RESEARCH REPORT

Institute of Transportation Studies | UNIVERSITY OF CALIFORNIA

Sanitization of Transportation Data: Policy Implications and Gaps

Matt Bishop, Ph.D., Professor, Department of Computer Science, University of California, Davis

November 2021

ITS UCDAVIS INSTITUTE OF TRANSPORTATION STUDIES

Report No.: UC-ITS-2020-04 | DOI: 10.7922/G2NS0S6B

Technical Report Documentation Page

1. Report No. UC-ITS-2020-04	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Sanitization of Transportation Data: Policy Implications and Gaps		**5. Report Date** November 2021 **6. Performing Organization Code** ITS-Davis
7. Author(s) Matt Bishop, Ph.D., https://orcid.org/0000-0002-7301-7060		**8. Performing Organization Report No.** UCD-ITS-RR-21-62
9. Performing Organization Name and Address Institute of Transportation Studies, Davis 1605 Tilia Street Davis, Ca 95616		**10. Work Unit No.** N/A **11. Contract or Grant No.** UC-ITS-2020-04
12. Sponsoring Agency Name and Address The University of California Institute of Transportation Studies www.ucits.org		**13. Type of Report and Period Covered** Final Report (October 2019 – September 2020) **14. Sponsoring Agency Code** UC ITS
15. Supplementary Notes DOI:10.7922/G2NS0S6B		

16. Abstract

Data about mobility provides information to improve city planning, identify traffic patterns, detect traffic jams, and route vehicles around them. This data often contains proprietary and personal information that companies and individuals do not wish others to know, for competitive and personal reasons. This sets up a paradox: the data needs to be analyzed, but it cannot be without revealing information that must be kept secret. A solution is to *sanitize* the data—i.e., remove or suppress the sensitive information. The goal of sanitization is to protect sensitive information while enabling analyses of the data that will produce the same results as analyses of the unsanitized data. However, protecting information requires that sanitized data cannot be linked to data from other sources in a manner that leads to desanitization. This project reviews typical strategies used to sanitize datasets, the research on how some of these strategies are unsuccessful, and the questions that must be addressed to better understand the risks of desanitization.

17. Key Words Data, traffic data, data sharing, data cleaning, data fusion, data privacy, computer security, transportation planning		**18. Distribution Statement** No restrictions.	
19. Security Classification (of this report) Unclassified	**20. Security Classification (of this page)** Unclassified	**21. No. of Pages** 33	**21. Price** N/A

Form Dot F 1700.7 (8-72) Reproduction of completed page authorized

About the UC Institute of Transportation Studies

The University of California Institute of Transportation Studies (UC ITS) is a network of faculty, research and administrative staff, and students dedicated to advancing the state of the art in transportation engineering, planning, and policy for the people of California. Established by the Legislature in 1947, ITS has branches at UC Berkeley, UC Davis, UC Irvine, and UCLA.

Acknowledgments

This study was made possible through funding received by the University of California Institute of Transportation Studies from the State of California through the Public Transportation Account and the Road Repair and Accountability Act of 2017 (Senate Bill 1). The authors would like to thank the State of California for its support of university-based research, and especially for the funding received for this project. The author would also like to thank Prof. Karl Levitt and Dr. Michael Clifford for useful conversations.

Disclaimer

Sanitization of Transportation Data: Policy Implications and Gaps

Matt Bishop, Ph.D., Professor, Department of Computer Science, University of California, Davis

October 2021

ITS UCDAVIS
INSTITUTE OF TRANSPORTATION STUDIES

Report No.: UC-ITS-2020-04 | DOI: 10.7922/G2NS0S6B

Table of Contents

Table of Contents

Executive Summary

Executive Summary

Transportation data provides information vital to city planners, transportation companies, and governments. The data collected by transportation companies—such as ride-hailing companies—typically consist of information about the passenger, when and where the trip started and ended, route taken, the number of miles driven, etc. City planners can use this information to look for heavily-trafficked routes, to predict future transportation needs, and to build infrastructure for them. Companies use this information to improve their services, for example decreasing waiting-times for passengers and determining the best routes, how much to charge passengers, and where to position vehicles for maximizing customers and profits. Governments use the data in collaboration with city planners and to potentially develop methods of charging vehicles, such as road-use charges or tolls, to pay for infrastructure or congestion control.

A major challenge and inherent conflict when sharing data is between (a) including all the information needed for a desired analysis and (b) excluding information that could be used by adversaries to violate the privacy of passengers or to harm the business interests of the company that collected the data. For example, an adversary may use personally identifiable information to track the movement of a person or use proprietary information on common pick-up locations to take business from a ride-hailing company. The process of removing sensitive information from a dataset is referred to as *sanitization*, or, when personal identifiable information is removed, *anonymization*. However, for sanitization to be successful, not only must the sensitive information be suppressed or removed from the dataset, so must ancillary information that would allow an adversary to deduce the sensitive information.

If the data is to be shared with a specific, limited group of people, the inherent conflict between protecting sensitive information and maintaining the usefulness of the data is achieved through legal means, such as contracts that restrict the uses of the data.

If the data is to be made available to the public, then the focus shifts to protecting the suppressed information. There are two mutually exclusive assumptions involved in the approach to suppressing information to sanitize the dataset. The first is that the only data in the dataset is available to recover the suppressed information. The second is that an adversary can access other sources of data and link that data with the unsuppressed information in the dataset of interest to reveal the suppressed information, thereby desanitizing that dataset. Therefore, the sanitization must take into account the external data.

This report reviews some of the typical strategies used to sanitize datasets and the research on how some such strategies are unsuccessful, leaving datasets vulnerable to desanitization. The following are the current gaps—or questions that must be answered to better understand the risks of sanitized data being desanitized:

1. What are the specific threats that data sanitization is to guard against, and how can one validate that those are the correct threats?

2. How do we prevent false inferences from being drawn? (For example, when an adversary draws a false conclusion about where a person regularly goes.)
3. How do we determine what external data is needed to desanitize the target data?
4. What information must one or more external datasets have so that an adversary could desanitize enough data in the sanitized dataset for their purpose?

Contents

Introduction

As society increases in complexity, so do the transportation needs of its inhabitants. To meet these needs, planners, transportation agencies, companies, and government agencies need to know how people move about. The ability to gather data on such movements is key to planning.

