Legally Anonymizing Location Data Under the GDPR

Cameron D. Bale*, **Jordan L. Fischer****, **Matthew J. Schneider***, **Steven Weber*****, **Suzanne Chang****

*Lebow College of Business, Drexel University, Philadelphia, PA 19104, USA.

**Thomas R. Kline School of Law, Drexel University, Philadelphia, PA 19104, USA.

***Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA.

E-mail: cdb327@drexel.edu, jlf324@drexel.edu, mjs624@drexel.edu, spw26@drexel.edu, sc3887@drexel.edu

Updated March 21, 2023.

Abstract. In the last decade, regions and countries adopted distinct data protection legislation to govern the collection and processing of personal data. Most of these legislative frameworks recognize that data can be personal and non-personal; but, there is a lack of definitive criteria for when personal data has become non-personal data, as well as an understanding of the practical consequences of applying such criteria. This creates uncertainty as to whether organizations are truly compliant with privacy law. In this paper, we propose a framework of anonymization criteria to evaluate the conversion of personal data to non-personal data under the GDPR, which can be used to assess whether data falls outside of the scope of privacy legislation. We test our framework on two logical statistical solutions applied to the location data of COVID-19 patients in South Korea. We show that these anonymization solutions are capable of producing legally anonymous location data, but with varied degrees of usefulness. Further, we show the importance of accounting for context when balancing privacy and utility in legally anonymized location data.

Keywords. Privacy Law, Anonymization, Statistical Analysis, Location Data

1 Introduction

The current privacy legislative landscape is complex and disjointed. In the past decade, a legal evolution placing an increasing emphasis on data protection has matched the growing reliance on data within the global economy. Within this evolution, the European Union (EU) led the charge with the adoption of the General Data Protection Regulation (GDPR) in 2016 [26]. Since then, a number of regions, including Japan, Brazil, Canada, California, Colorado, Connecticut, Virginia, and Utah, have followed suit with similar, and sometimes dissimilar, privacy-oriented laws.[1]

[1]See, eg, Brazil's Lei Geral de Proteçao de Dados (LGPD), Law No 13,709, of 14 August 2018, amending Law No 12,965, of 23 April 2014 [10]; Japan's Act on the Protection of Personal Information, Act No 57 of 2003, as amended in 2016 [35]; see also [49], pp. 431–48, discussing the impact of the GDPR across numerous regions on the world, and [37].

Generally, under each regional privacy law exists a concept of personal data or personally identifiable information. The definition of personal data encompasses two concepts: data that directly identifies the individual, or direct identifiers, and data that indirectly identifies the individual, or indirect identifiers (*e.g.*, GDPR Art. 4(1) [26]; California Consumer Privacy Act (CCPA), 1798.140(o)(1) [11]). Additionally, these regulations recognize that data can be non-personal. In essence, the regulations have provided for a spectrum of data: data that is clearly personal, data that is reasonably likely to identify an individual, and data that is non-personal, *i.e.*, non-identifiable to the individual who is related to the data collected.[2]

Data that directly identifies an individual is generally easily defined and identified (*e.g.*, name, address, social security number). Defining personal data that indirectly identifies an individual (*e.g.*, gender, date of birth, physical characteristics) is more challenging. Indirect identifiers sit on a sliding scale, from identifiable information (and therefore subject to data protection regulation) to non-identifiable information (and therefore outside of the scope of data protection regulations) (Id.). Built into the concept of personal data is the concept of reasonableness or proportionality: whether the identifiers provided are reasonably likely to identify an individual (*e.g.*, GDPR Recital 26 [26]). These can be identifiers that are individually considered personal data or in combination with other identifiers that together create identifiable information. Finally, these indirect identifiers also often relate to Inferences that can be made from the data itself.

Privacy laws also recognize, either expressly or inherently, that techniques exist to convert personal data to non-personal data. This non-personal data is often exempt from data security requirements (GDPR, Recital 26 [26]; CCPA § 1798.140(O(3)) [11]).[3] However, these laws typically disregard that truly removing identifiable information from data is difficult, while re-identifying the individuals is becoming easier. Thus, data controllers and processors are presented with two options for complying with regional privacy laws: implement the required privacy and security controls for personal data (and bear the risks of that personal data), or attempt to convert personal data into non-personal data.

