of privacy in terms of decreased efficiency of the mobility sharing system? In this paper, for the first time, we address the privacy issues under this point of view, and show how location privacy-preserving techniques could affect the performance of mobility sharing applications, in terms of both *System Efficiency* and *Quality of Service*. To this extent, we first apply different data-masking techniques to anonymize geographical information, and then compare the performance of *shareability networks*-based trip matching algorithms for ride-sharing, applied to the real data and to the privacy-preserving data. The goal of the paper is to evaluate the performance of mobility sharing privacy-preserving systems, and to shed light on the trade-off between data privacy and its costs. The results show that the total traveled distance increase due to the introduction of data privacy could be bounded if users are willing to spend (or "pay") for more time in order to share a trip, meaning that data location privacy impacts both efficiency and quality of service.

KEYWORDS

Privacy preserved shared mobility; Transportation privacy price; K-anonymity; Obfuscation; Cloaking

1. Introduction

Mobility sharing applications are becoming more and more popular both in urban and extra-urban settings, for either short and medium-long trips. People get together on the same vehicles to share costs, reduce congestion and pollution, and, whenever possible, socialize (Blumenberg and Smart, 2010; Liao et al., 2018; Librino et al., 2018, 2019; Neoh et al., 2017). By providing these applications with both spatial and temporal data on their activities, users could reveal personal habits, preferences and behaviors. In a typical mobility sharing scenario, a driver who intends to drive from a given location to a destination is willing to transport other potential riders for a part of, or the whole, journey. In the context of ride-sharing, the drivers are professionals who will be paid for the services they offer; in case of carpooling, the users are people having same (or similar) destinations, and they carpool in order to share the cost of the trips, or split the burden of driving and using their (personal) car. One typical example of carpooling is the one used for home-work commuting, where users share trips on a regular basis.

Correspondence Address M.Elena Renda, Istituto di Informatica e Telematica – CNR, 56124 Pisa, Italy. Email: elena.renda@iit.cnr.it, erenda@mit.edu

Mobility sharing applications can vary in how much data they require and/or collect from users, but usually they store information about where users are going, like frequent destinations, how long they stay there, and origins of trips. Data collected from mobile apps could be used for service improvement, for location-based marketing, information or alerts, through well-known mechanisms like geofencing and geotargeting (Lian et al., 2019; Maiouak and Taleb, 2019), or, depending on laws and regulations, is required to be collected and retained by Countries, even for extended periods of time, in order to provide copies of such information to governmental or other authorities (see, e.g., (Feiler, 2010)). Some companies might simply share general usage statistics, while others may send user-identifying data to third parties (Cleary, 2018; Jennifer Valentino-DeVries and Krolik, 2018). As mobility companies further blur the line between providing technology and transportation services, how they use the amount of data they collect on users is becoming a growing concern among citizens, authorities, and companies. Furthermore, data breaches are increasing, as cybercriminals find company databases to be populated with treasured personal information: hackers often sell information to professional scammers over black data markets (The Manifest, 2019). Given that data breaches and misuse of personal information have become more and more common over the past few years, users could be keen not to share the exact location of their origins and/or destinations.

The data privacy problem requires to find a trade-off between data privacy and *data utility*, a metric usually related to the performance, cost and Quality of Service (QoS) based on the effectiveness of the underlying data. On one hand, masking location data in order to avoid the identification of users in case of data leakage, misusage and/or breaches increases user privacy; on the other hand, the loss of information could lead to data utility decreases and to poor quality or low efficiency of the location-based system. Some ride-sharing systems provide options for the users to meet at a pick up point instead of providing their real address (Aïvodji et al., 2016), or have studied methodologies to mask users' physical coordinates before the system acquires them (Aïvodji et al., 2018; Luo et al., 2019).

But *what is the price of privacy control in terms of data utility? And how would this affect the efficiency of mobility sharing systems, such as total miles driven to serve all travelers?* Here for the first time, we address privacy issues under this point of view, and show how data privacy-preserving techniques applied to location data, with increasing granularity of data anonymization, could affect the performance of mobility sharing applications. We consider both transportation efficiency, namely Vehicle Miles Traveled (VMT), and quality of service, namely users' waiting and riding time. In particular, we address a specific mobility sharing scenario, carpooling for home-work commuting, because it is the largest contributor to traffic congestion and pollution, but also because this kind of "regular" trips have the potential to reveal users' repetitive route patterns and their recurrent time schedules. To this end, we use the data collected through a survey issued in 2016 to people living and/or working in Pisa (Italy) and its surroundings (City of Pisa, 2016), in which users provided exact home and work addresses, time schedules, and flexibility on departure times. To the best of our knowledge, this is one of the few datasets including such a level of geographic and time details.

The transportation community earnestly needs a clear definition of property right of mobility data, and a set of precise rules that govern the ownership, use, transaction, and erasure of digital information related to mobility. This paper focuses on one specific component of this broad mobility data discussion: the value of locational privacy quantified from the supply perspective, i.e., the impacts of maintaining the different

degrees of locational privacy upon the system performance of the shared mobility service.

To the best of our knowledge, this is the first paper trying to (i) quantify the effects of privacy control on ride-sharing applications in terms of quality of service and system efficiency loss; (ii) compare different techniques and different levels of location data anonymization granularity; (iii) show the trade-off between data location privacy and data utility.

In order to anonymize location data and guarantee privacy to carpoolers, we use three different data-masking techniques, namely *k-anonymity*, *obfuscation* and *cloaking*, and fine-tune their settings to achieve different levels of anonymization granularity (Section 3.3). To match trips, we apply the *Carpooling Shareability Networks-based trip matching algorithms* presented in (Librino et al., 2018). Carpooling shareability networks connect trips (the nodes in the network) satisfying spatial (physical location) and temporal (user's departure time and flexibility) constraints (see Section 3.2). The presence of a link between two nodes in the network means that the related trips are shareable. Over the generated network we compute the optimal trip matching to maximize the saved VMT.

First, we compare the *privacy-preserving shareability network*, generated using the masked data, with the *original shareability network*, generated using the real location data, to quantify how many trips becomes shareable in the former while they are not in the latter (*false positive*), and how many are recognized as not sharebale while they are in reality (*false negative*). Then we compare the optimal matching over the original network and the matching over the privacy preserving network, showing that choosing *bad matches* in the latter, i.e. false positive links, will degrade the performance in terms of VMT and waiting/riding time. We use the difference in the system performance as a way to quantify the price of privacy. The results clearly show that the price to pay to guarantee data privacy within a carpooling system is to increase its total carbon footprint, i.e. the total traveled miles, but this effect could be mitigated if users are willing to spend (or "pay") for more time in order to share a trip on a daily basis.

In this paper we are assuming that the ride-sharing system will only receive the masked data and will never be able to know the real position of the users even after the (privacy-preserving) matching has been done. On a real implementation, though, matched users should share their reciprocal information once the system notifies them that their trips could be shared. Nevertheless, at that time users themselves will chose to (privately) communicate to the matched carpooler their exact address, or find an agreement on a meeting point.

The paper is organized as follows: in the next Section we present general approaches for preserving privacy, in different application scenarios, and then focus on the transportation and shared mobility scenario; Section 3 introduces the data masking mechanisms we identified as the most suitable for the mobility sharing scenario, and we describe the methodologies used for the simulations, the data, and the experiment settings; Section 4 presents the results we obtained, while in Section 5 we conclude by discussing the results' implications and suggesting future direction in the implementation of privacy-preserving mobility sharing systems.

2. Related work

In this paper, we present different methodologies for masking users' location data, so not to provide the exact location to the ride-sharing system and to the other riders as

well. To prevent or mitigate privacy breaches, many location data masking techniques have been proposed in literature to hide real users' locations while providing their approximate information used in Location Based Services. *k-anonymity* technique was introduced in the context of relational database (Sweeney, 2002), where data are stored in a table and each row of this table corresponds to one individual. The basic idea of the k-anonymity model is to guarantee that the information of every data subject cannot be distinguished from the information of other $k-1$ data subjects. In this context, for example, the dummy location insertion (Kido et al., 2005; Shankar et al., 2009) generates $k-1$ dummy points and makes a user's location indistinguishable among a set of k locations, which provides k-anonymity.

The spatial and temporal cloaking techniques (Bamba et al., 2008; Gedik and Liu, 2005; Gruteser and Grunwald, 2003; Mokbel et al., 2006) choose a sufficiently large region that includes k indistinguishable locations to achieve k-anonymity.

To further improve privacy properties of the k-anonymity mechanism, the l-diversity concept has been introduced (Machanavajjhala et al., 2006): the cloaked region containing the k individuals must include l different POIs. This may, of course, sensibly reduce the accuracy of the actual position and lead to an excessive degradation of the quality of service (Bamba et al., 2008; Xue et al., 2009).

A specific concept of *obfuscation*, similar to both k-anonymity and the spatial cloaking technique proposed in (Gruteser and Grunwald, 2003), has been defined in (Duckham and Kulik, 2005): here, obfuscation means "*deliberately degrading the quality of information about an individual's location in order to protect that individual's location privacy*". The geographic space is modeled as a graph in which the nodes are locations and the edges represent connectivity and/or proximity between locations. The actual location is obfuscated by providing a subset of the nodes whose cardinality could be tuned to obtain the desired level of privacy.

