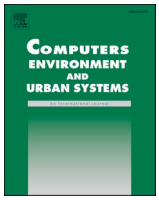
Computers, Environment and Urban Systems 110 (2024) 102091

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Understanding the protection of privacy when counting subway travelers through anonymization

Nadia Shafaeipour [a], Valeriu-Daniel Stanciu [b], Maarten van Steen [b], Mingshu Wang [c,*]

[a] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, the Netherlands
[b] Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, the Netherlands
[c] School of Geographical & Earth Sciences, University of Glasgow, United Kingdom

ARTICLE INFO

Keywords:
K-anonymity
Travelers
Privacy preservation
General data protection regulation
Public transportation

ABSTRACT

Public transportation, especially in large cities, is critical for livability. Counting passengers as they travel between stations is crucial to establishing and maintaining effective transportation systems. Various information and communication technologies, such as GPS, Bluetooth, and Wi-Fi, have been used to measure people's movements automatically. Regarding public transportation applications, the automated fare collection (AFC) system has been widely adopted as a convenient method for measuring passengers, mainly because it is relatively easy to identify card owners uniquely and, as such, the movements of their card holders. However, there are serious concerns regarding privacy infringements when deploying such technologies, to the extent that Europe's General Data Protection Regulation has forbidden straightforward deployment for measuring pedestrian dynamics unless explicit consent has been provided. As a result, privacy-preservation techniques (e.g., anonymization) must be used when deploying such systems. Against this backdrop, we investigate to what extent a recently developed anonymization technique, known as detection k-anonymity, can be adapted to count public transportation travelers while preserving privacy. In the case study, we tested our methods with data from Beijing subway trips. Results show different scenarios when detection k-anonymity can be effectively applied and when it cannot. Due to the complicated relationship between the detection k-anonymity parameters, setting the proper parameter values can be difficult, leading to inaccurate results. Furthermore, through detection k-anonymity, it is possible to count travelers between two locations with high accuracy. However, counting travelers from more than two locations leads to more inaccurate results.

1. Introduction

Residents and visitors depend on public transportation in cities and towns worldwide. Analyzing public transportation data helps in understanding and improving services. One beneficial use case is counting passengers as they move between locations. Measuring passenger movements has now become relatively simple for many modern public transportation systems, as users check in and out of subways and buses using customized smart cards. The information obtained from measuring passengers' behavior is essential for overall fleet management and effective transport scheduling, leading to improving the quality and reliability of the public transportation service, identifying travel patterns, or emergency preparedness (Boreiko & Teslyuk, 2016; Brauer, Mäkinen, Forsch, Oksanen, & Haunert, 2022; Dunlap, Li, Henrickson, & Wang, 2016; Patlins & Kunicina, 2015; Wirz et al., 2012).

In order to measure passenger movements, their locations must be known. The growing use of location-enabled technologies allows an increasing number of parties to access this information. Consequently, people are concerned about their geoprivacy. Geoprivacy is a subset of information privacy that involves a person's right to decide how, when, and to what extent location data about himself or herself is shared (Billen, Joao, & Forrest, 2006). Unfortunately, many people have limited knowledge of how the underlying technology for using location information works, such as what can(not) be inferred from an individual's location over time. Asking for a person's consent can therefore be asking something too complicated. We take the standpoint that privacy should be protected upfront such that no consent is needed: no party has, by design, access to a person's sensitive geographical data.

* Corresponding author at: School of Geographical & Earth Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom.
E-mail addresses: z.shafaeipoursarmoor@utwente.nl (N. Shafaeipour), v.stanciu@utwente.nl (V.-D. Stanciu), m.r.vansteen@utwente.nl (M. van Steen), Mingshu.Wang@glasgow.ac.uk (M. Wang).

https://doi.org/10.1016/j.compenvurbsys.2024.102091
Received 11 June 2023; Received in revised form 30 December 2023; Accepted 17 February 2024
Available online 14 March 2024

In the case of measuring passenger behavior for public transportation, the main challenge faced is a significant risk of privacy violations when using smart cards (Ardagna et al., 2008; Asadpour & Dashti, 2011). This is because each smart card is individually recognizable. In other words, anonymization of the data on a card is not enough. These cards often contain personal information, such as card identifiers, travel patterns and locations visited. Consequently, the public needs to trust the organizations that provide the cards and those that use their data for further analyses. For this reason, the analyses of the data extracted from the use of such cards are generally strictly regulated by privacy laws, such as Europe's General Data Protection Regulation (GDPR) (Georgiadou, de By, & Kounadi, 2019; Voss, 2016).

The current strategies for preserving privacy rely on replacing actual identifiers with pseudonyms, which still allows tracking over time and space. In 2013, the largest Japanese train company announced its intention to sell its passenger dataset to third-party companies (Avoine, Calderoni, Delvaux, Maio, & Palmieri, 2014; Geuss, 2013); they planned to anonymize the data by replacing sensitive information such as the card owner's name and residence with an anonymous ID. Obviously, this is not enough to protect privacy. By simply analyzing patterns of an individual card and combining those patterns with other datasets, it has been shown that identifying an individual is still possible (Avoine et al., 2014). Numerous studies have been conducted on re-identifying formerly anonymized individuals, and they have shown that it is often not difficult to do so (Avoine et al., 2014; Fechner & Kray, 2012). It has been demonstrated that auxiliary data can be used to re-identify individuals in datasets that appeared perfectly anonymized on their own (El Emam, Jonker, Arbuckle, & Malin, 2011; Fechner & Kray, 2012). More is needed. Machanavajjhala, Kifer, Gehrke, and Venkitasubramaniam (2007) proposed a new concept of privacy, referred to as l-diversity, which necessitates that the distribution of a sensitive attribute in each equivalence class has a minimum of 'well represented' values. One issue with l-diversity is that it is restricted in its supposition of adversarial knowledge.

T-closeness (Li, Li, & Venkatasubramanian, 2006) is a technique that attempts to make the distribution of a sensitive attribute within any group of individuals similar to the overall dataset distribution, with the disparity limited to a certain threshold (t). Despite its advantages, t-closeness has some drawbacks, such as being sensitive to the chosen threshold (t).

K-anonymity (Samarati & Sweeney, 1998; Wang, Xie, Zheng, & Lee, 2014) and differential privacy (Hutchison et al., 2010; Mir, Isaacman, Cáceres, Martonosi, & Wright, 2013) are two of the more common approaches used in the geospatial sciences to maximize the value of dataset containing location information while minimizing the chances of identifying individuals or groups in the data.

Differential privacy, initially suggested by Dwork (2006), is a powerful technique for safeguarding the privacy of individuals in datasets while still permitting meaningful statistical analysis.The main concept is that the results of mechanisms that are differentially private remain the same, even when there are slight modifications, such as adding or taking away a single item from the dataset. This steadiness of results presents a major difficulty for any attackers who are trying to get information about particular people. Achieving privacy in this framework requires adding noise to the data, with the challenge that more noise is needed for queries involving fewer individuals to maintain the same level of privacy (Dwork, McSherry, Nissim, & Smith, 2006).