Computers and sensors aid in this data collection. They can record a multiplicity of attributes of transportation and, using techniques to analyze "big data," can deduce information useful to planning. Call these systems "data gathering systems." Different data gathering systems record different information, according to the needs of their customers. The information gathered usually includes the following:

- Information about who is traveling;
- Time of departure;
- Time of arrival;
- When the ride was requested;
- When the ride started;
- When the ride ended;
- Origin (starting point of travel): usually nearest intersection in latitude and longitude (GPS);
- Destination (end point of travel): usually nearest intersection in latitude and longitude (GPS);
- Miles driven;
- Waiting time (how long between the request and the beginning of the ride);
- Trace data (trace of route); and
- Number of passengers.

Call one set of this information a *record*, and each element of the record a *field* or *attribute*. A record records information about one or more segments of travel. A field records one datum about the segment(s).

The type of ride and provider affects what can be collected. For a ride-hailing service such as Uber or Lyft, all the above information can be gathered. For public transit such as buses, the information recorded will be about routes, times, and numbers of passengers rather than individuals.

Private vehicles are another matter. Currently no such information is collected. There has been discussion of taxing vehicles by trip, because with the advent of electric cars, the gasoline tax revenue will decrease, as those cars do not use gas and so the owners do not pay the gas tax. To determine the amount of tax, legislation will determine the data to be gathered. At a minimum, it will consist of distance and some attribute tying that to the vehicle. It may also include other information such as origin and destination.

The granularity of the data affects its utility. If the data only includes the above information, the data is unlikely to provide information about the route, unless there is only one route from the origin to the destination; then

one can deduce the route from the above information. So route data can be gathered and used to suggest to drivers routes with the least traffic [1].

This data serves several purposes, the majority of which relate to urban planning:

- Predicting transportation needs [2,3];
- City planning [4];
- Identifying traffic patterns [5,6,7];
- Identifying current traffic to be routed around traffic jams and other hindrances [1];
- Determining how far a car travels for taxation purposes [8,9]; and
- Determining for how long passengers are waiting for a ride between two points [10].

Sanitizing data refers to suppressing values of fields that the sharer of the data does not wish others to know. This may be proprietary information or personal information. Suppressing information about people such as names and addresses is also called *anonymization*. Sanitizing data complicates sharing the information. The main problem is that sanitized data may be combined with external information to reveal *sensitive information*—i.e., information that needs to be kept confidential. As examples from other domains, in 2006, AOL released anonymized search queries, and two days later removed this from the web. Two New York Times reporters were able to identify one person using information obtained from maps, home ownership records, and other external data [11]. To confirm their identification, they interviewed her, and she confirmed the queries were hers. Netflix released a set of records with customer names anonymized and offered a $1,000,000 prize to anyone who could build a movie recommendation system that worked better than Netflix's. Researchers noticed the fields present in the release were similar to fields in the Internet Movie Database (IMDb). Working with data of around 50 customers, they were able to de-anonymize two users in the Netflix data [12]. While these are not transportation data, they show the difficulty of anonymizing data irrecoverably.

The purpose of this report is to examine the problem of sanitizing transportation data and to identify gaps in existing work in this area. Specifically, what are the gaps that need to be filled to answer the following questions:

1. How does one determine what external data is necessary to reverse the sanitization?
2. How does one reduce the probability of this given a set of external data?

The paper is organized as follows. First, we present the system and threat models to make clear the basis for the work done in the rest of the paper. We then examine what privacy is, in the context of transportation data and keeping information hidden. We review some related work, and then present an approach that combines work done in statistics with sanitization. We conclude with a list of gaps between solving the problem posed above and the current state of research.

System and Threat Models

The benefits of data sharing go beyond transportation data. The National Institutes of Health says, “We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health” [13]. To provide guidance on data sharing, the European Commission wrote, “Data-driven innovation is a key enabler of growth and jobs in Europe. The importance of data collected online, the growing importance of data generated by objects connected to the Internet of Things (IoT), the increasing availability of Big Data analytics tools and the emergence of broad availability of certain Artificial Intelligence applications are key technical drivers” [14]. More generally, Karl Popper, the influential philosopher of science, presented data as necessary for reproducing experimental results, a foundation of science: “non-reproducible single occurrences are of no significance to science” [15, p. 66].

However, data to be shared often contains information that needs to be kept confidential. Such information will be called “sensitive.” The sensitive data may need to be kept confidential from everyone, or only from some set of entities (people and organizations). For example, if one mobile service does lots of business ferrying people from one locale to another, a competitor might position vehicles and advertise heavily in that area to obtain more traffic for itself and, consequently, less for the first carrier. So the first carrier might consider its traffic patterns to be sensitive data with respect to the second carrier, but not to urban planners.

Clearly, the original data cannot be circulated as is. Two approaches are used to share the data.

In the first approach, the holder of the data can share it with a limited number of entities and bind them, usually contractually, not to reveal the data and, in many cases, constrain what they can do with the data. In this way, the holder of the data controls what the recipients can do with it. A good analogy is digital rights management. The originator of the data controls all rights to the data, and can constrain its distribution and use, just as a movie studio distributes movies on Blu-Ray or DVD, with the requirement that the movies not be copied for another and are for home use only. (The formal access control model is called “ORCON”, for ORiginator-CONtrolled access control [16].)

The second approach is to publish the data, making it freely available to everyone. Holders that use this method have no control over who accesses the data or what they do with it. There is no relationship, contractual or otherwise, between the holder and the recipient. This poses problems in that the original holders of the data cannot control what others do with it—specifically, whether they try to deanonymize it. As an example, the “information about who is traveling'' is recorded as unique information and anonymized by hiding the person's name—for example, by replacing the name with a number that is random or computed by a mathematical function of the name. But this is insufficient because by knowing the starting and ending points, one can use external data to identify the addresses, and from that deduce information about the individual, in some cases even a name. Hence sharing the data openly will compromise the identity of the individual. A contractual agreement on the use of the data would give the data provider some recourse.

This report assumes the data is to be published as in the second approach, and hence anyone can see it. Even if the first method is followed, a serious problem lies in how the data is handled. If the recipient stores the data and it is compromised, or the data is compromised as the recipient moves it or computes with it, the sensitive information has leaked. Further, it is vulnerable to an insider attack—an attack where a trusted person betrays that trust, in this case leaking the data. Thus, it is appropriate to assume the data is to be published.