The notion of *differential privacy* first appeared in (Dwork, 2006) in the context of statistical databases. The key concept is that an aggregated response to a query does not change if a single user data is modified. In other words, the probability of a response r to a query sent to database D and the probability of a response r' sent to database D' (where D and D' are adjacent, namely they differ only in a single individual data) should be less than e^{ϵ}. Typically this is achieved by adding random noise to the query output. In (Ho and Ruan, 2011), a differential privacy approach has been applied for interesting geographic location discovery using a region quadtree spatial decomposition. A more refined concept of location privacy (similar to differential privacy) has been introduced in (Andrés et al., 2013): this new notion is named *geo-indistinguishability* and is based on a bi-variate version of the Laplace function, to perturb the actual location. The case study application is the point-of-interest retrieval, and the proposed framework takes into account the desired level of protection as well as the side-information that the attacker might have.

2.1. Privacy in Transportation

Many papers on location privacy are focused on location-based applications such as point of interest retrieval, mapping applications, GPS navigation and traffic patterns, and location-aware social networks.

In transportation, one of the first contexts in which the privacy problem arose has been vehicular networking. Typical applications in this field are: traffic monitoring, road safety and infotainment. In a vehicular network, vehicles communicate with each

other and with road infrastructures towards back-end servers by including in the messages their position, speed and other information that could lead to a privacy leakage. One way of approaching privacy in vehicular networks is the use of pseudonyms (Baik Hoh et al., 2006; Bettini et al., 2009; Buttyán et al., 2007; Eckhoff et al., 2011; Sucasas et al., 2016). Another way is to encrypt the mobility traces as in (Bilogrevic et al., 2014; Xi et al., 2014), or to blur them into a region with lower resolution (both in space and time) as in (Cheng et al., 2006). In (Hoh and Gruteser, 2005), a path perturbation algorithm is proposed to avoid the tracking of individual users: the key concept is to cross paths in those regions where at least two users meet, so decreasing the possibility that an adversary could distinguish the trajectories of different users.

In (de Montjoye et al., 2013), an interesting result has been given by analyzing the mobility phone traces of more than one and half million of individuals. They found that only four spatio-temporal points are sufficient to uniquely identify human trajectories of 95% of individuals. Given that the resolution of their dataset was quite coarse (the space resolution is given by the zone covered by an antenna, and the temporal resolution is one hour), it is clear that it is rather simple for an attacker to infer private information. A similar study has been presented in (Gao et al., 2019). Here, the authors used a data set, generated by a LPR (License Plate Recognition) system composed by 516 detectors, consisting of 260 million spatio-temporal transactions generated by 14 million vehicles in one month. They showed that five spatio-temporal records are enough to uniquely identify about 90% of individuals even when the temporal resolution is 12 hours. Moreover, they proposed two procedures to be applied to LPR data sets before publishing in order to preserve privacy: one is based on suppressing sensitive records, while the other one is a generalization solution using a bintree-based adaptive time interval cloaking algorithm. In (Ma et al., 2013) instead, the focus is on the amount of side information that could be sufficient to an attacker for identifying an individual mobility trace from a data set of anonymous trajectories. They show that just 10 pieces of side information (referred as "location snapshots") are enough for identifying between 30% and 50% of the victims. A "Privacy-by-Design" approach by means of virtual trip lines (VTLs) is proposed in (Sun et al., 2013). The key concept is that mobile data is collected only when the privacy is preserved. They extend the concept of VTL by considering a traffic modeling system of VTL *zones* able to apply different filtering approaches depending on the desired level of privacy.

The *risk-utility* trade-off of mobility data sets has been analyzed in depth in (Calacci et al., 2019), by evaluating the balance between the value of high-resolution location data to the private market and to the public interest, and its implications for individual privacy and personal risk. This paper is quite interesting since it also provides real-world examples of how high resolution location data, with different degrees of privacy protection, can be used in the public sphere and how it is currently used by private firms.

In (Stenneth and Yu, 2010), the concept of *transportation mode homogeneity* is introduced: a user submitting a location based query is aggregated with at least $k-1$ other mobile clients traveling with the same transportation mode. The reason for suggesting this kind of k-anonymity is that by simply grouping k users within a region is not sufficient to guarantee privacy: in fact, if $k-1$ mobile clients are moving in that region for instance by walking or cycling, and one is driving a car, it is very easy for an adversary to identify them in case of continuous queries (a continuous query is when the client asks a service for a certain period of time).

In ride-sharing applications, the problem of data privacy preservation is of central importance because in order to match drivers and riders, their position and departure

time must be communicated to the system. In (Aïvodji et al., 2016), an integrated approach consisting in advanced privacy technologies combined with multimodal shortest path algorithms has been proposed for computing meeting points that are mutually interesting for both drivers and riders. In (Luo et al., 2019), a privacy-preserving ride-matching scheme is proposed: riders and drivers are matched by the server according to their distances in the road network without revealing their location. In (Goel et al., 2016), the proposed privacy-aware ride-sharing scheme is based on the obfuscation technique applied to riders' location. In (Sherif et al., 2017) instead, autonomous vehicles-based ride-sharing is considered: in the proposed scheme, the ride-sharing region is divided in cells, each one corresponding to a bit in a binary vector, the trip data is coded trough these binary vectors, encrypted before being sent to the server, and a secure k-nearest neighbors algorithm is applied to find the most similar trips. A similar work is presented in (Nabil et al., 2019). A privacy preserving carpooling scheme is presented in (Li et al., 2019) which uses anonymous authentication and build a private blockchain into the carpooling system to record carpooling records for data auditability. As in the current paper, the papers presented in literature try to achieve privacy preservation through mechanisms usually borrowed from other contexts, but while previous literature mainly focuses on achieving good levels of computational performance in matching rides while protecting privacy, or demonstrating the security properties of the privacy protocol adopted, here we try to make a step forward, by evaluating the effects of privacy control on the overall system. In particular, in this paper we try, for the first time in literature, to quantify the effects of data-masking on ride-sharing applications both in terms of QoS (users' riding time and waiting time) and system efficiency (VMT). To do so, we compare the results of different techniques for location data anonymization, using different settings to evaluate different levels of privacy granularity, and highlight the trade-off between data location privacy and data utility.

3. Materials and Methods

The main goal of this paper is to capture the trade-off between data utility and data privacy within ride-sharing applications. In order to tackle this problem we build an evaluation model requiring: (i) trip matching algorithms efficient enough to be run several times on different data-masking algorithms and with different parameter settings; (ii) users' location data, with departure time and flexibility to compute efficiency and QoS costs; (iii) a certain number of data-masking methodologies allowing to fine-tune the privacy granularity; and (iv) a methodology for the evaluation of the system QoS and efficiency outcomes.

In this Section we present the methodologies we have chosen (and why) to model sharing opportunities within a carpooling system, the techniques we applied to anonymize location data in order to protect sensible users' information with different level of privacy granularity, and the data we used to test the algorithms and the methodologies.

3.1. Shareability Networks

To compute sharing opportunities we use the Shareability Networks (SNs) paradigm, firstly introduced in (Santi et al., 2014) to model ride-sharing in the context of taxi trips. The novel idea of the authors is to represent the trips as network nodes, where

nodes are connected only if the corresponding trips are shareable, based on given conditions, i.e. time and space constraints of the trips. The algorithm used to find an optimal solution for matching trips is a (weighted) graph matching, that is computationally very efficient, no matter which is the specific objective function to optimize.

To model carpooling opportunities here we also use a graph representation of the sharing opportunities named *carpooling shareability network* presented in (Librino et al., 2018), defined as $N = (T, E)$, where T is the set of nodes corresponding to the trips in the dataset, and E is the set of directed links between nodes corresponding to shareable trips. The network as defined herein is substantially different from the one defined in (Santi et al., 2014) to analyze taxi ride-sharing opportunities in New York City, since we use dissimilar criteria to determine the existence of a link in the network; furthermore carpooling links have a direction (once the driver is determined) in contrast to the bi-directional links used for taxi rides. In fact, different from taxi ride-sharing and Uber-like car-sharing programs, in which drivers roam the city picking up and dropping off passengers, in the context of carpooling travelers have a different role, namely that of *driver* and *passenger*. When carpooling, the traveler who is also the driver might incur a higher travel "cost" due to the extra time typically required to detour from the optimal route in order to pick the passenger up (and, possibly, drop her off). The maximum amount of extra time allowed by the driver for the detour needed for carpooling is called the *sharing delay* Δ.

Each trip T_i in T is defined by:

- $S(i)$ and $D(i)$, respectively the starting and the destination locations expressed as a pair of $(lat, long)$ coordinates;
- $s_t(i)$, the typical starting time of the trip;
- $t_t(i)$, the flexibility on the starting time. This value, together with $s_t(i)$, identifies a time interval $[s_t(i) - t_t(i), s_t(i) + t_t(i)]$ which represents the time window within which the traveler is willing to start the trip. Consequently, we set $s_t^{min}(i) = s_t(i) - t_t(i)$ and $s_t^{MAX}(i) = s_t(i) + t_t(i)$.