However, if we employ schemes where identifiers are randomized independently for each location, the ability to count travelers across multiple locations is compromised. This is because we face the challenge of having to match identifiers. Consider the scenario where an individual, denoted as "X", relocates from A to B and is assigned the random ID "000" at location A and a different random ID "111" at location B. In such instances, the inability to correlate these IDs impedes our ability to identify and count the individual as a traveler who moved from A to B. The key to effective counting across diverse locations is the establishment of a link between identifiers from one location to those from another. Achieving this linkage across different locations proved to be difficult when using techniques such as differential privacy. On the other hand, K anonymity may be more suitable.

K-anonymization is one of the most widely used methods for anonymizing identifiers and also trajectory in geoprivacy (Brauer et al., 2022); the Location Privacy-Preserving Mechanisms (LPPMs) have been developed to achieve k-anonymity for trajectory datasets by generalizing, suppressing, and distorting trajectory data (Shokri, Theodorakopoulos, Le Boudec, & Hubaux, 2011). The k-anonymity algorithm is the basis for many state-of-the-art LPPMs, and it is capable of preventing re-identification attacks. As a formal guarantee of privacy, it has limits.

The purpose of this paper is to explore this method, but not to analyze trajectory data; rather, we used this method to anonymize identifiers. Stanciu et al. (Stanciu, van Steen, Dobre, & Peter, 2020) have developed a technique based on k-anonymity that effectively ensures that every identifier is converted into a pseudonym assigned to at least *k*-1 other card identifiers. Data cannot be traced back to a single individual using this method; instead, data can be traced back only to a group of at least k individuals. We call this technique *detection k-anonymity*.

This paper examines to what extent and under which conditions we can accurately apply detection k-anonymity to counting passengers who travel on a particular subway line from one station to another while ensuring that the data cannot be traced back to an individual. In the case study, we applied detection k-anonymity on a dataset of trips made on the Beijing subway system using smart cards to check in or check out travelers. A trip is defined as the movement of travelers from one station (station A) to another (station B). We used this data as the ground truth to evaluate the balance between the degree of anonymity and accuracy when counting trips to see if this method works on data such as the Beijing dataset.

The paper is organized in the following way: the following section overviews the method used to protect privacy. The third section describes the data used in the case study, and the fourth section introduces the experimental setup and the research findings. The final section concludes the paper.

2. Protecting privacy through detection k-anonymity

K-anonymity is a technique used in data anonymization to protect individual identities in a dataset. It does this by making sure that each record is indistinguishable from at least *k*-1 other records with respect to certain quasi-identifiers. The aim is to create groups of at least *k* records, thus preventing the identification of any particular person. Detection k-anonymity is an advancement of the traditional k-anonymity approach. It takes into account the challenge of preserving anonymity across different single-column databases that contain the same kind of identifiers. Instead of just focusing on creating anonymized groups within one dataset, detection k-anonymity ensures that any combination of these single-column databases maintains k-anonymity for the shared identifiers, and if you combine this dataset later, you still have the same property. This extension is especially useful in situations where multiple datasets with shared identifiers need to be analyzed together, such as the Beijing dataset, while still protecting individual privacy.

There are many methods to measure pedestrian dynamics in public transportation, including manual counting and automatic passenger counting (APC) devices (Patlins & Kunicina, 2015; Tilg, Pawlowski, & Bogenberger, 2021). Another method is the automated fare collection (AFC) system. AFC has established a smart card system that metropolitan governments use worldwide to compute prices for various city transport lines, such as buses and subways. AFC is the method through which the Beijing subway data has been collected.

In a recent paper, smart cards were used to estimate origin-destination demand for public transportation using statistical pattern recognition. The method has been tested on a large dataset of Melbourne's transportation network (Hamedmoghadam et al., 2021).

However, geoprivacy concerns arise using passenger location data for transportation analysis (Keßler & McKenzie, 2018). For a more extensive overview of state-of-the-art geoprivacy, we refer the reader to (McKenzie, Romm, Zhang, & Brunila, 2022; Ogulenko, Benenson, Omer, & Alon, 2021; Swanlund, Schuurman, & Brussoni, 2020).

Another typical data collection tool that has become widely popular is detecting individual mobile devices through the Wi-Fi or Bluetooth signals they transmit. As these signals carry device-identifying information, such as a unique network address, they can, in principle, be used for tracking (Oransirikul, Nishide, Piumarta, & Takada, 2014). This method is deployed, for example, by Transport for London to monitor how passengers travel through the subway.

To demonstrate the significant potential of using WiFi probe request data for understanding mobility patterns in cities, a study (Traunmueller, Johnson, Malik, & Kontokosta, 2018) presents various patterns of mobility in cities. To estimate pedestrian activity, SmartStreetSensor collected WiFi data from mobile devices in 105 UK towns and cities (Trasberg, Soundararaj, & Cheshire, 2021). In order to achieve these objectives, they considered a full version of privacy-preserving tools in their works.

In this study, our goal is to measure pedestrian dynamics in a subway setting and how privacy preservation by detection k-anonymity works in this setting. For an AFC system, every person carries a smart card to use the subway; we have counters that detect the cards when passengers check in or check out, and each card is marked with an identifier that can be read by these counters.

In our approach, we demand that the check-in and check-out counters, which collect identifiers, timestamps, and locations of each card, are responsible for applying anonymization techniques immediately upon detecting a smart card. By collecting card identifiers at each counter during a small timespan and subsequently replacing such an identifier with a *k-anonymous pseudonym*, the system should, in principle, provide a sufficient degree of privacy.

Our description corresponds to how privacy preservation is performed in Wi-Fi detection systems, which led to the development of detection k-anonymity. When a traveler passes a sensor, the traveler's device identifier, a timestamp, and the sensor's identifier are logged (we assume that the actual location of the sensor is known). To successfully anonymize travelers, we collect data during an interval referred to as an epoch (e.g., 5 min). The length of the epoch is a parameter that could be adjusted based on the number of entering passenger. After an epoch has elapsed, we replace each traveler identifier with a pseudonym such that each pseudonym is used for at least k travelers detected during that epoch and record how many travelers have been detected per assigned pseudonym. This information is then sent to a central server.

3. Method

A privacy-preserving AFC passengers-monitoring environment consists of the following:

- A network of subway lines with each line consisting of a source and a destination, and counters at each source and destination gathering card identifiers; a counter acts as a sensor s; all counters form a set S.
- A set of E of N epochs, jointly spanning an elapsed time T during which the system runs, we should have enough data during each epoch to apply anonymization.
- A set *IDS* of M card identifiers detected by our system during T; we assume that each card identifier represents a passenger.