There are substantial benefits from converting personal data to non-personal data. Under Recital 26 of the GDPR, it states: "...the principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable," [26]. Therefore, if personal data is converted to non-personal data, also known as anonymized data, then none of the requirements within the GDPR would apply [4, 34]. This means that anonymized data can be stored indefinitely and does not require user consent to processing. Anonymized data is portable, meaning it can be used for any purpose and transported across borders. Further, data subjects are not required to be notified in the event that anonymized data is breached and any fines imposed under the GDPR would not apply.

Unfortunately, data privacy laws use numerous terms with different, and overlapping, definitions to distinguish personal from non-personal data, creating confusion and potential conflict between these legal frameworks. Within this backdrop of abstract legal definitions and the application of those legal definitions of personal data, a growing challenge exists: the lack of definitive criteria for when personal data has become non-personal, as well as an understanding of the practical consequences of applying such criteria. This cre-

[2] See GDPR, Recital 26 [26]; see also the California Consumer Privacy Act of 2018 ("CCPA"), § 1798.140(O(3) [11]).

[3] "De-identification or anonymization techniques can be used to remove the data set from the scope of regulations," [1].

ates uncertainty as to whether organizations are truly compliant with privacy law, and can make drawing the line between personal data and non-personal data a difficult endeavor.

For example, in Case 582/14 – Patrick Breyer v Germany ([14]), the European Court of Justice (ECJ) held that Internet Protocol ("IP") addresses, in certain circumstances, are considered personal data.[4] The ECJ made it clear that determining whether data is identifiable is context specific to the parties involved and the information available to those parties. For the website owner, it held that dynamic IP addresses were not personal data "since such an address does not directly reveal the identity of the natural person who owns the computer from which a website was accessed, or that of another person who might use that computer."¶38. However, it recognized that for a different party, *i.e.*, the internet service provider ("ISP"), a dynamic IP address could be considered personal data: "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person." ¶42. In essence, the ECJ recognized that the website owner alone did not maintain personal data via access to the dynamic IP address. However, since the ISP could be legally required to turn over the additional information necessary to identify an individual based on her IP address, then for both the ISP and the website owner, the dynamic IP address constituted personal data. ¶49.

This case, and subsequent interpretations of the boundaries of personal data, highlight the contextual challenges in distinguishing personal data from non-personal data. Overcoming these challenges requires privacy tools that help data controllers reason through the threats their data will be exposed to, what level of risk these threats pose to data subjects, and what amount of data transformation will be necessary to mitigate privacy risk while ensuring that data users receive useful data. However, there is a lack of clarity as to what levels of risk are acceptable in non-personal data, and which anonymization techniques achieve acceptably low risk. Depending on the dataset and the use case, this may result in datasets with widely varying privacy and security impacts and corresponding risks.

To help address these challenges, we propose a framework of criteria to evaluate the conversion of personal data to non-personal data under the GDPR. We choose to focus our framework in the context of trajectory micro-data, hereafter referred to as location data, to complement the many anonymization techniques for this type of data [28]. In doing so, we are the first to create a set of anonymization criteria for assessing whether anonymized location data has reasonably low privacy risk and falls outside of the scope of the GDPR. Given our framework, the questions then become: What are the practical, statistical solutions that could be applied to convert personal location data to non-personal location data according to the legal definition? How do differences in these solutions, which may be subtle, impact the actual privacy provided to individuals and the resulting use of the non-personal data?

To address these questions, we use our framework to evaluate two reasonable and straightforward statistical solutions applied to a unique location dataset from COVID-19 patients in South Korea in an attempt to convert personal location data to non-personal location data. This is the first assessment determining whether existing anonymization solutions for location data are capable of legally anonymizing location data, and how the differences in these solutions impact the privacy and utility of the resulting data. This assessment is vitally important to the legal privacy community because there is an absence of case law showing when data meet legal anonymization criteria, and practitioners are likely to start

[4] In Patrick Breyer v. Germany, the ECJ made its decision under the precursor to the GDPR, the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (OJ 1995 L 281, p. 31, [25]). However, its ruling still applies under the GDPR.

with simple and well understood solutions for anonymization. We find that both methods are viable, but result in differing impacts to both the privacy and usefulness of the data. While we will ultimately require case law to determine reasonable levels of anonymization, our work provides organizations with a tool to document and show reasonable efforts to anonymize personal location data [13].