For clarity, we define some terms as used in this paper:

- *raw data*: the original dataset; for example, transportation data consisting of entries for some number of rides;
- *record*: one datum in the dataset, which may itself contain several parts; for example, in transportation data, the record of one trip;
- *fields* or *attributes*: part of a record; for example, in transportation data, the start time, end time, and waiting time for a particular ride;
- *sanitization*: the process of transforming a dataset to conceal information deemed sensitive or private; the various forms of this word are defined similarly;
- *anonymization*: sanitization of data confined to information about people, but used in many works as a synonym for "sanitization";
- *redacted data*: data that is deleted or transformed to conceal the original value;
- *external data*: data that is not part of the original dataset, for example obtained from the web;
- *desanitiation*: the process of reversing the sanitization of a dataset, to reveal one or more redacted values;
- *deanonymization*: desanitization applied to redacted information about people, but again often used as a synonym for "desanitization";
- *sanitizers*: someone who sanitizes the data;
- *anonymizers*: someone who anonymizes the data;
- *adversary*: someone who is not authorized to see the raw data, and wants to desanitize the data so they can see (parts of) the raw data; and
- *analyst*: someone authorized to analyze either the raw or sanitized data (made clear from context).

We also make several assumptions:

- The data is composed of records, each with one or more fields containing sensitive information;
- The sensitive fields are redacted by transforming them into some other value or deleting them;
- The raw data is kept secret;
- The redacted data is released without constraint, so anyone can see it and use it; and
- An adversary wants to recover some redacted values from the redacted fields.

System Model

The system model describes how the data is handled. It covers the many different methods used in practice. Figure 1 shows the model of data sanitization.

When raw data is collected (lower left oval in Figure 1), the analysts will analyze it and obtain some results (upper left oval). The sanitizers will hide some information in the data, producing the sanitized data (lower right oval). Ideally, when the analysts analyze the sanitized data (upper right oval), the results obtained will be the same as the results from analyzing the raw data. The adversaries do not have access to the raw data but do have access to the sanitized data. They want to recover the raw data. The sanitizers want to prevent this, and so have instituted a privacy policy that says what data may be revealed and what data must not be revealed. The sanitization is designed to enforce this policy.

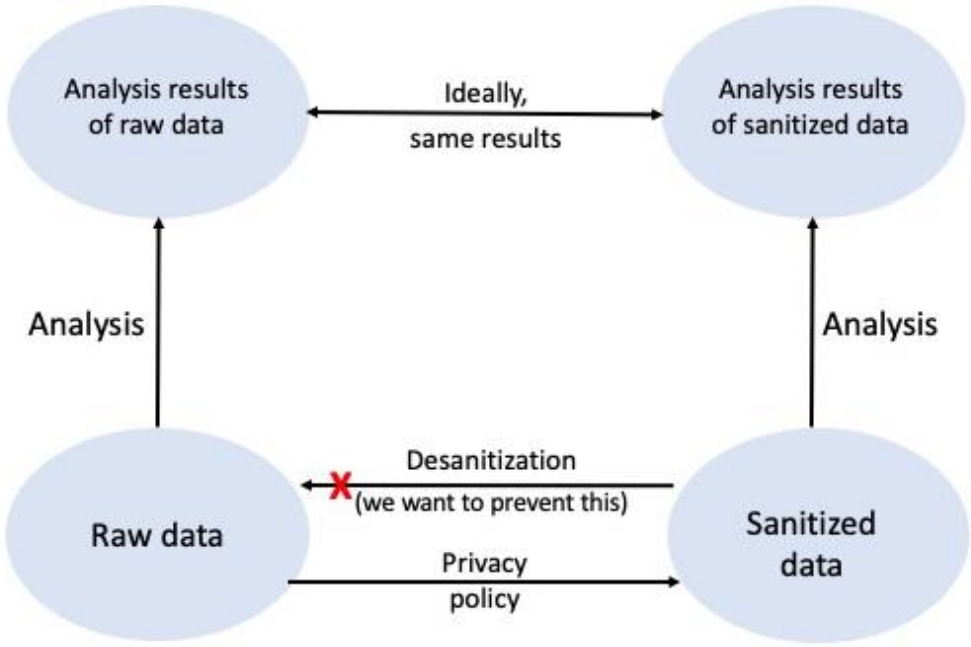

Figure 1. The system model

As an example, suppose analysts want to determine the most common routes used by vehicles and when they are taken. Records available to them contain information such as that listed in paragraph 2 of the Introduction: who is traveling; time and location of request, origin, destination; route information; and so forth. The privacy policy is that segments of routes cannot be tied to the path an individual vehicle takes. We view two analyses.

The first analysis uses raw data, and from that establishes the flows of traffic over each block, the times at which vehicles enter and exit the block, from which road they enter the block, and to which road they leave the block.

This information lies in the trace data, which includes the route and times of accessing each segment of the route. This produces the results of the analysis and corresponds to the left ovals in Figure 1.

The second analysis uses sanitized data. To satisfy the privacy policy that states what is to be kept confidential (and is represented by the bottom arrow in Figure 1), the sanitizers break each trace field into a sequence of data, each datum representing a block, and delete everything else from the record. They give this to a second set of analysts, who carry out the analysis on the redacted data. As they have sufficient information to perform the same analysis as the analysts who used the raw data, the results should be the same. This is represented by the right ovals in Figure 1, and the two-headed top arrow shows the (desired) equivalence of the results.

But an adversary, call her Addie, gets hold of the sanitized dataset. She wants to find the starting point and ending point of each trip, so she can then look up who lives nearby and construct a pattern of people's movements. Although the block information is scrambled, that information includes the time of entry onto and exit from the block. Addie gets a map of the city and uses that to match the entry time of one block with the exit time of the adjacent blocks. She is able to construct the routes taken. This is the “desanitize” arrow, and the “X” near the end symbolizes the data holder's desire that no-one be able to recover any sanitized raw data from the sanitized data (which Addie just did). Although Addie has not recovered all information from the records, she has uncovered enough information to suit her needs.

Threat Model

The threat model states what attacks sanitizing the data is to prevent. A breach occurs when one of these threats is realized. The threat is that the adversary uncovers sanitized information *using the dataset*, possibly in combination with other external data. That is, the adversary does not have access to any copy of the raw dataset.