A detection is a triplet $(id, s, e), id \in IDS, s \in S, e \in E$, representing a card uniquely identified by its identifier *id*, sensed by counter s during epoch e (Stanciu et al., 2020). Each detected card identifier is first mapped to an *N-bit pseudonym*, with the *PID* denoting the set of all possible pseudonyms. A pseudonym is derived from a card identifier through secure hashing, establishing that pseudonyms are uniformly distributed in the interval $[0, 2^N)$. We devise an anonymization procedure m to a new set of multipseudonyms *MPID*, such that for each detected $pid \in PID$, there are at least k-1 other detected pseudonyms $\{pid_1, \ldots, pid_{k-1}\} \subset MPID$ with $m(pid) = m(pid_i)$. As mentioned, we assume that each counter stores only multipseudonyms; we guarantee that for each stored multipseudonym, a counter detected at least k different travelers (i.e., pseudonyms) during each epoch. An example of such an anonymization procedure is the truncation operation *trunc(id, nb)* that removes all but the leftmost *nb* bits from the binary representation of *pid*, for all $pid \in PID$. In doing so, we are effectively mapping different pseudonyms to the same multipseudonym. To illustrate, imagine that we truncate psuedonyms to just two bits. The result would be that we have only four multipseudonyms ("00", "01", "10", "11") and that each pseudonym would be mapped to one of these four multipseudonyms.

It should now be clear that to ensure that at least k pseudonyms are mapped to the same multipseudonym, we need to carefully set a value for *nb*. If we keep too many bits, truncation of detected pseudonyms may leave us with multipseudonyms for which there are less than k detected pseudonyms. In that situation, we have no choice but to discard those multipseudonyms, which may significantly affect the accuracy of passenger counts. So, we need to figure out how many bits we need to keep in order to ensure k-anonymity.

As an alternative to discarding multipseudonyms (and thus detected pseudonyms), we deploy a systematic method to map k-anonymity-disobeying detected multipseudonyms and apply that method to all sensors. We addressed this problem with a correction method. Assume there are n disobeying multipseudonyms during an epoch. Each such multipseudonym has less than k detected pseudonyms. We first sort these multipseudonyms in an ascending order and subsequently keep only the first $[n/k]$ ones, systematically evenly spreading the $n-[n/k]$ counts from the discarded multipseudonyms over the multipseudonyms that we keep. Note that each kept multipseudonym will now have an associated count of at least k travelers.

To illustrate, consider the following five disobeying multipseudonyms sets after truncation by keeping four bits $(nb = 4)$ and $k = 2 : \{(0011, 1), (0111, 1), (1011, 1), (1100, 1), (0000, 1)\}$. There is a count of 1 associated with each of these multipseudonyms, which violates the constraint of at least two. To apply the correction, the disobeying multipseudonyms are sorted, leading to $\{0000, 0011, 0111, 1011, 1100\}$. We then keep only the first $[n/k] = [5/2] = 2$ entries, namely $\{0000, 0011\}$, and evenly spread the counts of the other-disobeying multipseudonyms, leading to the multiset $\{0000, 0011, 0000, 0011, 0000\}$, represented as $\{(0000, 3), (0011, 2)\}$.

After anonymizing data at a counter using detection k-anonymity, data is transmitted to a central server. That server houses two types of anonymized data, i.e. one corresponding to checked-in, another one to checked-out trips. Typically, a traveler moving from one station to another receives the same pseudonym due to uniform settings across all locations. However, there is a possibility that individuals may not receive the same pseudonym at different stations, particularly when the identifiers lack k-anonymity, especially in instances involving joiners from other stations. We address this by applying corrections when k-anonymity is not achieved, and this is the place IDs get different pseudonyms. The resulting data can then be effectively utilized for performing a counting method, as detailed below.

4. Dataset

In the case study, we adopted the weekday public transit smartcard records in April 2010 of the Beijing subway (Wang, Zhou, Long, & Chen, 2016; Zhou, Wang, & Long, 2017) to demonstrate how our methods work. This dataset contains 239,728 records that belong to trips that happened during one week. Each record contains a unique card identifier, the day, time, and location at which an individual checked in and later checked out. The smart cards that passengers use are usually

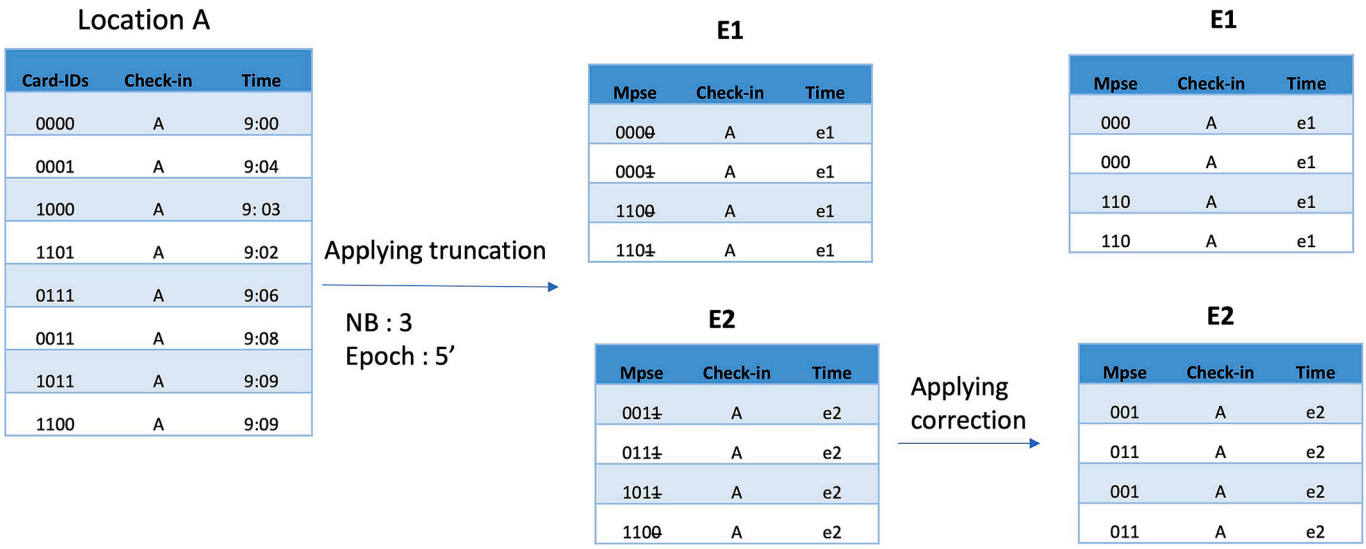


Fig. 1. Applying detection 2-anonymity on the example dataset with the parameter values of $k = 2$, $(nb = 3)$, and epoch: 5 min (Mpse: multipseudonyms).

purchased anonymously through resellers or automated machines and have a unique ID. They are generally unregistered as belonging to a specific individual, so they do not carry any personal information about identities. As a result, apparently, customers can consider these smart cards completely private while also keeping some of the benefits of personal travel permits, such as the capacity to be used many times or some offers from transportation companies for smart card holders. Note that using unique identifiers still allows to derive traveling patterns that may lead to the identification of an individual. It is for this reason that the GDPR renders pseudonymization insuffcient for protecting privacy. These unique smart cards allow us to evaluate the behavior and the number of passengers who travel between stations.