This paper proceeds as follows. In Sections 2 and 3, we outline the legal criteria used to determine when data is considered legally anonymized, examine existing work that quantifies legal anonymization criteria, and discuss how our work fits into the literature on protecting location data. In Section 4, we use legal anonymization criteria to create a framework of anonymization criteria for location data. This framework is a tool for determining whether or not location data is legally anonymized, and therefore becomes non-personal data, which falls outside of privacy legislation. In Section 5, we use our framework to evaluate the two simple statistical solutions applied to the location trajectories from COVID-19 patients in South Korea. Section 6 concludes with legal recommendations and future areas of research.

2 Legal Criteria for Converting Personal Data to Non-Personal Data

The GDPR recognizes the term pseudonymization, defining it (GDPR, Art. 4(5) [26]) as

> "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

While not defined with the GDPR itself, the EU uses the term "anonymous information" to describe data that falls outside of the requirements of the GDPR. Anonymous information consists of "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." (GDPR, Recital 26 [26]). It is important to note that while necessary to render data anonymous, pseudonymization alone produces data that still falls under the GDPR.

The European Data Protection Board ("EDPB"), charged with creating a cohesive interpretation and approach to data protection across the EU, adopted additional guidance on anonymization. The April 2020 Guidelines (¶15-16, [22]) on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, provide that (Id., at ¶15; see also GDPR, Recital 26)

> "Anonymisation refers to the use of a set of techniques in order to remove the ability to link the data with an identified or identifiable natural person against any "reasonable" effort. This "reasonability test" must take into account both objective aspects (time, technical means) and contextual elements that may vary case by case (rarity of a phenomenon including population density, nature and volume of data)."

These guidelines further outline three criteria used to determine the "robustness of anonymization" used to convert personal data to non-personal data (¶ 16, [22]):

> (i) singling-out (isolating an individual in a larger group based on the data);
>
> (ii) Linkability (linking together two records concerning the same individual); and
>
> (iii) Inference (deducing, with significant probability, unknown information about an individual).

These criteria are derived from earlier guidance from the EU, under the EU's precursor to the GDPR, the Article 29 Working Party Guidelines 05/2014 ("WP Guidelines" [7]). Both the Working Party and the EDPB Guidelines emphasize the applicability of the "reasonability test" to the concept of anonymization. The Working Party states that the test for whether an anonymization technique is sufficiently robust is "whether identification has become reasonably impossible," (¶ 2.2.2; see also 2020 April Guidelines [22], ¶¶ 15, 22). Therefore, these criteria are assessed under the reasonability test to determine if data is in fact anonymized, or remains personal data subject to the GDPR. Presently, there is an absence of case law for meeting these criteria which we seek to investigate in this paper.

Robustness of anonymization is key to ensuring that personal data becomes, and remains, non-identifiable. The risk of reidentification is a growing concern (¶19, [22]). Paul Ohm [44] argues that "reidentification science exposes the underlying promise made by these [privacy] laws– that anonymization protects privacy–as an empty one, as broken as the technologists' promises." But, the law does not require complete inability to re-identify data. It relies on a reasonableness criterion that creates an added layer of nuance to the anonymization process.

The WP Guidelines [7] provide further guidance on anonymization processes, stating:

> "...anonymisation techniques can provide privacy guarantees and may be used to generate efficient anonymization processes, but only if their application is engineered appropriately - which means that the prerequisites (context) and the objective(s) of the anonymization process must be clearly set out in order to achieve the targeted anonymization while producing some useful data. The optimal solution should be decided on a case-by-case basis...," (pp. 3-4).