While the adversary may not be able to recover the sanitized information, they may be able to draw inferences about that data. There is a dual problem here. First, if the inferences are correct, then the adversary has learned something about the sanitized data, and that may be sufficient for their purposes. As an example, suppose the adversary is trying to determine how much someone is spending on rides from transport data in which the name of the rider is redacted. The adversary knows that the starting point may be one of 4 places, all within 2 blocks of one another. They look for routes that begin at those four locations and end at the company. The amounts that these routes cost differ but are within $5 of each other. Now the adversary knows approximately what the target spends on rides to work.

An alternate problem arises if the inferences are wrong. Suppose a number of routes begin at one point and end at a hotel. The adversary may infer they belong to the same person, and the target is having an affair, because they go to the hotel almost every day. But an equally logical explanation is that the location is where an intercity bus stops, and people arriving from outside the city spend the night in the hotel during the time in question. The routes are indeed the same, but the riders are different—and the inference is wrong and, if disclosed, could

cause unnecessary grief for the specific person the adversary is investigating. The current research does not examine methods for preventing such erroneous inferences.

The specific threat this study assumes is the ability to uncover any redacted data. In practice, this is usually the identity of the person who is traveling. But other fields may be deemed sensitive under certain conditions. To continue the above example, an extortionist might want to know the name of a hotel with the most traffic (arrivals) to determine where to set up their camera to take pictures to blackmail people. In that case, the destination needs to be redacted along with the traveler's identity.

Privacy

There are many definitions of *privacy*, and people consider different things private. In the context of this work, privacy is the prevention of disclosure of information that has been sanitized or anonymized. With a record of a trip, there are at least 2 entities that may desire privacy: the passenger and the transport company. But their interests may differ. The passenger may not want anyone to know that they used the mobile service. The transport company may want to protect the passenger's identity, and not necessarily the fact that the passenger used the mobile service.

In this context, *disclosure* means revealing information that would otherwise not be known. Two definitions make this more precise. Although both deal with statistical databases, any data can be treated as elements of such a database.

In 1977, Dalenius [17] defined a statistical disclosure as occurring when releasing data, even if sanitized, enables someone to determine a redacted value more accurately than they could without the data being released.

More precisely, such a disclosure takes place when the release of a set of statistics *S* makes it possible to determine the actual value of a redacted value of an attribute more accurately than when *S* is not released. This is a good definition, but it is an ideal.

In 2006, Dwork [18,19] presented an alternate definition of disclosure, called *differential privacy*. An example will show how this works. Consider two databases with identical information *except* that one has the sensitive record and the other does not. If someone asks the same question of both databases, the difference in the answers can be made as small as one likes.

In mathematical terms, represent a database as a set of tuples over some domain D^n. Suppose two such databases $d_1, d_2 \in D^n$ differ only in the value of one attribute of one element. One value is the sanitized value of that attribute for that element; the other is the raw value. A transcript *t* is a sequence of queries and their responses. Select a transcript $t(d_1)$ and its corresponding transcript $t(d_2)$ Then the probability that the transcripts differ can be made arbitrarily small. Formally, the databases are ε-indistinguishable if

$$P(t(d_1) = t) \leq e^{\varepsilon} P(t(d_2) = t)$$

where ε is a parameter chosen by the sanitizers. The transcripts mean the databases are interactive.

Consider a database that is *not* interactive, for example, a set of tables containing the data of interest, and once sanitized as appropriate, is released—this is typically how transportation data is shared. For this non-interactive release, there is always external information that, when combined with the sanitized released information, will enable one to deduce the original values of the information that was sanitized from the released data [19]. The key question is how the adversary can determine *what* information is needed, and *how* they can get it.

Related Work

Transportation data, and more generally mobility data, is becoming widely used, and along with that expansion come concerns about privacy. The ability to track movement and associate that movement with an individual or some small set of individuals enables the development of individual tracking, which in turn can lead to problems if someone does not want others to know where they go.

Routes are often unique, and so transport data can be de-anonymized using little external information. de Montjoye et al. [20] used a dataset of 1.5 million users of a mobile phone company to demonstrate this. Unfortunately, they did not say how the dataset was anonymized.

Some transportation data contains information that allows deductions to be made that will reveal private information—i.e., information about individual lifestyles, medical conditions, and other information generally considered personal and not to be known widely.

Social networking expands the types of non-private information that can be used to deduce private information, as many such networks provide locations of postings, especially where posted pictures were taken. Connecting this with mobility data enhances the individual's characterization, leading to violations of privacy. Ruiz Vicente et al. [21] has a good overview of the problem.

Given a graph that indicates social connections among people and related transportation data, the two can be structurally correlated to identify individuals. A small study tested this with three datasets drawn from social networks. The experimenters used proximity (defined slightly differently for the different datasets) to define social connection. They correlated this with a public social network (Facebook for two sets, and the DBLP authorship database in which co-authorship provided the social relationships for the third). From these, they created contact and social graphs. By comparing structures, they achieved over 80% accuracy in deanonymizing the mobility datasets [22].

Other work has inferred social relationships from mobility data [23,24], which would enable matching. Variants include using location, co-location, and identity in social networks that enable users to publish location data in real time, leading to similar inferences [21].

Preserving Privacy through Anonymization

There are two approaches to providing sanitized transport data: synthesizing data and sanitizing raw data.

Synthesizing data has been proposed as a way to eliminate privacy violations. The idea is to generate from an actual trace a second trace with artificial data in such a way that the utility of the raw data is preserved. Machanavajjhala et al. [25] discuss this approach in detail. Some tools to do this have been studied [26].

This method requires the data holders to know to what use the data will be put. For example, if a city wants to analyze which routes are most often taken, then the statistics of interest are about that route (the number of times the route is taken, the mean time to transit the route, and so forth). Knowing this, the data holder can create artificial data whose statistics match those of the real data. However, other statistics that one might want to obtain may, or may not, be the same for the artificial data and the raw data.

The second method is replacement or deletion of the values in some fields. The usual approach is to apply differential privacy in some form. Many papers discuss this (see for example Xiong et al. [27], Chen et al. [28], and Yin et al. [29]). Researchers have used differential privacy to anonymize a large transportation data set by perturbing the data [28].