The dataset contains precise information on which card was checked in at a specific location and was later checked out at another given location. In other words, we have accurate ground truth data on actual passenger dynamics. In this sense, the Beijing dataset is much better for evaluating our anonymization method than possible with Wi-Fi-based datasets. Apart from the fact that Wi-Fi detection is subject to many failures (caused by, for example, interferences, erratic detection and transmission ranges, varying signal strengths, and randomization of MAC addresses), attaining the ground truth is extremely difficult. The latter involves knowing which devices are carried by whom and subsequently physically tracking an individual.

5. Experiments

For our experiments, we simulate two scenarios: (1) counting travelers from one location to another and (2) counting travelers from two locations to a common destination. We are conducting this scenario to determine to what extent we can count travelers when they check in at a location and move straight to a destination (A to B). A second scenario involves adding another source to the common destination (A to Z and B to Z) to determine how counting passengers from two different sources interferes with the common destination, as it may be more difficult to reliably associate a multipseudonym at the destination with the original source.

For our goal, counting the number of devices detected at location A during many successive epochs and later at location B over again a series of epochs, we applied detection k-anonymity for different values of k,nb, and different epoch lengths. First, we consider an isolated line, only those trips between two specific locations (A to B). To counting the number of trips between these two locations, each counter applies our privacy preserving algorithm with the same values for all parameters (that is, k,nb, and the epoch length); we consider that each counter stores only pseudonyms and only during the length of an epoch to assign pseudonyms to multipseudonyms subsequently. After applying detection k-anonymity over epoch e, all pseudonyms gathered during e are discarded, and the multipseudonyms, along with their respective counts, are sent to a central server.

To associate multiple pseudonyms with a single multipseudonym, we could ideally apply truncation to the original card identifier. However, truncation works only if we can assume that detected card identifiers are uniformly distributed over the entire possible space of card identifiers. To this end, each counter first applies a globally agreed upon secure hashing function that generates a unique yet uniform random pseudonym for each detected card identifier. We then apply detection k-anonymity on such pseudonyms to produce multipseudonyms. The uniform distribution of pseudonyms guarantees that when constructing a multipseudonym by truncation, there is no built-in bias toward which multipseudonyms are constructed, nor is there a bias toward the actual number of associated pseudonyms for each multipseudonym. A counter keeps track of how many pseudonyms have been assigned to a single multipseudonym to later send the pairs *multipseudonym, number of detections* to a central server.

5.1. Simulated environment for one line

To get a clear understanding of the behavior of the anonymization process, we tested the design on subway trips from Beijing in various settings. As the dataset was relatively sparse, we pretended that all registered trips occurred on the same day.

Many parameters shape the experiments, such as values of k, the truncation parameter nb, and the epoch length. We tested various values for each parameter during our experiment to examine in which situations we still have high accuracy in counting detected devices on a specific line.

Fig. 1 shows how detection k-anonymity works on an example of eight trips between stations A and B. We perform this experiment for epochs lasting 5 min. We keep three bits $(nb = 3)$ and set $k = 2$. In our example, we consider two successive epochs. Applying truncation during the first epoch e_1, keeping only the three leftmost bits, transforms the detected set of pseudonyms $\{0000, 0001, 1100, 1101\}$ to the multipseudonyms $\{(000, 2), (110, 2)\}$. In other words, we record that we have detected multipseudonym 000 through two actual pseudonyms. The same holds for multipseudonym 110.

The situation is different for epoch e_2 where we have the pseudonyms

Table 1
Matching trips between two locations, A and B, based on the data on the central server, resulted in counting eight trips.

Mpse -in	L1-A	Mpse-out	L2-B	Mpse	L1	L2
0000 0000	A	0000 0000	B	0000 0000	A	B
0000	A	0000 0000	B	0000 0000	A	B
1111	A	0111 1111	B	0111 1111	A	B
1111	A	0111 1111	B	0111 1111	A	B
0110	A	1011 0110	B	1011 0110	A	B
0110	A	1011 0110	B	1011 0110	A	B
0001	A	1100 0001	B	1100 0001	A	B
1100 0001	**A**	_	_	_	_	_

Table 2
The number of epochs and devices recorded during each epoch for locations A and B (check-in/out).

location A	A1	A2	A3	A4	A5	A6	–	–
N-trips	107	88	87	86	83	94	–	–
Location B	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**	**B9**	**B10**
N-trips	38	83	91	90	89	77	71	5

$\{0011, 0111, 1011, 1100\}$. If we similarly apply truncation, we are left with the multipseudonyms $\{(001,1),(011,1),(101,1),(110,1)\}$. Each of these multipseudonyms has an associated count of 1, i.e., each disobeys the constraint that the count should be at least 2. Applying the correction, we keep, after sorting, only the first $\lceil n/k \rceil = \lceil 4/2 \rceil = 2$ entries, and evenly spread the counts of the other-disobeying multipseudonyms, leading to the set $\{(001,2),(011,2)\}$. Note that in this way, we have not lost any counts (the total count during e_2 is still 4).

After applying detection k-anonymity at the end of each epoch, the anonymized data is sent to a central server. At the central server, we have two tables containing the data received from the check-in and check-out locations. The question arises, how do we count the number of people going from one location to another? We do so using a simple matching algorithm: see if a multipseudonym during an epoch at location A has also been recorded during an epoch at location B. The algorithm for matching is shown in Table 1. After applying detection k-anonymity, but now for illustration purposes with a large value for nb ($k = 2, nb = 8$), we have two tables, one belonging to location A during the epoch e_1 (09:00 to 09:05) and one belonging to location B during the epoch e_3 (09:10 to 09:15). Here, we can incorporate the average travel time into the counting process. This will give us the range of epochs in which we should expect to see the multipseudonyms from A to B. By knowing the average travel time, we can more easily identify relevant departure and arrival epochs, yet strictly speaking, we need not know the length of the trips. In this example, if we consider an average travel time of 10 min, we expect to find multipseudonyms from A to B during epoch e_3 (10 min after check-in time). We pick the multipseudonym "0000 0000" from location A and start searching to find the same multipseudonym at B; in the first row of Table 1 location B, we have the same multipseudonym; thus, we match this multipseudonym as a trip that has occurred by the multipseudonym "0000 0000" from A to B, we do the same for other multipseudonyms to find a match for them as well. There may be times when it is impossible to match a multipseudonym; for instance, for the second occurrence of "1100 0001" in table A (which we indicated in bold), there is no match anymore for "1100 0001" in Table B, which means this multipseudonym will have arrived during another epoch (or possibly, at another station).

As mentioned, we first take an isolated line from the Beijing subway dataset (M122 to M113) with 545 trips. In Table 2, the number of check-in/outs is shown during each epoch (A1 means location A, epoch e_1) in two tables for locations A and B (check-in/out). The travel time for this line is generally between 23 and 30 min; we take the epoch length as 10 min for both stations. Knowing that the average travel time is at least 23 min, we expect to see travelers from A start to record at B during epoch e_3. For epochs e_1 and $e_2$at location B, we know that the number of devices is zero because travelers have not yet arrived. We chose to use fixed-length epochs for all stations in order to keep our query formulation straightforward and easy to understand. We are mainly looking into the possibility and accuracy of counting travelers. If we had used variable length epochs, it would have made the query more complicated and the analysis more difficult. Additionally, our query objective is to prevent situations where a traveler could appear in multiple places at the same time. The final step is to apply the detection k-anonymity to each epoch.