Let d_{AB} be the distance between the locations A and B, as determined by the underlying road network; and τ_{AB}, the time required to travel from A to B. Travel time estimations have been obtained through a public, OpenStreetMap-based API.

Using the above notation, for a given trip T_i, its length is denoted by $d_{S(i)D(i)}$, and its duration is $\tau_{S(i)D(i)}$.

3.2. Shareability Conditions

Given a driver trip T_1 and a passenger trip T_2, we say that T_2 is shareable with T_1 if the three following conditions hold:

$$\tau_{S(1)S(2)} + \tau_{S(2)D(2)} + \tau_{D(2)D(1)} \leq \tau_{S(1)D(1)} + \Delta \tag{1}$$

$$s_t^{min}(2) - s_t^{MAX}(1) \leq \tau_{S(1)S(2)} \leq s_t^{MAX}(2) - s_t^{min}(1) \tag{2}$$

$$d_{S(1)D(1)} > d_{S(1)S(2)} + d_{D(2)D(1)} \tag{3}$$

The first condition states that the required detour time for the driver to pick the passenger up and drop her off does not exceed the threshold value Δ. The second

condition ensures that the two starting time windows are properly overlapped. While the second condition establishes the temporal compatibility of the two trips, this is not enough to ensure that carpooling actually contributes to reduce traffic and carbon footprint. In fact, it is possible that the driver performs a very long detour to pickup the passenger, thus making the length of the shared trip longer than the sum of the lengths of the two individual trips. So, we add the third condition that ensures that the above described possible negative side effects of shared mobility are avoided.

To sum up, in the carpooling shareability network $N = (T, E)$ a link $e_{1,2} \in E$ between T_1 and T_2 exists if and only if the three conditions (1), (2) and (3) stated above are satisfied.

Similarly to what have been done by (Librino et al., 2018; Tachet des Combes et al., 2017) and (Santi et al., 2014), in the remainder of this paper we focus on pair-wise, static matching of rides: i.e., at most two trips can be combined into a single ride and, when the shared ride is formed, no further modification to the route is possible. Although simplified with respect to some ride-sharing models and existing services, this approach has proven to be computationally effective and to provide impressive opportunities for ride-sharing.

The reason why we have chosen the Shareability network-based algorithms for matching the trips is that, to the best of our knowledge, this is the only methodology accounting for shared traveled distance, detour time (i.e. extra travel time due to sharing), and flexibility (i.e. waiting time), while performing trip matching, in a very efficient way, allowing to easily compare data-masking techniques and different parameter settings.

They are also fast enough to allow testing different algorithms and sets of parameters (Section 3.5), both for matching (e.g. 16 delta detour times) and privacy (e.g. the granularity given by the parameters radius and k for k-anonymity). This allowed us to exactly analyze which is the impact of privacy masked-data not only on traveled distance but also on traveling time, represented by the detour time parameter Delta (Par 3.2, Equation 1), and waiting time, represented by the time windows in the sharebility conditions (Par 3.2, Equation 2).

3.3. Introducing Data Privacy in Shared Mobility

To study the effects of data privacy control in carpooling applications, we have implemented 3 different techniques of location data anonymization:

(1) *obfuscation*;
(2) *k-anonymity*;
(3) *cloaking.*

The obfuscation mechanism we implemented consists in generating a random location within an area centered in the real location with a given radius.

K-anonymity is a very common technique and consists in generating $k - 1$ dummy locations within an area centered in the real one with a given radius (as in the obfuscation mechanism) and then randomly selecting a location among the dummy ones plus the real one.

In the cloaking technique, the real location is replaced by a point (like the center or the centroid) of a given region. For this purpose, we wanted to use similar regions, at least in terms of people, a stable set of geographic units for the presentation of data. For our study we computed the centroid locations of the census blocks (ISTAT, Istituto Nazionale di Statistica, 2011) to which the real position belongs to. Census

blocks are statistical areas bounded by visible features, such as streets, roads, streams, and railroad tracks, and by non-visible boundaries, such as selected property lines and city, township, school district, and county limits. Generally, census blocks are small in area; for example, a block in a city is bounded on all sides by streets. Census blocks nest within all other tabulated census geographic entities and are the smallest unit for all tabulated data. Census blocks in suburban and rural areas may be large, irregular, and bounded by a variety of features.

Figure 1 graphically shows how the three mechanisms work, with the real position represented with a star and the circles representing the masked locations.

3.4. Data

The *MobilitandoPisa* (City of Pisa, 2016) joint initiative of the National Research Council with the city of Pisa led to an anonymous mobility survey whose objective was to understand the daily commuting habits within the metropolitan area of the city of Pisa.

The survey data we have chosen for this study has the unique feature of containing not only detailed information on origins and destinations of daily car commuters and their departure and arrival times from home to work and vice versa, but also a precise quantification of their flexibility in departure and arrival time. This information provides the opportunity to also evaluate the effects of data privacy on users waiting time. To the best of our knowledge, there is no other available datasets providing information on origins, destinations, departure time and flexibility, and we believe that a dataset based on a real situation could provide a better understanding of reality.

The survey was extensive and structured into different parts; the most relevant for this study are reported below:

- *Demographic Data*, includes information such as gender, age, marital status, educational level, profession, and owned transportation means;
- *Workers Commuting Data*: transport mode(s) used for commuting, commuting hours, typical travel times, home and work addresses, working times, and their potential flexibility in departing earlier/later from/to home;
- *Attitude towards*: car/ride-sharing/pooling.

Among the 6,200+ respondents, we extracted the subset of $N_P = 1965$ commuters who daily commute by car, agrees on carpooling and, most importantly, provided detailed information on their home and work addresses, departure times and flexibility.

3.5. Experiments

We have run extensive experiments to analyze the effect of privacy control on carpooling. In particular, we want to understand how masking the real coordinates of carpoolers' positions will affect the global system efficiency (in terms of carbon footprint, i.e. VMT) and the quality of service (in terms of the extra time required to the users in waiting for a ride or riding). We perturbed both origin and destination coordinates to be as more general as possible.

We built the SNs over the real location data and the privacy-preserving one, i.e. using the real coordinates in the former case, and the masked ones in the latter. For the detour time Δ we used values ranging from 0 to 15 minutes.

For both obfuscation and k-anonymity we used three radii: 100, 200 and 500 meters,

while for the latter k was set to 5 and 10. Since in obfuscation and k-anonymity the point is randomly selected, we run the same setting 20 times, and for each we will report the average. We computed the performance in terms of number of matched nodes (trips) and saved traveled distance with respect to the sum of all the single ride distances, this last being a measure strictly correlated with carbon emissions (one of the metrics against which the hidden costs of privacy will be evaluated).

For matching the trips over the SN, different methodologies could be used depending on the objective we want to achieve. If the objective is to maximize the number of matched trips and minimize the number of circulating vehicles, the *maximum cardinality matching* (M_c) can be used. If the objective is to reduce carbon footprint, by maximizing the amount of kilometers that could be avoided by matching the trips instead of performing single trips, the *maximum saved distance matching* (M_d) algorithm can be used. In the former the optimal solution is given by computing the maximum matching on the SN, while in the latter we compute a maximum (distance-) weighted matching on the SN.

Since in this paper we focus on carbon footprint, we have chosen M_d; in Section 4 we only report the results obtained with M_d in terms of saved traveled distance.

4. Results

As said, SNs connect trips whose spatial and temporal constraints satisfy all the shareability conditions presented in Section 3. But what happens when spatial information is masked to provide privacy to the users? In Figures 2 and 3 we show two examples of what could happen in the SN creation and in matching the trips when masking the geographical coordinates of the trips, and introduce the concept of "*good matches*" and "*bad matches*".

In Figure 2 we have 8 trips to be paired: after verifying the shareability conditions introduced in Section 3.2, we obtain the SN in Figure 2a; the optimal solution for maximizing the saved traveled distance is represented by the matching in Figure 2b. In Figure 2c on the right we depicted the SN generated by the privacy-preserving data, where some of the links correspond to the original ones (true positive), some links disappeared, meaning that with the masked data the trips result to be not shareable (false negative), while other nodes have been connected through links even though they were not in the original SN, meaning that they did not meet all the shareability conditions on the real data, but they do on the privacy-preserving network (false positive). In this example none of the false positive ("bad links") has been chosen for the matching: we define these "*good matches*". Nevertheless, the solution is not the optimal one in the original network. This implies that the total saved traveled distance of the matching solution computed in the privacy-preserving network of Figure 2d is lower than that achieved by the optimal solution reported in Figure 2b.

In Figure 3, we have a SN with 10 nodes. As for the previous example, after masking the geographical coordinates, we do have true positive, false positive, and false negative, but in this case during the matching phase some of the false positive links (between trips $T_2 - T_3$, $T_4 - T_9$, $T_6 - T_7$, $T_8 - T_{10}$) have been selected (see Figure 3b): in this case we are in presence of "*bad matches*", and including them in the final result means not only that the returned matching would not correspond to the optimal solution in terms of saved traveled distance, but also that for these matched trips either the traveled distance is greater than performing the single trips, and/or they require more flexibility on the departure time, and/or they require a greater detour time than

the nominal one – recall conditions (1),(2),(3) in the definition of SN. To quantify the price of introducing privacy control in carpooling applications, we first compared the SNs obtained with the privacy-preserving data with those obtained with the *real* data (Section 4.1), and then the corresponding matchings (Section 4.2).