Sometimes differential privacy is combined with other mechanisms like *k*-anonymity, in which *k* people are combined into a set that is then protected; the idea is that one can identify the set but not the individual to whom the data applies. For example, Soria-Comas et al. [30] combine *k*-anonymity with differential privacy to focus on not impeding the utility of the data. Other privacy mechanisms have been proposed, such as LKC [31,32], which focuses on anonymizing trajectories (routes). Local differential privacy [33,34,35] perturbs each record in a different way, but ensures the resulting dataset will produce the same statistics as the original dataset.

RAPPOR [33] was developed for cloud providers to collect statistics on their users and software. It is designed for interactive use. It uses a two-step process: first, the data is randomized (permanent randomized response). For each query, the value in the previous step is further randomized (instantaneous randomized response). That final value is reported. Each step uses techniques to ensure differential privacy holds. The communication cost can be high, because each client has to send a vector of values to the server.

Oltenau et al. [36] examine the effects of co-location information of 2 or more individuals on location privacy, where the location is sanitized but the fact of 2 people being together is not.

Work has also been done on matching different mobility datasets. Although not strictly deanonymization, this will become critical to our identification of gaps. Kondor et al. [37] studied this problem using two very large datasets drawn from mobility traces and transportation smart card usage of several million people. They correlated events across both databases to develop a matching algorithm, and with one week of data, they could match 16.8% of the records; with four weeks, over 55%. Shao et al. [38] use a similar method, but they correlated perturbed data internally. Their method, iTracker, is based on machine learning techniques.

Liu et al. [39] deal with dependencies, but only within the same set; they do not consider external data. In a blog post [40], McSherry disputes their results and points out that “differential privacy’s guarantees only mask the presence of those records received from each user, they do not mask larger statistical trends that may reveal information about each user.” Matching with the right external data will provide those trends, and thus reveal information about each user.

Record Linkage and Privacy

Consider two databases, one with records of people before their name change, and one with people after their name change. All other data for each person in both databases is the same. Then *record linkage* refers to correlating the records in the database so those with the person's original name are connected to those with the person's changed name. Newcombe et al. [41] were among the first to propose applying this to vital records.

In practice, the problem is complex. For example, the records may contain variations of a person's name. One may refer to "Matt", the other "Matthew". Values may be misspelled ("Mathew") or entered incorrectly. Thus, something more complex than simple matching is needed. Fellegi and Sunter [42] provide a framework for using computers to link records in the face of such problems. Further research has studied a number of techniques to link records in the face of inaccurate or incomplete information, including machine learning [43].

An *identifier* is some information identifying what is to be sanitized—for example, a person's name or address. A *quasi-identifier* is a set of non-redacted data from which an adversary can reconstruct the sanitized data.

As an example, consider a record that contains a person's name, ZIP code, gender, and birthday. The name would be the identifier because that uniquely identifies the entity whose identity is to be redacted. Now consider the same record but with the name replaced by a 10-digit random number. So the identifier is suppressed. But the person can probably be identified by the ZIP code, gender, and age [44,45]. All three are necessary. So those three elements—ZIP code, birthday, and gender—form a quasi-identifier.

Suppressing identifiers is easy. But determining what fields form a quasi-identifier is much more complicated, because the relationships with the entity may be obscure.

If some fields of the sanitized record match those of the external record, then the redacted values can be compared to the unredacted values in the external database, and the resulting linkage will desanitize those fields.

In addition to (or perhaps instead of) deleting fields, consider perturbing the data, as is customarily done in methods using differential privacy. In that case, rather than seeking an exact match, the adversary must seek a "close enough" match. What "close enough" means must be determined from the type of the field, the possible values of the field, and other factors.

More precisely, consider a database *D*. Each record R is composed of *n* fields, so $R = \{ r_1, \cdots, r_n \}$. When the record is sanitized, some fields may be removed. Without loss of generality, reorder the fields so the last $n - k$ fields are deleted. Thus, the record consists of *k* fields that are visible and $n - k$ fields that are not.

Now consider a second database D'. Each record $R' \in D'$ consists of fields $R' = \{ r_1, \cdots, r_k, x_{k+1}, \cdots, x_m \}$. By linking R' with any *R* such that the first *k* fields match, we have at least one of $r_{k+1}, \cdots, r_n$ that matches one of $x_{k+1}, \cdots, x_m$.

Let f_i be the perturbation function for field r_i. Then, instead of the record *R*, there is a modified record $R^* = \{ f_1(r_1), \cdots, f_k(r_k)\}$. Now the elements of the redacted data set no longer match the elements in other databases. So, for simplicity, we assume that there is a series of δs such that $\left\| f_i(r_i) - x_j \right\| \leq \delta_{ij}$ for $j = k + 1, \cdots, n$, and the δ_{ij} are defined in such a way that perturbations meeting this requirement will not affect the utility significantly. If no such δ_{ij} exist, then the utility of the data is affected. We note that if the data is categorical, the sanitizers must define the difference and the norm in terms of the categorical values. Thus, to desanitize the data, an approximation method is required. In this case, we compute $\delta'_{ij} = \left\| f_i(r_i) - x_j \right\|$ for $j = k + 1, \cdots, n$ and select the x_j with the smallest δ'_{ij}.

Note that each field will have a type. For example, one field may be the time of departure; another, the miles driven. Comparing the values in these two fields does not make sense. So in the above, not every x_j need be compared to every $f_i(r_i)$. The time of departure would need to be compared with the time fields, and the miles driven with numerical fields.

Closed World and Open World Assumptions

Privacy can be viewed as a set of constraints on the records and fields. When these constraints are satisfied, the information cannot be tied to that which is being redacted. The problem comes when one balances privacy with how the data is to be used.

Sanitizing data so it is still useful requires meeting two sets of constraints: the *privacy constraints* that protect the sensitive information and the *utility constraints* that ensure the resulting sanitized data can be analyzed as appropriate. Satisfying the utility constraints is usually simple but doing so while satisfying the privacy constraints requires suppressing both identifiers and quasi-identifiers.

Let $P = \{ p_1, \cdots, p_n \}$ be a set of privacy constraints, and $U = \{ u_1, \cdots, u_m \}$ a set of utility requirements that must be satisfied for the data to be usable. The sanitized data must satisfy $C = p_1 \wedge \cdots \wedge p_n \wedge u_1 \wedge \cdots \wedge u_m$. If this cannot be satisfied, then at least two attributes of the data conflict, and the conflict needs to be resolved. We assume this has been done, and the elements of *P* and *U* are consistent with respect to their elements and the elements of the other set. Satisfying *U* is simple. But satisfying *P* requires suppressing both identifiers and quasi-identifiers.