Table 3
Comparison of detection k-anonymity with the ground truth for different settings: $k = 2$, epoch length 10 min, and the number of bits to keep (NB).

K-2	A1-B3	A2-B4	A3-B5	A4-B6	A5-B7	A6-B8
Ground Truth	37	25	28	24	28	24
NB: 2	38	70	79	78	78	76
NB: 4	38	66	69	68	74	70
NB: 12	30	20	23	24	32	28
NB: 13	34	22	27	36	28	22
NB: 14	32	28	23	20	32	22
NB: 15	38	28	32	26	25	34
NB: 16	32	22	32	18	26	18
NB: 17	38	30	21	20	26	16
NB: 18	38	24	28	24	24	20
NB: 19	34	24	24	22	32	24
NB: 20	36	22	26	24	33	20
NB: 24	34	22	22	24	22	24
NB: 27	28	22	19	18	20	24

Table 3 shows the results of counting detected passengers at location A and later at location B during different epochs; then, according to the matching algorithm, the multipseudonyms are matched as trips between these two locations. We interpret the results as follows: the first row is the number of devices recorded for the ground truth. They checked in during epoch e at A and later checked out during epoch e' at B (we know the actual number of trips because we computed it based on the original card identifiers). In the following, to achieve detection k-anonymity, we kept different numbers of bits to compare the accuracy of our design in various settings with the ground truth. As shown for $k = 2$, when we increase the numbers of bits to keep, our counts come closer to the ground truth in counting the number of trips between two locations.

To further clarify, consider columns A2-B4 (i.e., trips that started at A during epoch e_2 and arrived at B during e_4). We have a known ground truth of 25 trips. For $nb = 2$, the algorithm counted 70 trips, which is considerably higher than the ground truth. We know that 45 out of these 70 trips actually arrived at B during other epochs than e_4 ($70 - 25 = 45$). The reason for our large number of counts is that for $nb = 2$, because of truncation, we have only four multipseudonyms $(00, 01, 10, 11)$ in A and B. Using detection k-anonymity, we have the multiset $\{(00:16),(01:19),(10:28),(11:25)\}$ at A and $\{(00:21),(01:27),(10:22),(11:13)\}$ at B. Our algorithm matched all these multipseudonyms in source and destination; for example, for multipseudonyms "00," we have 16 of them at A and 21 at B, so the algorithm counts 16 (smallest value) trips made by multipseudonym "00" from A to B, and it repeats the same procedure with the other three multipseudonyms, leading to a total of $16 + 19 + 22 + 13 = 70$ trips.

In the case of $nb = 27$, the algorithm counted 22 trips from A2 to B4, which is closer to the ground truth than for $nb = 2$. All bits, and thus pseudonyms, were retained at A and B, so only the correction step of detection k-anonymity was applied. Once k-anonymity was established, the matching algorithm looked for multipseudonyms from A to B. The matching algorithm found only 22 trips out of 25; half of the original multipseudonyms have been replaced with other, smaller multipseudonyms. It became smaller because we first sorted all multipseudonyms and effectively kept only the smallest ones for matching.

To get a better understanding of where these numbers originate from we counted true positives, false positives, false negatives, and true

Table 4
NB:2.

Epoch	GT	N-trips	TP	FP	FN	TN
A1-B3	37	38	37	1	0	70
A2-B4	25	70	25	45	0	63
A3-B5	28	79	28	51	0	59
A4-B6	24	78	24	54	0	62
A5-B7	28	78	28	50	0	55
A6-B8	24	76	24	52	0	70

Table 5
NB:12.

Epoch	GT	N-trips	TP	FP	FN	TN
A1-B3	37	30	15	15	22	70
A2-B4	25	20	11	9	14	63
A3-B5	28	23	14	9	14	59
A4-B6	24	24	12	12	12	62
A5-B7	28	32	16	16	12	55
A6-B8	24	28	13	15	11	70

Table 6
NB:18.

Epoch	GT	N-trips	TP	FP	FN	TN
A1-B3	37	38	22	16	15	70
A2-B4	25	24	12	12	13	63
A3-B5	28	28	17	11	11	59
A4-B6	24	21	12	12	12	62
A5-B7	28	21	12	12	16	55
A6-B8	24	20	10	10	14	70

Table 7
NB:27.

Epoch	GT	N-trips	TP	FP	FN	TN
A1-B3	37	2	16	12	21	70
A2-B4	25	22	11	11	14	63
A3-B5	28	19	9	10	19	59
A4-B6	24	18	9	9	15	62
A5-B7	28	20	10	10	18	55
A6-B8	24	24	12	10	12	70

negatives. We define false and true counts as follows:

- *True positives*: the number of trips we were able to count that actually occurred.
- *False positives*: the number of trips we counted that actually did not occur.
- *True negatives*: the number of trips we did not count, and that indeed did not happen.
- *False negatives*False negatives: the number of trips we did not count but actually occurred (i.e., we missed them).

Table 4: Counting the number of trips for $k = 2$ and four different *NB*, in addition to the number of true positives, false positives, false negatives and true negatives.

The four Tables 4, 5, 6, and 7 show the results of our counting method for four different values of *nb* and $k = 2$. For each table, the first column shows epochs for which we want to count how many people moved from A and arrived at B during those epochs; the second column shows the actual number of trips that happened (i.e., the ground truth) to which the results of our matching algorithm in the third column are compared with. Other columns include true positives, false positives, false negatives, and true negatives. The sum of *TP* and *FP* column values is equal to the number of trips our method counted (N-Trips column),

Table 8
Counting the number of multipseudonym that appear at each epoch of A and later in two epochs of B.

K-2	A1, B3, B4	A2, B4, B5	A3, B5, B6	A4, B6, B7	A5, B7, B8	A6, B8, B9
Ground Truth	95	79	84	82	77	89
NB: 2	106	88	87	86	83	94
NB: 4	103	88	87	86	82	94
NB: 5	93	86	86	86	81	91
NB: 7	87	80	80	78	75	77

where *TP* indicates the number of trips our algorithms counted correctly. False positive represents the number of trips that were mistakenly recorded but did not occur, and these trips do not appear in *GT*. As we mentioned before, when only truncation is performed, such as with $nb = 2$, the number of different multipseudonyms is low at both locations. A small number of multipseudonyms leads to incorrectly matching many trips that did not happen, leading to many false positives. Table 4 illustrates this. The number of false positive trips decreases when the number of bits is increased, as shown in the other tables. By increasing the number of bits to keep, we have more different multipseudonyms, leading to more corrections (because multipseudonyms do not have enough associated trips). When applying corrections, we lose trips as false negatives.