4.1. Comparing shareability networks

To compare the SNs, we introduce the concepts of "*good link*" and "*bad link*": a link in the privacy-preserving SN is *good* (also called *true positive*) if the same link is present in the real SN, namely the two related trips are effectively shareable; on the other hand the link is *bad* (also called *false positive*) if it is not present in the real SN, namely the two related trips are identified as shareable with the anonymized location data, but not with the real data.

Here we report, for all the privacy techniques applied, the percentage of true positive (TP) with respect to the real SN and the percentage of false positive of the privacy-preserving network (Figures 4 and 5). We do not report the percentage of false negative links in the privacy-preserving SN, i.e. all the links connecting trips in the real SN resulted unshareable with the masked data, since they are easily derived as $1 - TP$. Both results show that low values of detour time Δ imply high loss in terms of data utility; the same happens if we zoom out and mask more users' locations by increasing the level of privacy through higher radii (for obfuscation and k-anonymity) or the value of k (for k-anonymity). Cloaking performs better than the other approaches in terms of true positive links when Δ is really low (< 2), or when obfuscation and k-anonymity have radius higher than 100 meters, behaving similarly to the 200 meters radius-based methods when $\Delta > 5$, with higher degrading performance when $\Delta > 10$; in terms of false positive links, cloaking returns similar results as the 200 meters radius methods.

Considering for instance the value of $\Delta = 5$, we lose from 40% – using k-anonymity with 100 meters radius and k=5 – to 65% of good links – using obfuscation with 500 meters radius (see Figure 4); the methods have the same behavior in identifying bad links (see Figure 5), i.e. links in the privacy-preserving SN connecting 2 trips not shareable in the real SN: about 53% of the total links of the privacy-preserving network are "wrongly" introduced by k-anonymity with 100 meters radius and k=5 due to the masked data, a percentage that increased up to 68% when using obfuscation with 500 meters radius.

4.2. Comparing matchings

To compare the matchings, we introduce the concepts of "*good match*" and "*bad match*": a match over the privacy-preserving SN is *good* if the corresponding link is present in the real SN, namely the link is a true positive one, meaning that the two related trips are effectively shareable if considering the real location data; on the other hand, the match is *bad* if chosen among the false positive links.

Here we show how "good" (or "bad") the matching obtained with the privacy-preserving SN could be with respect to the real one. To this end, we do not only compute the percentage of *good matches* through the privacy-preserving SN, but also the total traveled distance saved with privacy-preserving carpooling, and compare it with the optimal savings that would be achieved having exact knowledge of the origin and destination of trips. A value of 80%, for instance, would indicate that the optimal privacy-preserving carpool matching achieves only 80% of the total distance savings

achievable without privacy preservation.

In Figure 6, the percentage of good matches is reported: even if cloaking outperforms all the other mechanisms for $\Delta \leq 1$, its performance gradually degrades to eventually return similar results as the 200 meters radius-based methods when $\Delta > 5$; the k-anonymity mechanism with 100 meters radius and k=5 is the best when $\Delta \geq 2$, while the worst is obfuscation with 500 meters radius.

In Figure 7 we report the saved traveled distance obtained over the privacy-preserving SN when considering only the good matches, comparing all the mechanisms: the trend is exactly the same as the one showed for the good matches percentage (Figure 6), with k-anonymity mechanism with 100 meters radius and k=5 being the best in general, and obfuscation with 500 meters radius being the worst. We also report the traveled distance savings comparing cloaking and the 100 meters-based mechanisms (Figure 8), cloaking and the 200 meters-based mechanisms (Figure 9), and cloaking and the 500 meters-based mechanisms (Figure 10). As for the SN comparisons, when comparing the matchings low values of detour time Δ imply high loss in terms of system efficiency, i.e. saved traveled distance; the same happens if we zoom out and mask more users' locations by increasing the level of privacy through higher radii (for obfuscation and k-anonymity) or the value of k (for k-anonymity). Cloaking outperforms all the other methods when these consider a radius of 500 meters (quite clear from Figure 10).

If we consider, for instance, $\Delta = 5$, k-anonymity with 100 meters radius and k=5 could achieve a traveled distance saving of about 74% of the one achievable without privacy preservation (Figure 8), while obfuscation with 500 meters achieves only a 57% of distance savings (Figures 10); these saving values could be increased up to 88% and 80%, respectively, if $\Delta = 10$, and 93% and 88% if $\Delta = 15$.

In Figure 11 we show how different granularity levels of privacy applied to k-anonymity could affect system efficiency: increasing privacy control, either through the radius or k, clearly worsens the performance, while higher values of Δ reduce the losses in terms of saved traveled distance, with a maximum result of 93% achievable with 100 meters radius k-anonymity and $\Delta = 15$.

We have also computed the total saved traveled distance of the optimal matching over the privacy-preserving SN, counting all the matches provided by the matching algorithm, regardless of whether a match is good or bad. To quantify the performance degradation due to privacy preservation, the saved traveled distance of a bad match is computed by using the *real* start and destination locations of the related trips. When comparing the results reported in Figure 12 with those reported in Figure 7, the overall trend of the different methodologies is the same, but for low values of Δ the total saved distance is higher, with a maximum saving around 97% for the total saved distance against a 93% for the good matches-based saved distance. This happens because a bad match in the privacy-preserving SN could be so because one or more of the shareability conditions introduced in Section 3.2 are not satisfied when using the real data. If we have a bad match because the Condition 3 is not satisfied (i.e. the length of the shared trip is longer than the sum of the lengths of the two individual trips), the correspondent saved distance will be negative. On the other hand, let's assume we have a bad match m_x for which Condition 3 is satisfied (i.e. the length of the shared trip is shorter than the sum of the lengths of the two individual trips) the total saved distance will be higher even when selecting m_x; but, since m_x is a bad match, it doesn't satisfy either Condition 1 (i.e. the required detour time for the driver to pick the passenger up and drop her off does not exceed the threshold value Δ) or Condition 2 (the starting time windows of driver and passenger are properly

overlapped), or both. In this case we are in presence of the second price of privacy control: the user time, both in terms of detour time and departure time flexibility.

In Figure 13 we compare, for k-anonymity with 100 meters radius and k=5 only, the good matches saved traveled distance and the total saved traveled distance over the privacy-preserving matchings. The difference between these two values is quite high for low values of detour time Δ, and becomes smaller while increasing Δ. On the other hand, in Figure 14 we report the break down of percentage of bad matches depending on which of the three conditions (detour time Δ, the time window, and/or the saved distance) is violated.

For instance, if we select $\Delta = 5$ we have a 74% good matches saved distance against a 95% of the overall matching saved distance (Figure 13), but for the same value we have more than 88% of bad matches due to an increased detour time (Figure 14), meaning that the effective detour time required to share those trips is higher than the nominal Δ. The higher is the nominal Δ, the lower is the percentage of bad matches due to its related condition (Condition 1), and the higher is the percentage of bad matches due to non-overlapping time windows (Condition 2) and/or negative saved distance (Condition 3). This is even clearer from Figures 15 and 16, where we compare all the mechanisms by reporting the percentage of bad matches present in the privacy-preserving matching due, respectively, to an increased detour time or to an increased traveled distance for the shared trip. In particular, we selected 4 different values of Δ, namely 0, 5, 10 and 15, and for each method and each Δ we show how increasing the privacy control through the radius value affects the bad matches in terms of detour time (Figure 15) and traveled distance (Figure 16). Since cloaking is the only method not based on radius, in these Figures we report the same value for all the radii, while for k-anonymity, for ease of reading, we only report the results for k=5.

In terms of percentage of bad matches due to an increased detour time, cloaking performs better for $\Delta = 0,\ 5$ no matter the radius used in the other methodologies, while for $\Delta = 10,\ 15$ cloaking is outperformed by obfuscation and k-anonymity when the radius is greater than 200 meters (see Figure 15). K-anonymity and obfuscation have similar trend and behavior, even if the latter is always slightly better than the former. Higher values of Δ perform better for all the methods, i.e. reduce the percentage of bad matches due to higher effective detour time. This increase of detour time is quantifiable in about 3' for $\Delta = 5$, on average, and about 5' for $\Delta > 10$, on average, for all the mechanisms considered (cloaking included).

In terms of percentage of bad matches due to an increased distance traveled when performing the shared trip, cloaking performs better for $\Delta = 15$, no matter the radius used in the other methodologies, while for $\Delta = 0,\ 5,\ 10$ cloaking is outperformed by obfuscation and k-anonymity when the radius is greater than 200 meters (see Figure 16). Also here k-anonymity and obfuscation have similar trend and behavior, but in this case k-anonymity is always slightly better than obfuscation. Lower values of Δ perform better for all the methods, i.e. reduce the percentage of bad matches due to higher traveled distance.