There are two key takeaways from this statement. First, the objective of an efficient anonymization process should be some level of targeted anonymization that also produces useful data. Any reasonability test that is useful in practice will therefore weigh the privacy of the data against its utility in determining whether data is non-personal. An apt description of this balancing act is given by [23]:

> "Ideally, we'd like to have a data set that has both maximal privacy protection and maximal usefulness. Unfortunately, this is impossible. Like Goldilocks, we want to fall somewhere in the middle, where privacy is good, but so is data utility," (pp. 22).

An anonymization process that disregards utility will likely fail to produce useful data. Second, in order to balance privacy and utility, two aspects need to be accounted for: the context and the objectives of the anonymization. A one-size-fits-all approach that does not account for these aspects would not allow for optimal solutions to be decided on a case-by-case basis. The Working Party [7] further states that removing directly identifying elements from data is not sufficient for anonymization, and that, "It will often be necessary to take additional measures to prevent identification, once again *depending on the context*

and purposes of the processing for which the anonymised data are intended," (pp. 9, emphasis added).

A proper reasonability test will identify both the context and objectives of the anonymization process, and determine the level of risk at which identification has become reasonably impossible and at which useful data is produced. If anonymized data achieves this level of risk, then it falls outside the scope of privacy law. Ultimately, this reasonableness concept makes it impractical to define a single set of criteria for anonymization within the opaque guidance of the law. The EU's definition of anonymization also presents an added challenge to the usefulness of data, since maintaining good statistical inferences about individuals reduces the chances of successfully converting personal data to non-personal data. This challenge may be exacerbated in contexts with very high privacy risk, *i.e.*, public data release. In any context, anonymization processes directly impact the usefulness of data, which can range from pure noise with strong privacy guarantees, to well-preserved statistics with weak privacy guarantees.

The complete exploration of all privacy regulations, definitions, and practical implications are beyond the scope of this paper. Instead, we utilize the EDPB's three criteria of anonymization in the context of the GDPR. These criteria appear to encapsulate many of the criteria that various legal definitions consider when determining if data is non-personal[5]. Exploration of other regional data protection regulations also provides little additional insight into the statistical interpretation of anonymization. We consider the EDPB's definition of anonymization criteria as exhaustive for the variety of legal frameworks and regions and focus our study on these criteria for the remainder of this paper.

3 Related Work

The privacy risks associated with location data are well documented (*e.g.*, [40]) and have spawned a growing literature examining anonymization techniques for such data. Two factors motivate our study of a location dataset in the context of legal data anonymization: (1) Location data has a wide array of uses, ranging from location based targeting [32] to the study of disease dynamics; and (2) it is notoriously difficult to anonymize due to the uniqueness of individuals' location trajectories [28]. Existing work has examined linkage, homogeneity, and probabilistic attacks against location data, as well as the privacy principles that oppose these attacks, namely *indistinguishability* and *uninformativeness* [28]. Privacy criteria, such as k-anonymity and differential privacy, have also been applied to location data to quantify these privacy principles. What is lacking is a connection between the privacy principles in the literature on anonymizing location data and the anonymization requirements dictated by privacy law, which we seek to address.

Evaluating an anonymization process using the EDPB's anonymization criteria requires criteria that distinguish personal from non-personal data. To date, there has been relatively little work in this area. While [12] provide an in-depth interpretation of Singling Out which aims to provide a universal standard, they do not discuss interpretations of Linkability or Inference, and do not incorporate the reasonability test or the context of the anonymization process into their interpretation. Their mathematical interpretation of Singling Out is designed with the worst-case scenario in mind, *e.g.*, the data is subjected to a computationally unbounded adversary with complete knowledge of the data generating distribution. However, this is not likely to occur in real-world anonymization scenarios. Preventing such an

[5]See the appendix for the analysis of regional privacy laws which led us to this conclusion.

adversary from re-identifying data requires anonymization methods that severely reduce data utility in many cases, *e.g.*, differential privacy [8, 48, 28], and likely goes above and beyond what would be determined as reasonable. Often, a data controller wishes to share personal data with processors who themselves have security controls for their data. Legal anonymization should incorporate this and other factors, such as the technical abilities of the recipient of anonymized data and the amount of time that the anonymized data will be available, to define context-specific anonymization requirements, as recommended by the Working Party. The authors in [29] take such an approach and interpret Singling Out, Linkability, and Inference in the context of protecting queries submitted to a dataset. We also take a context-specific approach, and define a framework of anonymization criteria in the context of location data.