In the correction phase, as we explained before, for having detection k-anonymity, some multipseudonyms will be replaced by others (but allowing counts greater or equal to k). Consequently, we lose some multipseudonyms, and those multipseudonyms are no longer available to match with arrivals or departures. In fact, lost multipseudonyms are matched by using their replacements and thus lead to counting trips that did not occur. These inaccurate matches count as false positives. At the same time, each lost multipseudonym will also mean that we miss trips, leading to false negatives.

Due to the connection between *FPs* and *FNs*, the best scenario is that an equal number of *FNs* will compensate all the *FPs*. In other words, we can achieve the highest level of accuracy by having only a correction phase (and having sufficient data). The most accurate counting occurs when false positives and false negatives are equal or close to each other. In Table 7 where $nb = 27$, for the last row (A6-B8), the *GT* is 24; the algorithm found 24 trips during this epoch. From these 24 trips, 12 counted correctly as *TP*, and 12 of them should be *FP*, so if 12 of them are false positive, then ideally, the same value should be the number of trips that happened, but our algorithm did not count (*FN*). We have $FP = FN = 12$ in this row, so that is the best we can attain.

The value for *TN* is the same in all tables regardless of the detection k-anonymity setting; the reason is that the algorithm cannot count a trip that does not occur between two locations. Therefore, it does not matter how the algorithm is set; the trips that have not happened can not be counted.

We must consider several consecutive epochs at the destination in order to find as many identifiers as possible from those who left during one specific epoch and arrived at the destination. This is because we do not know precisely during which epoch a passenger will actually check out. For example, in Table 2, 107 passengers checked in during epoch e_1 while only 38 passengers appeared during epoch e_3 at B. We can conclude that passengers who left at A during e_1 may have arrived at B during other epochs than e_3. To test this assumption, for each departure epoch e_i at A, and the expected arrival epoch e_{i+2} at B, we also looked at the next epoch e_{i+3} at B and checked potential arrivals from A who left during e_i. In Table 8, we displayed this as follows. For epoch e_1 at A (A1), we looked at arriving multipseudonyms at B during epochs e_3 as well as e_4(denoted as B3 + B4). In this way, we found more multipseudonyms from A at B. We repeated this approach for other departure/arrival epochs.

We also looked at the effect of the epoch length, as shown in Fig. 2.

a

Epoch length: 10 min

Epoch	A-B	GT	N-Trips	TP	FP	FN	TN
A1-B3	107-38	37	38	37	1	0	70
A2-B4	88-83	25	70	25	45	0	63
A3-B5	87-91	28	79	28	51	0	59
A4-B6	86-90	24	78	24	54	0	62

b

Epoch length: 15 min

Epoch	A-B	GT	N-Trips	TP	FP	FN	TN
A1-B3	146-38	37	38	37	1	0	109
A2-B4	136-121	22	121	22	99	0	114
A3-B5	122-143	24	122	24	98	0	98
A4-B6	141-126	26	126	26	100	0	115

Fig. 2. Comparison results for different epoch lengths (10 min, 15 min) from *A* to *B*, nb $= 2$, k $= 2$.

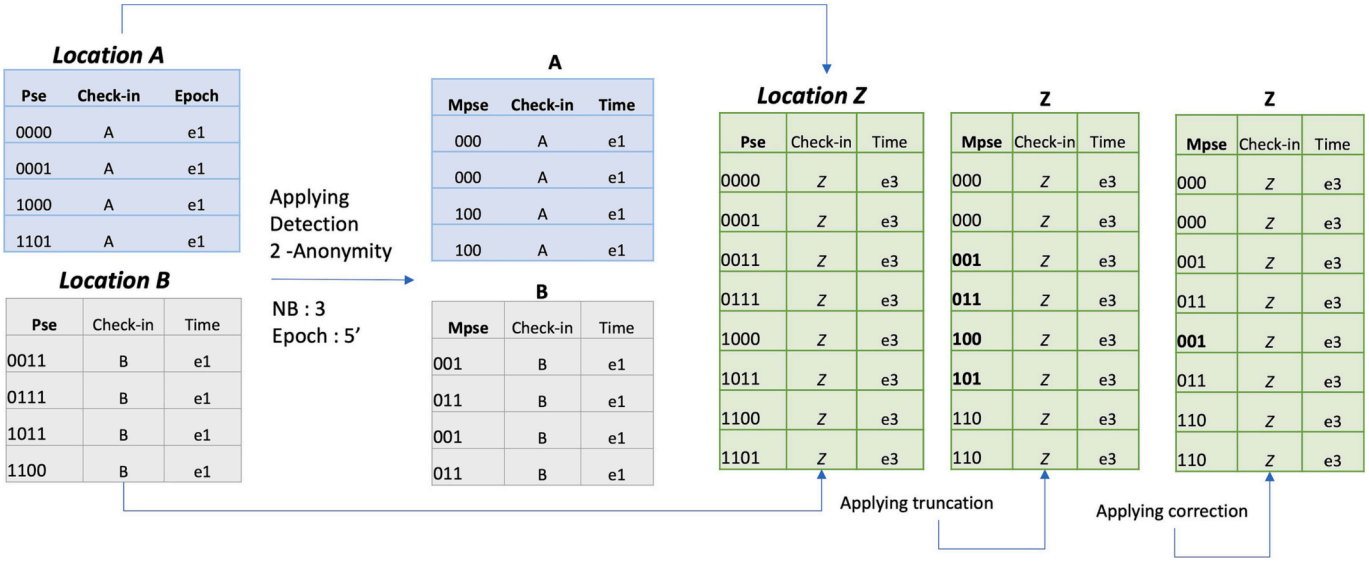


Fig. 3. Matching trips between two locations, A and B, based on the data on the central server, resulted in counting eight trips.

Notably, when epochs are small compared to the expected travel time, the number of travelers that the matching algorithm counts are closer to the ground truth than with large epochs. This happens when *nb* is very low, so all the theoretically existing multipseudonyms are used at both source and destination. In addition, larger epochs have a higher chance of having travelers from earlier epochs, as we witnessed before (Table 8), so we should see more false positives.

Large epochs mean we have more data for applying detection k-anonymity compared to having small epochs, so truncation allows us to keep more bits and requires less correction. However, having large epochs (relative to the time a trip takes) also means correctly matching departures to arrivals becomes more difficult. When applying detection k-anonymity, there is a higher risk of losing identifiers because of fewer data per epoch. If sufficient detections can be guaranteed, epoch lengths become less critical.

5.2. Simulated environment for combining trips

Matters may quickly get out of hand when combining trips: accuracy may drop to unacceptable levels. The problem can be salvaged to a limited extent if we keep a large number of bits. We explore this situation further in this section. First, we consider the situation of counting passengers moving from A to Z, with a counter at location A gathering a set of pseudonyms *PIDA*, mapping them to the multiset *MPIDA*. At destination location Z, we have *PIDZ* and *MPIDZ*, respectively. Now consider that we have another source, B, representing people who travel from B to Z, resulting in a set of pseudonyms *PIDB* and multipseudonyms *MPIDB*, respectively, for counter B. The counter at *Z* detects through *PIDZ* precisely the passengers moving from A to Z and B to Z, respectively. However, there are situations when the multipseudonyms in *MPIDZ*, corresponding to travelers from A, will have been "contaminated" by travelers who moved from B and arrived at Z. Let us look at this situation.