The combination of all the results presented here clearly shows that a trade-off between privacy preservation, system efficiency (traveled distance) and quality of service (time) can be fine tuned for privacy-preserving carpooling systems.

5. Discussion

In this paper, for the first time in literature, we try to quantify the effects of privacy control on ride-sharing applications, and capture the trade-off between data privacy and data utility, both in terms of QoS (users' riding time and waiting time) and system efficiency (VMT). We have compared different techniques for location data anonymization, achieving different levels of privacy granularity by fine-tuning their settings.

The analyses reported in this paper provide useful insights into the trade-off between user privacy and data utility in the context of home-work carpooling. The analyses allow a careful quantification of the effects of different privacy-preservation techniques and of their granularity on total VMT, showing that better VMT values can be obtained if users agree to trade convenience with privacy, more in terms of detour time (i.e. travel time) than time flexibility (i.e. waiting time). For instance, with a value of detour time Δ of at least 5 minutes, the VMT increases only by less than 10% in case of privacy-preservation. Thus, by compromising on convenience, it is possible to preserve privacy by only minimally impacting VMT. This observation might be especially useful for city authorities and policy makers to find a good compromise between the citizens' individual right to privacy, and the societal need of reducing VMT. For instance, introducing more flexibility in working hours could facilitate the achievement of the above compromise in urban contexts.

In terms of the relative strength of the different privacy-preserving techniques considered, k-anonymity is found to consistently outperform obfuscation, while cloaking becomes the most effective method when the spatial granularity of k-anonymity becomes large.

The focus of this paper has been on masking both origin and destination of trips. However, if the carpooling system is provided by the employer, it is clear that the real work destination could be used, implying that privacy-preservation could be applied only to one end of the trip. This is one avenue for our future work.

Another interesting future study could also be the application of privacy to the temporal dimension of the problem, namely by masking the time of departure. The data-masking techniques used for this study have been selected since they offer the option of changing the granularity of privacy and capture the trade-off we were looking for. For the future, we plan to evaluate and compare more sophisticated location privacy and anonymization techniques, such as differential privacy, or encryption and protocol-based methodologies, such as those introduced in (Aïvodji et al., 2018; Luo et al., 2019).

Users are becoming more and more concerned and aware of privacy issues, especially after data breaches happen. Currently, users grant whole data ownership and rights to the companies, since otherwise they would not be able to use the specific service/technology. If the current scenario will change (e.g. for new regulations), companies should start offering users benefits and rewards (e.g. lower cost, priority, higher score, etc.) to nudge them to fully or partially opt-out from the "privacy option", allowing the system to fully access their location data, or reduce the level of privacy users were initially granted. If the user could set a desired level of privacy or decide to do not require privacy at all, this will lead to different levels of data privacy within the same privacy-preserving system. Performing tests on the sensitivity of the system efficiency and QoS with respect to the riders penetration rate and their geographical distribution could be another interesting research direction to investigate.

References

Aïvodji, U. M., Gambs, S., Huguet, M.-J., and Killijian, M.-O. (2016). Meeting points in ridesharing: A privacy-preserving approach. *Transportation Research Part C: Emerging Technologies*, 72:239–253.

Aïvodji, U. M., Huguenin, K., Huguet, M.-J., and Killijian, M.-O. (2018). Sride: A privacy-preserving ridesharing system. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '18, pages 40–50, New York, NY, USA. ACM.

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. (2013). Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 901–914, New York, NY, USA. ACM.

Baik Hoh, Gruteser, M., Hui Xiong, and Alrabady, A. (2006). Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5(4):38–46.

Bamba, B., Liu, L., Pesti, P., and Wang, T. (2008). Supporting anonymous location queries in mobile environments with privacygrid. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 237–246, New York, NY, USA. ACM.

Bettini, C., Mascetti, S., Wang, X. S., Freni, D., and Jajodia, S. (2009). *Anonymity and Historical-Anonymity in Location-Based Services*, pages 1–30. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bilogrevic, I., Jadliwala, M., Joneja, V., Kalkan, K., Hubaux, J., and Aad, I. (2014). Privacy-preserving optimal meeting location determination on mobile devices. *IEEE Transactions on Information Forensics and Security*, 9(7):1141–1156.

Blumenberg, E. and Smart, M. (2010). Getting by with a little help from my friends and family: immigrants and carpooling. *Transportation*, 37(3):429–446.

Buttyán, L., Holczer, T., and Vajda, I. (2007). On the effectiveness of changing pseudonyms to provide location privacy in vanets. In Stajano, F., Meadows, C., Capkun, S., and Moore, T., editors, *Security and Privacy in Ad-hoc and Sensor Networks*, pages 129–141, Berlin, Heidelberg. Springer Berlin Heidelberg.

Calacci, D., Berke, A., Larson, K., and Pentland, A. (2019). The tradeoff between the utility and risk of location data and implications for public good. *arXiv*, 1905.09350.

Cheng, R., Zhang, Y., Bertino, E., and Prabhakar, S. (2006). Preserving user location privacy in mobile data management infrastructures. In Danezis, G. and Golle, P., editors, *Privacy Enhancing Technologies*, pages 393–412, Berlin, Heidelberg. Springer Berlin Heidelberg.

City of Pisa (2016). Mobilitando Pisa Initiative. Last accessed: August 24th 2017.

Cleary, G. (2018). Mobile privacy: What do your apps know about you? Last accessed: February 2020.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1376).

Duckham, M. and Kulik, L. (2005). A formal model of obfuscation and negotiation for location privacy. In *Proceedings of the Third International Conference on Pervasive Computing*, PERVASIVE'05, pages 152–170, Berlin, Heidelberg. Springer-Verlag.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eckhoff, D., German, R., Sommer, C., Dressler, F., and Gansen, T. (2011). Slotswap: strong and affordable location privacy in intelligent transportation systems. *IEEE Communications Magazine*, 49(11):126–133.

Feiler, L. (2010). The legality of the data retention directive in light of the fundamental rights to privacy and data protection. *European Journal of Law and Technology*, 1(3).

Gao, J., Sun, L., and Cai, M. (2019). Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data. *Transportation Research Part C: Emerging Technologies*, 104:78 – 94.

Gedik, B. and Liu, L. (2005). Location privacy in mobile systems: A personalized anonymization model. In *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, pages 620–629.

Goel, P., Kulik, L., and Ramamohanarao, K. (2016). Privacy-aware dynamic ride sharing. *ACM Trans. Spatial Algorithms Syst.*, 2(1):4:1–4:41.

Gruteser, M. and Grunwald, D. (2003). Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*, MobiSys '03, pages 31–42, New York, NY, USA. ACM.

Ho, S.-S. and Ruan, S. (2011). Differential privacy for location pattern mining. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, SPRINGL '11, pages 17–24, New York, NY, USA. ACM.

Hoh, B. and Gruteser, M. (2005). Protecting location privacy through path confusion. In *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05)*, pages 194–205.

ISTAT, Istituto Nazionale di Statistica (2011). Basi territoriali. Last accessed: May 2019.

Jennifer Valentino-DeVries, Natasha Singer, M. H. K. and Krolik, A. (2018). Our apps know where you were last night, and they're not keeping it secret. Last accessed: February 2020.

Kido, H., Yanagisawa, Y., and Satoh, T. (2005). Protection of location privacy using dummies for location-based services. In *21st International Conference on Data Engineering Workshops (ICDEW'05)*, pages 1248–1248.

Li, M., Zhu, L., and Lin, X. (2019). Efficient and privacy-preserving carpooling using blockchain-assisted vehicular fog computing. *IEEE Internet of Things Journal*, 6(3):4573–4584.

Lian, S., Cha, T., and Xu, Y. (2019). Enhancing geotargeting with temporal targeting, behavioral targeting and promotion for comprehensive contextual targeting. *Decision Support Systems*, 117:28 – 37.

Liao, F., Molin, E., Timmermans, H., and van Wee, B. (2018). Carsharing: the impact of system characteristics on its potential to replace private car trips and reduce car ownership. *Transportation*, Online:1–36.

Librino, F., Renda, M. E., Resta, Giovanni Santi, P., Duarte, F., Ratti, C., and Zhao, J. (2018). Social mixing and home-work carpooling (extended abstract). In *Proceedings of the Transportation Research Board 97th Annual Meeting*, pages 1–5.

Librino, F., Renda, M. E., Santi, P., Martelli, F., Resta, G., Duarte, F., Ratti, C., and Zhao, J. (2019). Home-work carpooling for social mixing. *Transportation: Planning - Policy - Research - Practice*. To appear.

Luo, Y., Jia, X., Fu, S., and Xu, M. (2019). pRide: Privacy-preserving ride matching over road networks for online ride-hailing service. *IEEE Transactions on Information Forensics and Security*, 14(7):1791–1802.

Ma, C. Y. T., Yau, D. K. Y., Yip, N. K., and Rao, N. S. V. (2013). Privacy vulnerability of published anonymous mobility traces. *IEEE/ACM Transactions on Networking*, 21(3):720–733.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24.

Maiouak, M. and Taleb, T. (2019). Dynamic maps for automated driving and uav geofencing. *IEEE Wireless Communications*, 26(4):54–59.