To do this, it is necessary to look at sources of data, which in turn are controlled by one of two hypotheses.

The first asserts the privacy constraints need only be satisfied by the data in the dataset; this is the *closed world* hypothesis [46]. The privacy constraints can then be tested using that dataset. Given *only* that data, the sanitization can be tuned to maximize the number of utility constraints that are satisfied by the data, or the number of privacy constraints, or some combination of those constraints.

The second hypothesis asserts that the privacy constraints must be satisfied for *any* data, regardless of whether that data resides in the dataset or is external to it; this is the *open world* hypothesis [46]. As transportation data

is not isolated, and its viewers have access to the World Wide Web and other public sources of data, this hypothesis is the more realistic of the two for our work.

Differential privacy is ideal for this, because it takes into account the external data. Since Dwork first proposed differential privacy, numerous methods for achieving it have been proposed. All are based on perturbing the data in some fashion, for example by adding noise using Laplacian mechanisms or exponential mechanisms. The perturbation must be done so as not to affect the utility of the data. When strong privacy guarantees are required for analyzing big data, the loss of accuracy is sufficiently large as to affect the results of analysis.

To fully realize the strength of differential privacy, one must also consider relationships that identify linkages between records. One paper [47] examined the interdependencies of fields with internal and external data. These interdependencies often reveal information about redacted fields regardless of how they are redacted. An example will show how this is done.

The Massachusetts commission responsible for health insurance for state employees, GIC, collected around 100 attributes on patient encounters for state employees. The data was believed to be anonymous, as all identifying information about the patient except for the ZIP code, gender, and birth date was sanitized. So GIC made it available to the public. Sweeney purchased the voter registration lists for Cambridge, MA, which as noted were publicly available. Realizing that the governor of Massachusetts lived in Cambridge, Sweeney compared the voting lists for Cambridge with the redacted medical information. She knew the governor's birth date, which matched six medical records. The governor was male, eliminating three of these. And she knew the ZIP code where the governor lived, which matched only one record [44]. Thus, the external information enabled Sweeney to break the anonymization by correlating the sets of data.

The manner in which redaction was carried out was irrelevant; indeed, in this case, the patient identities were completely suppressed. Yet they were reconstructed by comparing the internal, unredacted information, intended to be used for analysis, with publicly available information. This shows that the perturbation methods, while necessary to achieve ε-indistinguishability, may not be sufficient.

In a closed-world scenario, where the adversary only has the redacted data available, one need only consider that data in analyzing the constraints controlling privacy and utility. But closed-world scenarios are no longer common, because of the information available in the World Wide Web and other sources. Thus, the proper question to ask is:

> *Given the balance of privacy and utility desired, how does one determine what external information is needed to assure the level of privacy desired?*

Or, in more formal terms, what is the probability that an adversary can obtain and use data external to the database that will cause

$$P(t(d_1) = t) > e^{\varepsilon} P(t(d_2) = t)$$

thereby preventing the dataset from satisfying ε-indistinguishability. Dwork demonstrates that there is always such information, as noted above. The question is, can the adversary obtain it?

This fundamentally alters the nature of data sanitization. It is not a matter of whether an adversary can desanitize redacted data. It is a question of whether the adversary can determine the data needed to do the desanitization, and then locate it and use it.

Summary of Related Work

Combining record linkage with quasi-identifiers, the dataset can be mined to determine which combination of fields lead to quasi-identifiers. This may lead to quasi-identifiers that identify a single record, or to "quasi-quasi-identifiers" that identify some small set of records. The data mining will identify the fields that need to be suppressed.

Now factor in external information. Clearly, any external databases with raw data corresponding to the sanitized data can reveal the sanitized data if the records can be linked. Further, any set of external data allowing the linkage can reveal the same information. More precisely, if any information from external sources can be added to the redacted data to produce quasi-identifiers not in the dataset, then the sanitization can be reversed (at least, to some degree).

Thus, the problem is how to identify what external information is needed.

The Gaps

This section discusses topics not addressed or insufficiently addressed (the *gaps*) in current work on sanitizing transportation data. With respect to this data, the focus traditionally has been on identifying people. But an attacker might not care about the people. For example, the attacker may want to know which of several businesses in town rely on car services such as Lyft or Uber. So, the first gap lies in what is to be protected:

> **Gap 1**. What is the threat? That is, how can an adversary use the released data to discover information that is false or that is true but the sanitizers *or the people to whom the record applies* do not want the adversary to infer?

One approach is to ask what an adversary could learn from the sanitized data rather than ask whether specific information could be recovered. The sanitizers, or some organization like an Institutional Review Board, must then decide whether the gains from releasing the information offsets any potential harm, considering who might be harmed.

An important part of this is that the inferences the adversary draws may be true or false. The research done so far focuses on preventing an adversary from drawing *true* inferences. But false inferences can also be drawn. The AOL data release provides a good example of this. The reporters who tracked down one of the anonymized names write [11]:

> At first glace [*sic*], it might appear that Ms. Arnold fears she is suffering from a wide range of ailments. Her search history includes “hand tremors,” “nicotine effects on the body,” “dry mouth” and “bipolar.” But in an interview, Ms. Arnold said she routinely researched medical conditions for her friends to assuage their anxieties. Explaining her queries about nicotine, for example, she said: “I have a friend who needs to quit smoking and I want to help her do it.”

False inferences can cause equal or greater harm than true inferences. Suppose the raw data shows someone making repeated trips to a particular place. One way to sanitize the data would be to generalize it; that is, show the person making repeated trips to an area rather than one place. If that area encompasses motels, one inference might be that the person is having an affair—but the raw data shows the “particular place” is the library, which leads to the very different conclusion: the person loves to read. This is the second gap:

> **Gap 2**. How does one determine the incorrect inferences that could be drawn from a set of sanitized data, and how does one prevent that without affecting (or minimally affecting) the utility of the data set?

The next is the question raised by the open-world scenario under consideration. The current research focuses on using differential privacy to provide a guarantee of some degree of privacy; this is done by perturbing the data. But that is insufficient, because of linkage. That changes the problem of desanitization into one of risk: what is

the likelihood of an adversary finding *external* data enabling them to desanitize the information? More simply, what specific external data is needed for the adversary to be successful:

> **Gap 3**. Given a set of sanitized data, how do we determine what data is necessary to desanitize it? The problem here is record linkage, where the data being linked is not explicitly in the records. The external data may form quasi-identifiers that can then be linked with the raw or perturbed data in the sanitized dataset.