Fig. 3 shows three stations: A, B, and Z; people moved from A and B to later arrive at Z. Based on this example, we want to show how pseudonyms from A at Z are contaminated by travelers who moved from B to Z. At all locations, and for each epoch, we applied detection k-anonymity on locations with $k = 2$ and $nb = 3$. For location A and epoch e_1, truncation alone was enough to reach 2-anonymity; for B (and again, e_1), an additional correction was needed. After passengers arrive at Z, detection k-anonymity resulted in what is shown in green. After truncation, we have four disobeying multipseudonyms: $\{001, 011, 100, 101\}$. We first sort these and apply the correction method. After sorting, the multipseudonyms $\{001, 011\}$ were used to correct 100 and 101, now leading to $\{(001, 2), (011, 2)\}$. However, note that multipseudonym 100 was originally from location A, while multipseudonym 001 originated from B. In other words, the correction yields that we will wrongfully match a trip from A to one coming from B. This mismatch is entirely due to mixing trips from B with those from A: the trips from B are said to contaminate those from A. To further analyze this situation, we gradually add travelers from B (called joiners) to those arriving at Z and coming from A. When keeping only a few bits when applying truncation, There was a large discrepancy between our counting and ground truth because travelers from A may count as

Table 9
Constantly added travelers who moved from B to Z to travelers from A to Z for 2-anonymity and *NB* : 4.

K-2	A1-Z3	B1-Z3	A2-Z4	B2-Z4	A3-Z5	B3-Z5	A4-Z6	B4-Z6	A5-Z7	B5-Z7	A6-Z8	B6-Z8
Ground Truth	37	45	25	40	28	43	24	34	28	38	24	31
0%	28	–	22	–	19	–	18	–	20	–	24	–
20%	40	6	72	11	75	17	81	12	75	14	76	17
50%	40	20	77	34	78	39	83	37	77	42	75	29
100%	41	42	79	78	80	81	84	70	78	78	78	68

Table 10
constantly added travelers who moved from B to Z to travelers from A to Z for 2-anoniymity and *NB* : 27.

K-2	A1-Z3	B1-Z3	A2-Z4	B2-Z4	A3-Z5	B3-Z5	A4-Z6	B4-Z6	A5-Z7	B5-Z7	A6-Z8	B6-Z8
Ground Truth	37	45	25	40	28	43	24	34	28	38	24	31
0%	28	–	22	–	19	–	18	–	20	–	24	–
20%	28	4	22	4	18	6	18	8	22	6	24	6
50%	28	16	22	14	18	26	22	16	22	18	24	14
80%	28	30	22	28	18	40	20	18	20	28	24	16
100%	28	41	22	30	18	48	20	30	20	32	24	28

Table 11
The number of trips during each epoch from A to Z and B to Z compared to Ground Truth.

K-2	A1-Z3	B1-Z3	A2-Z4	B2-Z4	A3-Z5	B3-Z5	A4-Z6	B4-Z6	A5-Z7	B5-Z7	A6-Z8	B6-Z8
Ground Truth	37	45	25	40	28	43	24	34	28	38	24	31
NB: 2	41	42	79	78	80	81	84	70	78	78	78	68
NB: 12	30	47	22	44	18	48	26	34	34	38	32	26
NB: 27	28	41	22	30	18	48	20	30	20	32	24	28

Table 12
Comparison $NB = 2$ with $NB = 27$ for two lines with $K = 2$.

		NB: 2					NB: 27				
Epoch	GT	N-trips	TP	FP	FN	TN	N-trips	TP	FP	FN	TN
A1-Z3	37	82	37	45	0	70	28	14	14	23	70
A2-Z4	25	88	25	63	0	63	22	11	11	14	63
A3-Z5	28	87	28	59	0	59	18	9	9	19	59
A4-Z6	24	86	24	62	0	62	20	11	9	13	62
A5-Z7	28	83	28	55	0	55	20	10	10	18	55
A6-Z8	24	94	24	70	0	70	24	12	12	12	70
B1-Z3	45	81	45	36	0	54	41	21	20	24	54
B2-Z4	40	90	40	50	0	50	30	15	15	25	50
B3-Z5	43	84	43	41	0	41	48	25	23	18	41
B4-Z6	34	73	34	39	0	39	30	16	14	15	39
B5-Z7	38	84	38	46	0	46	32	16	16	22	46
B6-Z8	31	78	31	47	0	47	28	16	12	15	47

travelers from B to Z. This situation is sketched in Table 9. The 2-anonymity detection in Table 10 was improved by retaining all N bits. Our results demonstrate that the algorithm is more effective in counting travelers from A to Z by accurately matching multipseudonym when all bits are kept.

In Table 11, we display the results for different values of $nb, k = 2$, and combine each epoch of Z with the next epoch for counting the number of trips between two lines, A to Z and B to Z. Based on the following two figures; we showed that setting the proper detection k-anonymity parameters can be difficult, leading to inaccurate results. So, by having an acceptable epoch length and many bits to keep, we can obtain a closer count to the *GT* in counting the number of passengers who move from one location to another even if we mix with another departure station, like B (Table 10).

It can be seen in Table 11 that by increasing the number of bits, the results come close to the ground truth, implying a higher degree of accuracy. By keeping all bits and having only the correction phase, the algorithm could count travelers with high accuracy for both lines A to Z and B to Z, which we show in Table 12.

6. Conclusion

Sustainable city planning relies heavily on counting passengers in public transportation systems. Tracking passenger flows can be done in many ways, yet all these approaches have drawbacks. A prevailing concern is the preservation of privacy. The present study was designed to determine the effect of preserving privacy when counting subway travelers. Our objective was to assess the extent and conditions under which detection k-anonymity can accurately count subway passengers while ensuring that individuals cannot be traced from the final dataset. In the current study, comparing the results of our algorithm with the ground truth showed that passengers between two locations can be counted accurately if the detection k-anonymity algorithm is appropriately configured. However, results quickly get worse when combining trips from several departure stations yet having the same destination. This is mainly caused by the inability to match multipseudonyms correctly, as a single multipseudonym at the destination may have been constructed from trips from both origins.

This finding answered the questions of other studies in this area that public transportation companies can record and count passengers and

protect the privacy of individuals. We showed that it is possible to use anonymization techniques that prevent tracing back to an individual. It should be noted, however, that in some situations applying privacy preservation may lead to a severe decrease in counting accuracy. In future work, we plan to explore more recent alternative techniques that may lead to protecting privacy at higher accuracies.