Mokbel, M. F., Chow, C.-Y., and Aref, W. G. (2006). The new casper: Query processing for location services without compromising privacy. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 763–774. VLDB Endowment.

Nabil, M., Sherif, A., Mahmoud, M., Alsharif, A., and Abdallah, M. (2019). Efficient and privacy-preserving ridesharing organization for transferable and non-transferable services. *IEEE Transactions on Dependable and Secure Computing*, pages 1–1.

Neoh, J. G., Chipulu, M., and Marshall, A. (2017). What encourages people to carpool? an evaluation of factors with meta-analysis. *Transportation*, 44(2):423–447.

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., and Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareabilityn etworks. *Proceedings of the National Academy of Sciences*, 111(37):13290–13294.

Shankar, P., Ganapathy, V., and Iftode, L. (2009). Privately querying location-based services with sybilquery. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 31–40, New York, NY, USA. ACM.

Sherif, A. B. T., Rabieh, K., Mahmoud, M. M. E. A., and Liang, X. (2017). Privacy-preserving ride sharing scheme for autonomous vehicles in big data era. *IEEE Internet of Things Journal*, 4(2):611–618.

Stenneth, L. and Yu, P. S. (2010). Global privacy and transportation mode homogeneity anonymization in location based mobile systems with continuous queries. In *6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010)*, pages 1–10.

Sucasas, V., Mantas, G., Saghezchi, F. B., Radwan, A., and Rodriguez, J. (2016). An autonomous privacy-preserving authentication scheme for intelligent transportation systems. *Computers & Security*, 60:193–205.

Sun, Z., Zan, B., Ban, X. J., and Gruteser, M. (2013). Privacy protection method for fine-grained urban traffic modeling using mobile sensors. *Transportation Research Part B: Methodological*, 56:50 – 69.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570.

Tachet des Combes, R., Sagarra Pascual, O. J., Santi, P., Resta, G., Szell, M., Strogatz, S. H., and Ratti, C. (2017). Scaling law of urban ride sharing. *Nature Scientific Reports*, 7(42868):web.

The Manifest (2019). Data privacy concerns: An overview for 2019. Last accessed: June 2019.

Xi, Y., Schwiebert, L., and Shi, W. (2014). Privacy preserving shortest path routing with an application to navigation. *Pervasive and Mobile Computing*, 13:142–149.

Xue, M., Kalnis, P., and Pung, H. K. (2009). Location diversity: Enhanced privacy protection in location based services. In Choudhury, T., Quigley, A., Strang, T., and Suginuma, K., editors, *Location and Context Awareness*, pages 70–87, Berlin, Heidelberg. Springer Berlin Heidelberg.

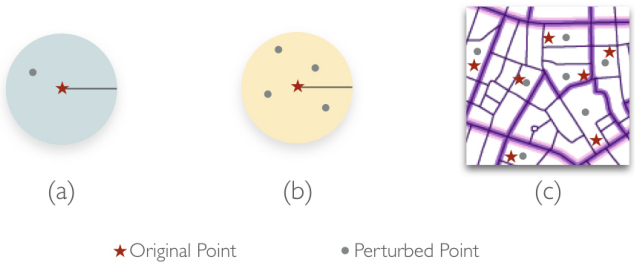
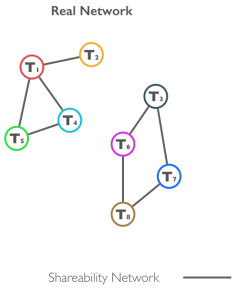
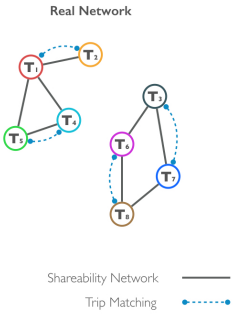


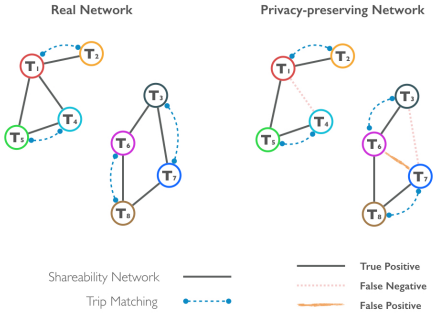
Figure 1. Example of privacy control techniques applied to the geographical coordinates of carpoolers: (a) Obfuscation; (b) K-Anonymity; (c) Cloaking.



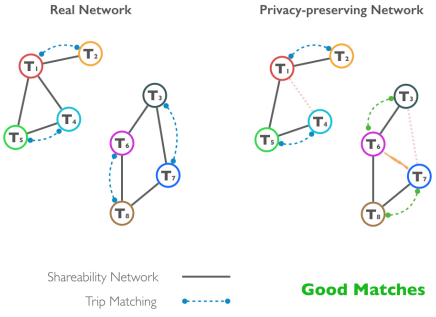
(a) SN with original trip data.



(b) Matching (dotted lines) over the original SN.

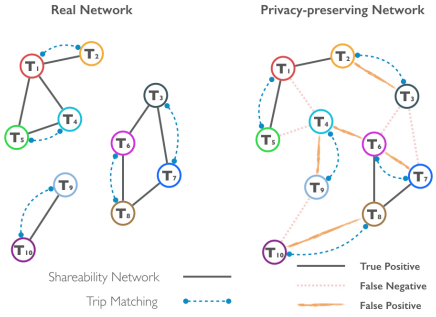


(c) The real SN and corresponding matching (on the left), and the SN and correspondent matching obtained by masking trip data (on the right). In this case only true positive are selected for the matching.

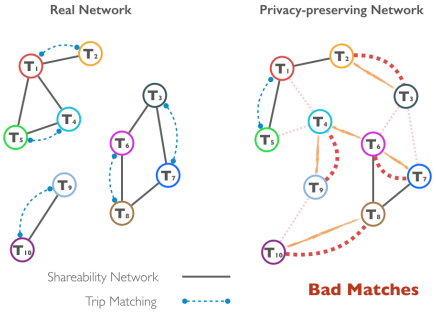


(d) Good matches are trips present in the original SN not selected for the original matching but selected for the privacy-preserving one.

Figure 2. Example of the optimal matching over the real Shareability Network (SN) and of a good matching over the privacy-preserving SN.



(a) SN (straight line) and matching (dotted lines) for the real data (on the left), and for the privacy-preserving data (on the right): some of the selected links for the matching in the latter are false positive, i.e. not present in the original network.



(b) Bad matches (bold dotted lines on the right) are trips not shareable in the original network because one or more of the shareability conditions were not satisfied.

Figure 3. Example of a bad matching over the privacy-preserving shareability network (SN) compared to the real SN and correspondent matching.

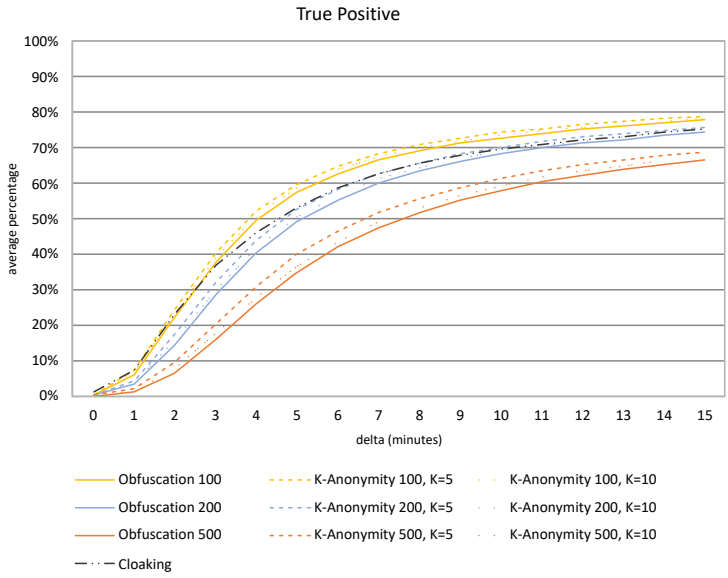


Figure 4. Percentage of true positive in the privacy-preserving shareability networks, i.e. trips identified as shareable in both networks.

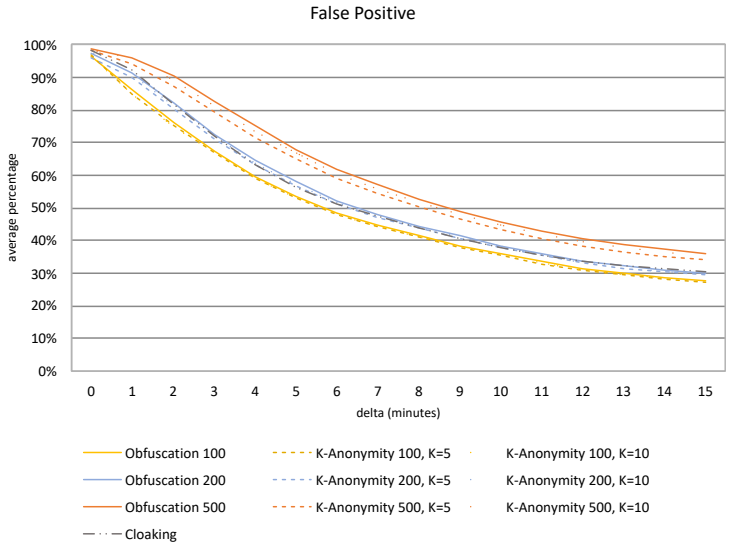


Figure 5. Percentage of false positive in the privacy-preserving SNs, i.e. trips identified as shareable with the masked data, but as not shareable with the real data.