The gap here speaks to the data being available in the same form as the raw data. In that case, linking the records is straightforward. If the data does not match exactly but is close enough ("close enough" being determined based on the data, its origin, and the need for precision), that may be sufficient.

The fourth gap generalizes this question, asking about records that do not match the sanitized data but, when combined with that data, enable reidentification. The key here is the creation of quasi-identifiers with attributes from both the sanitized and the external data.

> **Gap 4**. How do we identify the necessary attribute in order to make quasi-identifiers?

This is somewhat different than the question of finding quasi-identifiers in a dataset, which is a closed-world problem that generally uses some form of machine learning, the idea being that different combinations of the fields would be aggregated and the results used to try to identify someone [48,49,50]. The question here is about open-world analysis; what needs to be added to make fields that are not part of quasi-identifiers become quasi-identifiers, when taken together?

Summary and Conclusion

It is widely known that publishing transportation data creates a risk to both the originator of the data and the subjects of that data. The risk consists of an adversary being able to uncover or infer redacted data. Using differential privacy, researchers have focused on the transforms and perturbations necessary to provide a given degree of privacy.

That is not sufficient. Record linkage provides an alternate attack path and must be considered when using any privacy-enhancing mechanism. The linkage also requires considering *external* information, and therein lies the rub.

The problem is that the external information may be unknown to the sanitizers. It may also not exist at the time the data is released and may be created or made available publicly at some time *after* the data is released. This problem has occurred in the past, with sanitized data operating under a closed world model. Someone releases one version of the dataset, properly sanitized. Then, later, a second version of the same dataset, properly sanitized in a different way, is released. Separately, they cannot be desanitized. Put together, they can be.

The amount, type, and nature of what is available to the public via the web and other sources lead to the question of whether *anything* can be sanitized in such a way that it can never be desanitized and yet still provide the information one needs to do analytics. This is an open question. A related question is how to sanitize the dataset so that it is still useful yet would take the adversary the maximum amount of time to desanitize. This is similar to a question in cryptography. Ciphers need not be unbreakable; they only need to be unbreakable until the need for secrecy is gone. Perhaps a similar approach can be developed.

This paper has identified four gaps in the current work on sanitizing transportation data:

1. What are the specific threats data sanitization is to guard against, and how can one validate that those are the correct threats?
2. How do we prevent false inferences from being drawn?
3. How do we determine what external data is needed to desanitize sanitized data?
4. What fields or attributes must one or more external datasets have in order to form enough quasi-identifiers in the set of sanitized data for the adversary to be able to desanitize enough data for their purpose?

A final comment is in order. Even if complete and permanent sanitization is not possible, we live in a world with imperfections everywhere. Computers are the obvious example. We store highly sensitive data on computers, and yet computers have vulnerabilities. We remediate them as best we can, but attackers still compromise data that is highly sensitive. But it is more useful to store the data on computers where it can be analyzed rather than not doing so. Similarly, someone or some group must decide whether the utility of releasing the data outweighs

the damage done should the data be desanitized—and if so, how to minimize that threat of desanitization in light of the use to which the data is to be put. Who those people or groups are is a question for the body politic.

References

[1] Waze Mobile. Traffic Sucks: Know What's Ahead with Real-Time Help from Other Drivers. Waze, Mountain View, CA. https://www.waze.com/waze. Accessed April 17, 2021.

[2] Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., and Damas, L. Predicting Taxi–Passenger Demand Using Streaming Data. IEEE Transactions on Intelligent Transportation Systems, 2013. 14(3): 1393-1402.

[3] Pan, B., Demiryurek, U., and Shahabi, C. Utilizing Real-World Transportation Data for Accurate Traffic Prediction. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, 2012. 595-604.

[4] Zhong, C., Huang, X., Arisona, S. M., Schmitt, G. and Batty, M. Inferring Building Functions from a Probabilistic Model Using Public Transportation Data. Computers, Environment and Urban Systems, 2014. 48:124-137.

[5] Liua, Y., Wang, F., Y., Xiaoa, and Gaoa, S.. Urban land uses and traffic 'source-sink areas': Evidence from GPS-Enabled Taxi Data in Shanghai. Landscape and Urban Planning, 2012. 106(1):73-87.

[6] Zhu, D., Wang, N., Wu, L. and Liu, Y.. Street as a Big Geo-Data Assembly and Analysis Unit in Urban Studies: A Case Study Using Beijing Taxi Data. Applied Geography, 2017. 86:152-164.

[7] Kitamura, R., Fujii, S., and Pas, E. Time-Use Data, Analysis And Modeling: Toward The Next Generation Of Transportation Planning Methodologies. Transport Policy, 1997. 4(4):225-235.

[8] Boesen, U. Who Will Pay For The Roads? Fiscal Fact Report, 725 Tax Foundation, Washington DC, USA, 2020.

[9] California Road Charge. CalTrans, Sacramento, CA. http://caroadcharge.com. Accessed Jan. 14, 2021.

[10] Silalahi, B., Shilvia L., Handayani, P., and Munajat, Q. Service Quality Analysis for Online Transportation Services: Case Study of GO-JEK. In Proceedings of the Fourth Information Systems International Conference, 2017. 487-495.

[11] Barbaro, M. and Zeller Jr., T. A Face Is Exposed For AOL Searcher No. 4417749. The New York Times, 2006.

[12] Narayanan, A. and Shmatikov, V. Robust De-Anonymization of Large Sparse Datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy, 2008. 111-125.

[13] National Institutes for Health. Final NIH Statement On Sharing Research Data. National Institutes for Health, Bethesda, MD. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html. Accessed on Feb 7, 2020.

[14] Data Policy and Innovation Unit of the European Commission. Guidance On Private Sector Data Sharing. European Commission, Brussels, Belgium. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018SC0125&rid=2 . Accessed on Feb. 7, 2020.

[15] Popper, K. The Logic Of Scientific Discovery. Routledge Classics, New York, NY. 2002.

[16] Bishop, M. Computer Security: Art And Science. Addison-Wesley Professional, Reading, MA. 2019.

[17] Dalenius, T. Towards A Methodology For Statistical Disclosure Control. Statistik Tidskrift, 1977.15:429-444.