CRediT authorship contribution statement

Nadia Shafaeipour: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **Valeriu-Daniel Stanciu:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – review & editing. **Maarten van Steen:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Mingshu Wang:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Acknowledgments

We acknowledge the constructive feedback received from the editor and anonymous reviewers, which significantly improved the quality of our manuscript. This research was supported in part by the Dutch Research Council (NWO), "Measuring pedestrian dynamics: doing it the right way" (Grant No. 410.19.002).

References

Ardagna, C. A., Stavrou, A., Jajodia, S., Samarati, P., Martin, R., et al. (2008). A multi-path approach for k-anonymity in mobile hybrid networks. In *PILBA'08: Privacy in location-based applications: Workshop co-located with ESORICS 2008: Malaga, Spain, october 9, 2008: Proceedings (pp. 82–101). Null volume 397*.

Asadpour, M., & Dashti, M. T. (2011). A privacy-friendly rfid protocol using reusable anonymous tickets. In *2011IEEE 10th international conference on trust, security and privacy in computing and communications* (pp. 206–213). IEEE.

Avoine, G., Calderoni, L., Delvaux, J., Maio, D., & Palmieri, P. (2014). Passengers information in public transport and privacy: Can anonymous tickets prevent tracking? *International Journal of Information Management, 34*, 682–688.

Billen, R., Joao, E., & Forrest, D. (2006). *Dynamic and mobile GIS: Investigating changes in space and time*. CRC Press.

Boreiko, O., & Teslyuk, V. (2016). Structural model of passenger counting and public transport tracking system of smart city. In *2016 XII international conference on perspective technologies and methods in MEMS design (MEMSTECH)* (pp. 124–126). IEEE.

Brauer, A., Mäkinen, V., Forsch, A., Oksanen, J., & Haunert, J.-H. (2022). My home is my secret: Concealing sensitive locations by context-aware trajectory truncation. *International Journal of Geographical Information Science, 36*, 2496–2524.

Dunlap, M., Li, Z., Henrickson, K., & Wang, Y. (2016). Estimation of origin and destination information from bluetooth and wi-fi sensing for transit. *Transportation Research Record, 2595*, 11–17.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming* (pp. 1–12). Springer.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography: Third theory of cryptography conference, TCC 2006, New York, NY, USA, march 4-7, 2006. Proceedings 3* (pp. 265–284). Springer.

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLoS One, 6*, Article e28071.

Fechner, T., & Kray, C. (2012). Attacking location privacy: Exploring human strategies. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 95–98).

Georgiadou, Y., de By, R. A., & Kounadi, O. (2019). Location privacy in the wake of the gdpr. *ISPRS International Journal of Geo-Information, 8*, 157.

Geuss, M. (2013). *Japanese railway company plans to sell data from e-ticket records. Ars technica*.

Hamedmoghadam, H., Vu, H. L., Jalili, M., Saberi, M., Stone, L., & Hoogendoorn, S. (2021). Automated extraction of origin-destination demand for public transportation from smartcard data with pattern recognition. *Transportation Research Part C: Emerging Technologies, 129*, Article 103210.

Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Kobsa, A., Mattern, F., , … Rangan, C. P., et al. (2010). *Lecture notes in computer science*.

Keßler, C., & McKenzie, G. (2018). A geoprivacy manifesto. *Transactions in GIS, 22*, 3–19.

Li, N., Li, T., & Venkatasubramanian, S. (2006). T-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering* (pp. 106–115). IEEE.

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

McKenzie, G., Romm, D., Zhang, H., & Brunila, M. (2022). Privyto: A privacy-preserving location-sharing platform. *Transactions in GIS, 26*, 1703–1717.

Mir, D. J., Isaacman, S., Cáceres, R., Martonosi, M., & Wright, R. N. (2013). Dp-where: Differentially private modeling of human mobility. In *2013 IEEE international conference on big data* (pp. 580–588). IEEE.

Ogulenko, A., Benenson, I., Omer, I., & Alon, B. (2021). Probabilistic positioning in mobile phone network and its consequences for the privacy of mobility data. *Computers, Environment and Urban Systems, 85*, Article 101550.

Oransirikul, T., Nishide, R., Piumarta, I., & Takada, H. (2014). Measuring bus passenger load by monitoring wi-fi transmissions from mobile devices. *Procedia Technology, 18*, 120–125.

Patlins, A., & Kunicina, N. (2015). The new approach for passenger counting in public transport system. In , *vol. 1. 2015 IEEE 8th international conference on intelligent data acquisition and advanced computing systems: Technology and applications (IDAACS)* (pp. 53–57). IEEE.

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression*.

Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y., & Hubaux, J.-P. (2011). Quantifying location privacy. In *2011 IEEE symposium on security and privacy* (pp. 247–262). IEEE.

Stanciu, V.-D., van Steen, M., Dobre, C., & Peter, A. (2020). K-anonymous crowd flow analytics. In *MobiQuitous 2020-17th EAI international conference on Mobile and ubiquitous systems: Computing, networking and services* (pp. 376–385).

Swanlund, D., Schuurman, N., & Brussoni, M. (2020). Maskmy. Xyz: An easy-to-use tool for protecting geoprivacy using geographic masks. *Transactions in GIS, 24*, 390–401.

Tilg, G., Pawlowski, A., & Bogenberger, K. (2021). The impact of data characteristics on the estimation of the three-dimensional passenger macroscopic fundamental diagram. In *2021 IEEE international intelligent transportation systems conference (ITSC)* (pp. 2111–2117). IEEE.

Trasberg, T., Soundararaj, B., & Cheshire, J. (2021). Using wi-fi probe requests from mobile phones to quantify the impact of pedestrian flows on retail turnover. *Computers, Environment and Urban Systems, 87*, Article 101601.

Traunmueller, M. W., Johnson, N., Malik, A., & Kontokosta, C. E. (2018). Digital footprints: Using wifi probe and locational data to analyze human mobility trajectories in cities. *Computers, Environment and Urban Systems, 72*, 4–12.

Voss, W. G. (2016). European union data privacy law reform: General data protection regulation, privacy shield, and the right to delisting. *The Business Lawyer, 72*, 221–234.

Wang, M., Zhou, J., Long, Y., & Chen, F. (2016). Outside the ivory tower: Visualizing university students' top transit-trip destinations and popular corridors. *Regional Studies, Regional Science, 3*, 202–206.

Wang, Y., Xie, L., Zheng, B., & Lee, K. C. (2014). High utility k-anonymization for social network publishing. *Knowledge and Information Systems, 41*, 697–725.

Wirz, M., Franke, T., Roggen, D., Mitleton-Kelly, E., Lukowicz, P., & Tröster, G. (2012). Inferring crowd conditions from pedestrians' location traces for real-time crowd monitoring during city-scale mass gatherings. In *2012 IEEE 21st international workshop on enabling technologies: Infrastructure for collaborative enterprises* (pp. 367–372). IEEE.

Zhou, J., Wang, M., & Long, Y. (2017). Big data for intrametropolitan human movement studies a case study of bus commuters based on smart card data. *International Review for Spatial Planning and Sustainable Development, 5*, 100–115.