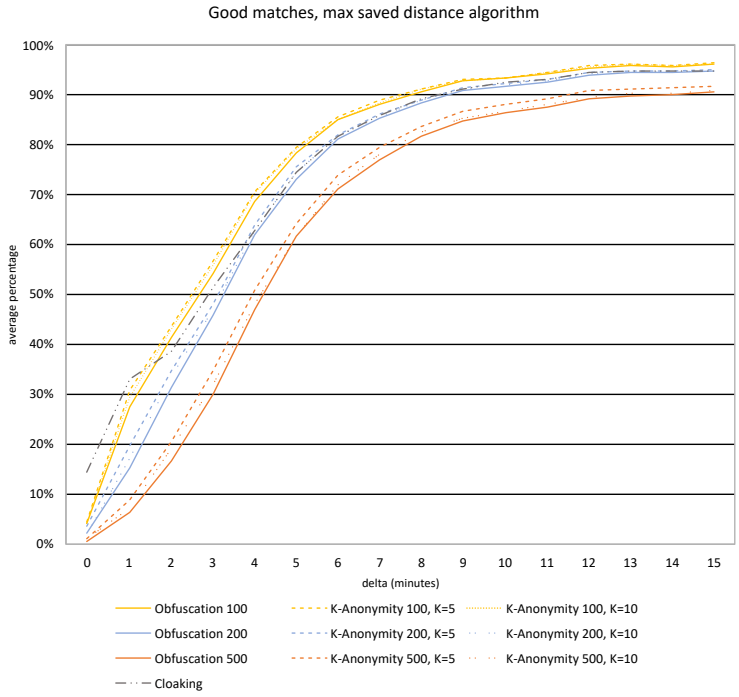


Figure 6. Percentage of good matches in the privacy-preserving SNs, i.e. trips identified as shareable in both networks, and selected for the optimal matching in the privacy-preserving SNs.

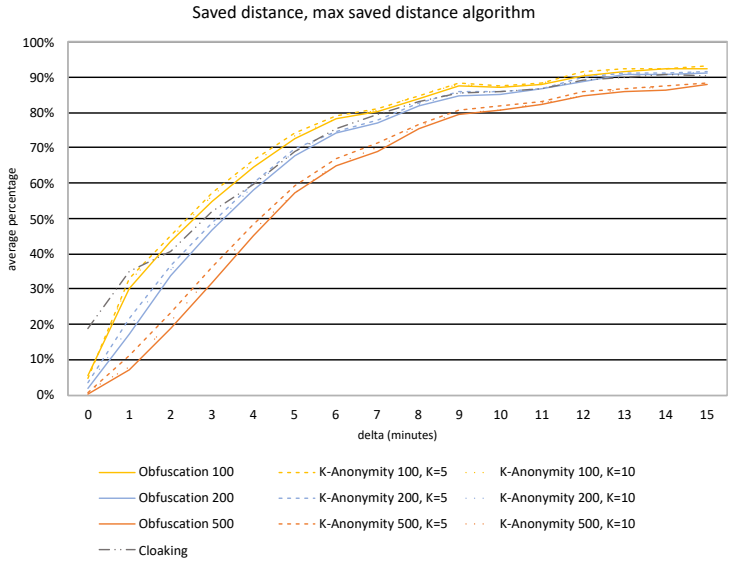


Figure 7. Matching saved traveled distance achievable in the privacy-preserving SN when only considering good matches, savings reported in terms of percentage with respect to the real SN.

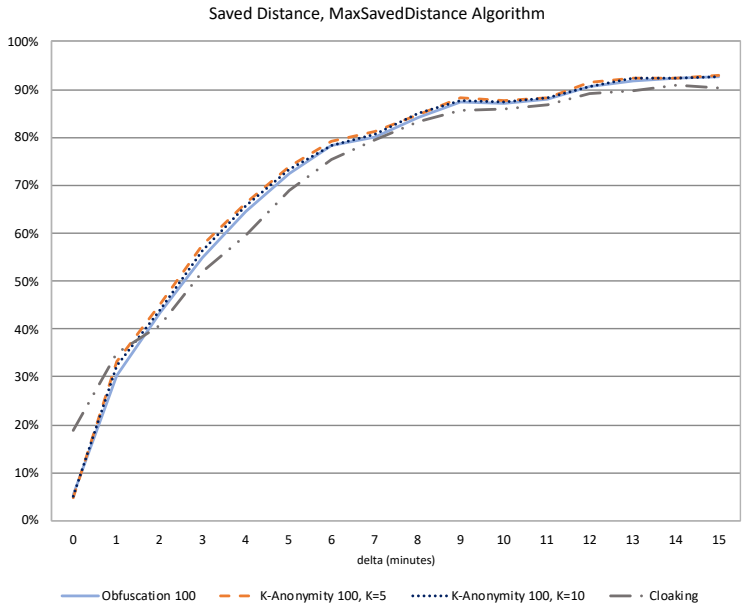


Figure 8. Comparing 100m radius-based algorithms and cloaking saved traveled distance over the privacy-preserving SN.

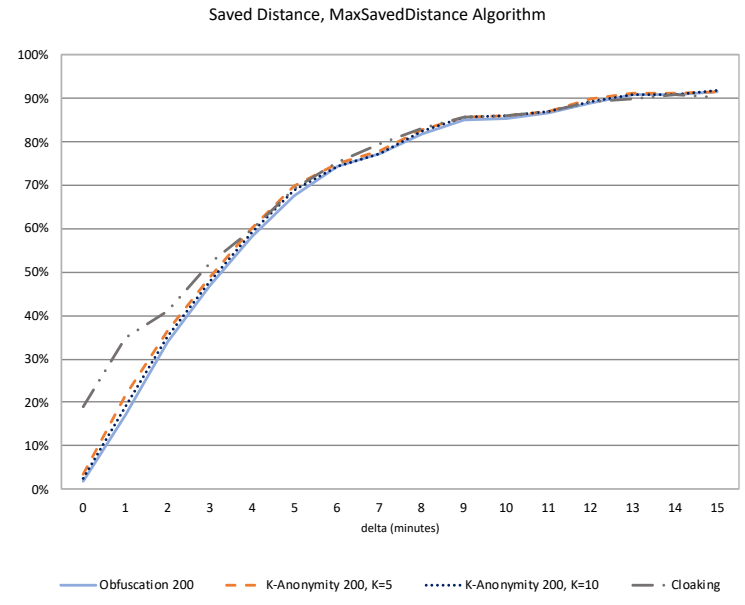


Figure 9. Comparing 200m radius-based algorithms and cloaking saved traveled distance over the privacy-preserving SN.

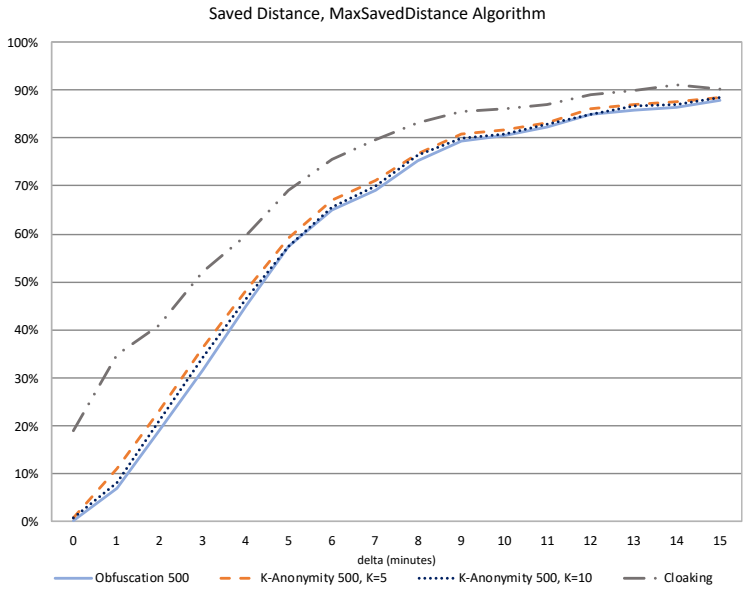


Figure 10. Comparing 500m radius-based algorithms and cloaking saved traveled distance over the privacy-preserving SN.