[18] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In Proceedings of the Third Theory of Cryptography Conference, volume 3876 of Lecture Notes in Computer Science, 2006. 265-284.

[19] Dwork, C. Differential Privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, volume 4052 of Lecture Notes in Computer Science, 2006. 1–12.

[20] de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. Unique In The Crowd: The Privacy Bounds Of Human Mobility. Scientific Reports, 2013. 3:1376:1-5.

[21] Ruiz Vicente, C. , Freni, D. , Bettini, C. , and Jensen, C. S. Location-Related Privacy In Geo-Social Networks. IEEE Internet Computing, 2011. 15(3):20-27.

[22] Srivatsa M. and Hicks, M.. Deanonymizing Mobility Traces: Using Social Network as a Side- Channel. In Proceedings of the 2012 ACM Conference on Computer and Communications Security, 2012. 628-637.

[23] Li, J., Zeng, F., Xiao, Z., Jiang, H., Zheng, Z., Liu, W., and Ren, J. Drive2Friends: Inferring Social Relationships From Individual Vehicle Mobility Data. IEEE Internet Of Things Journal, 2020. 7(6):5116-5127.

[24] Li, J., Zeng, F., Xiao, Z., Zheng, Z. Jiang, H., and Li, Z. Social Relationship Inference Over Private Vehicle Mobility Data. IEEE Transactions On Vehicular Technology, 2021.

[25] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhubber, L.. Privacy: Theory Meets Practice on the Map. In Proceedings of the IEEE 24th International Conference on Data Engineering, 2008. 277-286.

[26] Gursoy, M. E., Liu, L., Truex, S., Yu, L., and Wei, W. Utility-Aware Synthesis of Differentially Private and Attack-Resilient Location Traces. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018. 196-211.

[27] Xiong, P., Zhu, T., Niu, W., and Li, G. A Differentially Private Algorithm For Location Data Release. Knowledge And Information Systems, 2016. 47:647-669.

[28] Chen, R., Fung, B. C. M., Desai, B. P., and Sossou, N. M. Differentially Private Transit Data Publication: A Case Study on the Montreal Transportation System. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012. 213-221.

[29] Yin, C., Xi, J., Sun, R., and Wang, J. Location Privacy Protection Based On Differential Privacy Strategy For Big Data In Industrial Internet Of Things. IEEE Transactions On Industrial Informatics, 2018. 14(8):3628-3636.

[30] Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., and Martínez, S. Enhancing Data Utility In Differential Privacy Via Microaggregation-Based *k*-Anonymity. The VLDB Journal, 2014.23:771--794.

[31] Mohammed, N., Fung, B. C. M., and Debbabi, M. Walking in the Crowd: Anonymizing Trajectory Data for Pattern Analysis. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009. 1441-1444.

[32] Harnsamut, N. and Natwichal, J. Privacy Preservation for Trajectory Data Publishing and Heuristic Approach. In Proceedings of the 21st International Conference on Networked-Based Information Systems, 2017. 7:787-797.

[33] Erlingsson, Ú., Pihur. V., and Korolova, A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014. 1054-1067.

[34] Nguyen, T., Xiao, X., Yang, Y., Hui, S. C., Shin, H., and Shin, J. Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy. arXiv:1606.05053 [cs.DB], 2016.

[35] Zhao, P., Zhang, G., Wan, S., Liu, G., and Umer T. A Survey of Local Differential Privacy for Securing Internet of Vehicles. The Journal of Supercomputing, 2020. 76:8391-8412.

[36] Olteanu, A.-M., Huguenin, K., Shokri, R., Humbert, M., and Hubbaux, J.-P. Quantifying Interdependent Privacy Risks with Location Data. IEEE Transactions on Mobile Computing, 2016. 16(3):829-842.

[37] Kondor, D., Hashemian, B., de Montjoye, Y.-A., and Ratti, C. Towards Matching User Mobility Traces in Large-Scale Datasets. IEEE Transactions on Big Data, 2020. 6(4):714-726.

[38] Shao, M., Li, J., Yan, Q., Chen, F., Huang, H., and Chen, X. Structured Sparsity Model Based Trajectory Tracking Using Private Location Data Release. IEEE Transactions on Dependable and Secure Computing, 2020.

[39] Liu, C., Chakraborty, S., and Mittal, P. Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples. Proceedings of the 2016 Network and Distributed System Security Symposium, 2016.

[40] McSherry, F. Differential Privacy and Correlated Data. https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md. Accessed on January 21, 2021.

[41] Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. Automatic Linkage of Vital Records. Science, 1959. 130(3381):954-959.

[42] Fellegi, I. P. and Sunter, A. B. A Theory for Record Linkage. Journal of the American Statistical Association, 1969. 64(328)L1183-1210.

[43] Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, 2007. 19(1):1-16.

[44] Sweeney, L. *k*-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(5):557-570.

[45] Golle, P. Revisiting the Uniqueness of Simple Demographics in the US Population. In Proceedings of the 5th ACM Workshop on Privacy in Electronic Society, 2006. 77-80.

[46] Crawford, R., Bishop, M., Bhumiratana, B., Clark, L., and Levitt, K. Sanitization Models and Their Limitations. In Proceedings of the 2006 New Security Paradigms Workshop, 2006. 41-56.

[47] Bishop, M., Cummins, J., Peisert, S., Singh, A., Bhumiratana, B., Agarwal, D., Frincke, D., and Hogarth, M. Relationships and Data Sanitization: A Study in Scarlet. In Proceedings of the 2010 New Security Paradigms Workshop, 2010. 151-164.

[48] Soh, C. W., Njilla, L. L., Kwiat, K. K., and Kamhoua, C. A. Learning Quasi-Identifiers for Privacy-Preserving Exchanges: A Rough Set Theory Approach. Granular Computing, 2020. 5:71–80.

[49] Motwani, R. and Xu, Y. Efficient Algorithms for Masking and Finding Quasi-Identifiers. In Proceedings of the 2007 Very Large Data Base Conference, 2007.

[50] Pastore, M., Pellegrino, M. A., and Scarano, V. Detecting and Generalizing Quasi-Identifiers by Affecting *Singletons*. In Proceedings of Ongoing Research, Practitioners, Workshops, Posters, and Projects of the International Conference EGOV-CeDEM-ePart, 2020. 327–335.