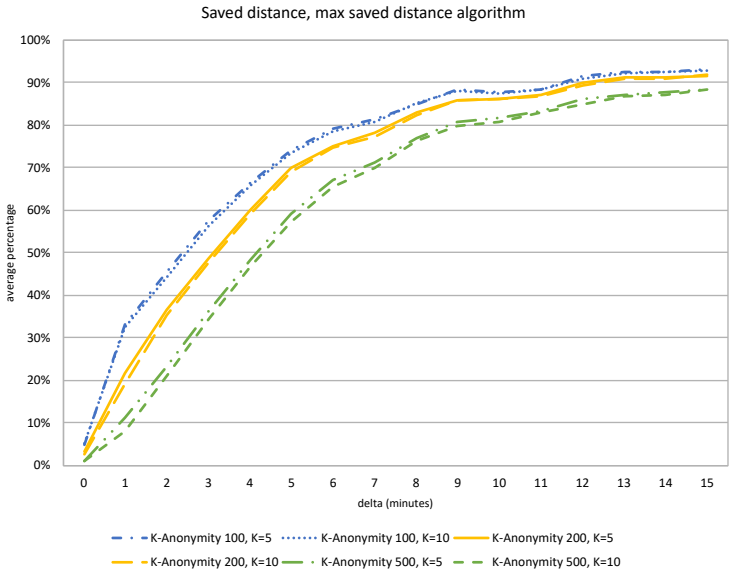


Figure 11. Comparing different granularity of privacy control within the same algorithm, K-Anonymity, in terms of saved traveled distance.

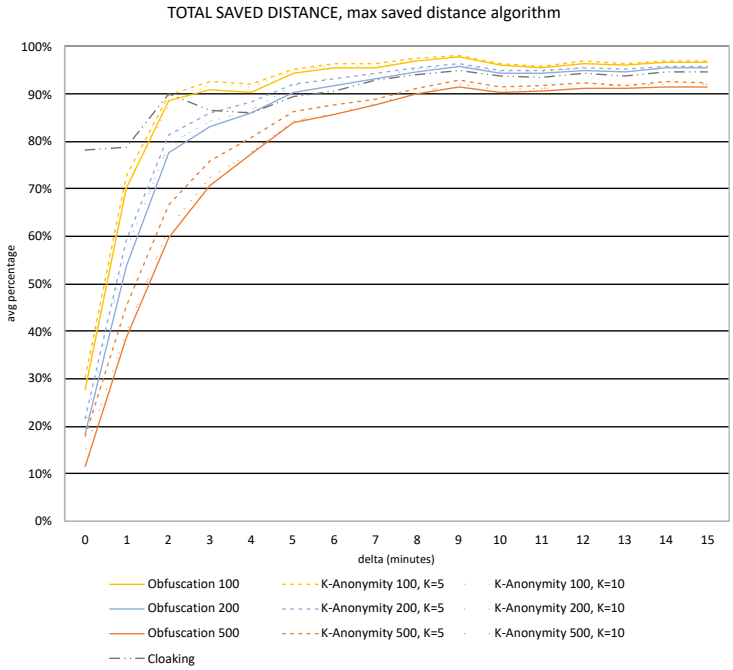


Figure 12. Matching saved traveled distance achievable over the privacy-preserving SN when also considering the bad matches.

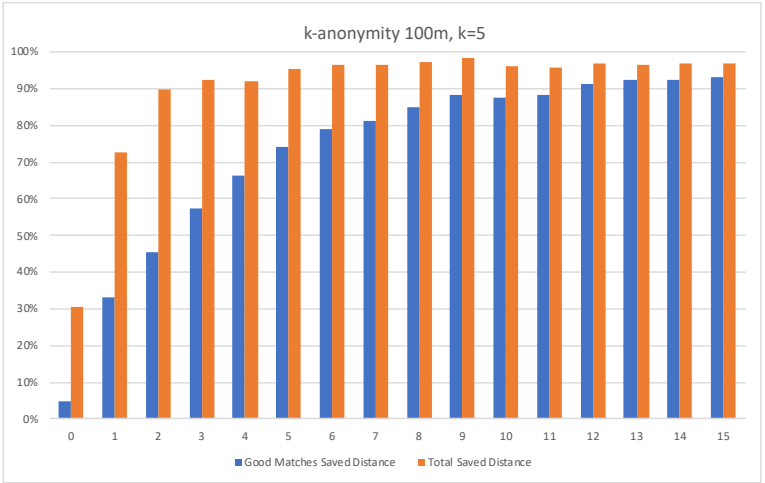


Figure 13. Comparing saved traveled distance and total saved traveled distance achievable over the privacy-preserving SN for k-anonymity with radius 100m and k=5.

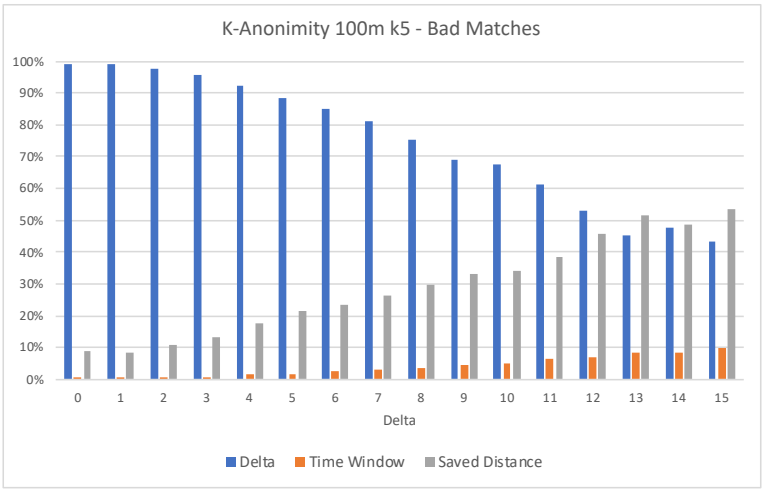


Figure 14. K-anonymity with radius 100m and k=5: percentage of bad matches due to incresed detour time Δ, to non-overlapping time windows, and/or increased distance for the shared trip when compared to the sum of the solo rides.

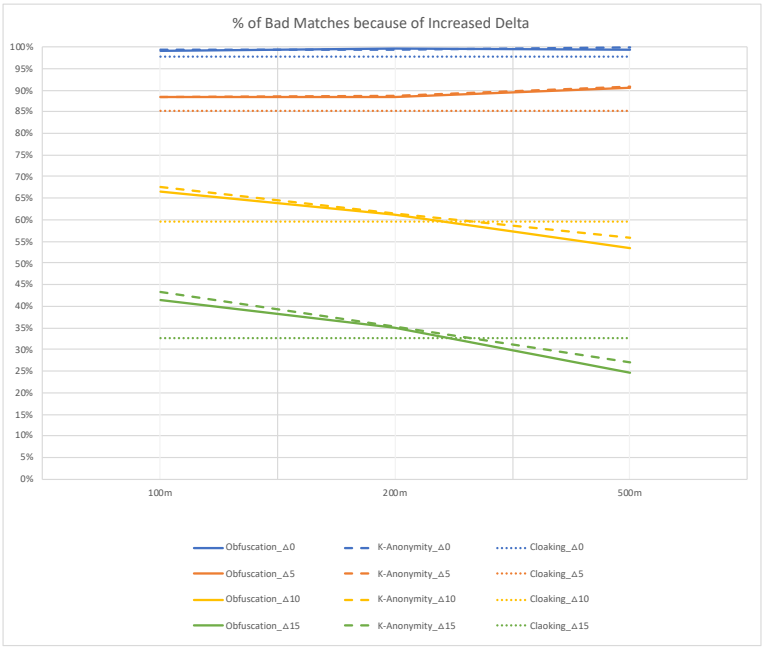


Figure 15. Percentage of bad matches present in the privacy-preserving matching due to increased detour time delta for the shared trip w.r.t. the nominal one.

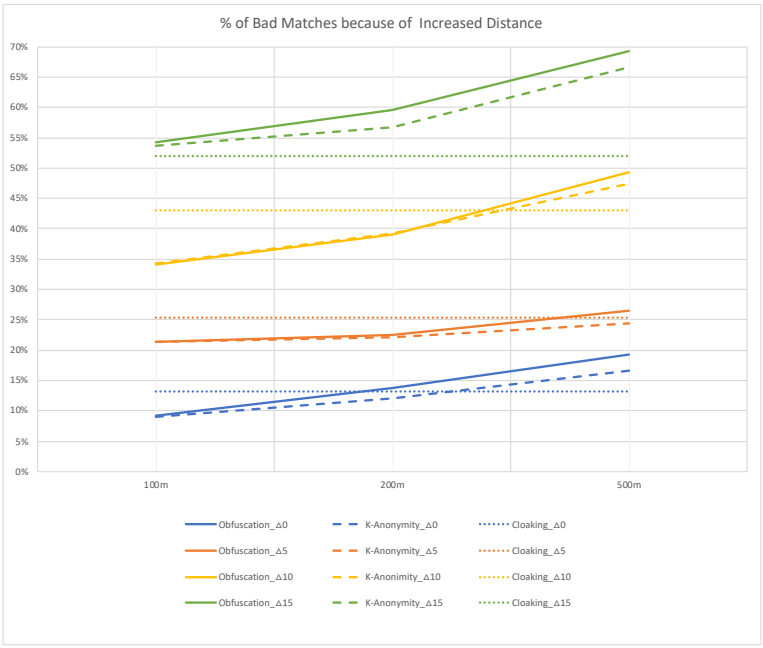


Figure 16. Percentage of bad matches present in the privacy-preserving matching due to increased distance for the shared trip wrt the single trips.