We highlight two other examples of context-based anonymization perspectives. The first is known as the Five Safes, which assesses the safety of an anonymization process in five areas: the project, people, setting, data, and outputs [6, 5]. These areas define the context of the anonymization process, predict the level of risk which data will be exposed to, and inform the amount of protection that is required to anonymize the data while maintaining its utility. Our work fits in the safe data portion of the Five Safes. We measure the risks inherent in personal location data using the criteria outlined by the law. Once these risks are defined, anonymization solutions can be used to reduce these risks to a legally acceptable level while retaining data utility.

The second example is the perspective on anonymization provided by the European Medicines Agency (EMA). The EMA has offered guidance on anonymizing clinical data based on legal frameworks and other standards including the WP Guidelines [24]. The EMA states that either Singling Out, Linkability, and Inference must be prevented, or the identification risks must be deemed acceptably low. Ultimately, a given dataset can be anonymized in different ways, each of which may be adequate depending on the risk of re-identification inherent in the context in which anonymized data is disclosed. In any context, the EMA supports maximizing the utility of the data for scientific study while adequately anonymizing the data. Similarly, in our empirical application, we evaluate two statistical anonymization solutions for location data, and examine the tradeoff between utility and privacy once the data is legally anonymized.

4 Framework of Legal Anonymization Criteria

In this section, we examine the three criteria used to evaluate the robustness of anonymization as described in the EDPB Guidelines [22]. We define when Singling Out, Linkability, and Inference have been prevented, and location data has been successfully converted from personal data to non-personal data. Per EDPB guidelines, the criteria constitute a framework that can be used to perform the reasonability test to evaluate anonymized data. Recall the three risks outlined by the EDPB:

> **Singling-out**: *isolating an individual in a larger group based on the data*
>
> **Linkability**: *linking together two records concerning the same individual*
>
> **Inference**: *deducing, with significant probability, unknown information about an individual*

We examine each of these criteria individually, highlighting their application using a simple example. For notation purposes, we consider a dataset $\mathbf{Y}$ which has been pseudonymized

through the removal of any direct identifiers. The dataset $\mathbf{Y}$ contains M rows, each row assumed to correspond to a unique individual, labeled by $[M] = \{1, \ldots, M\}$ and indexed by $m \in [M]$, and N columns, labeled $[N] = \{1, \ldots, N\}$ and indexed by $n \in [N]$. Note that the dataset $\mathbf{Y}$ is a collection of rows; $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]^\mathbf{T}$ where a row $\mathbf{y}_m = (y_{m,1}, \ldots y_{m,N})$ and $y_{m,n}$ denotes the value in row m and column n.

4.1 Singling-out

Singling Out can be described as "isolat[ing] some or all records which identify an individual in [a] dataset," [12] and Singling Out has occurred when the statement "There is exactly one user that has these attributes," is correctly made by someone examining the data [29]. For our purposes, we have assumed each individual has one record in $\mathbf{Y}$. Therefore, Singling Out an individual amounts to isolating the individual's record from the rest of the records in $\mathbf{Y}$. We assume isolation occurs based on the data values $y_{m,n}$ contained in each record - that is, if an adversary can correctly identify a record $\mathbf{y}_m$ in the anonymous data that differs from all other records based on a subset of the values $y_{m,n} \in \mathbf{y}_m$.

Define $\hat{\mathbf{Y}}$ as the matrix formed from the unique rows found in the dataset $\mathbf{Y}$. A row is unique if it differs from all other rows on at least one value $y_{m,n}$. Note that $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_I]^\mathbf{T}$ where $\hat{\mathbf{y}}_i$ is a unique row indexed by $i \in \{1, \ldots, I\}$ where $I \leq M$. Define a counting measure $\mu(\hat{\mathbf{y}}_i) = \#\{m \in [M] : \mathbf{y}_m = \hat{\mathbf{y}}_i\}$ as the number of times row $\hat{\mathbf{y}}_i$ appears in $\mathbf{Y}$, and define vector $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_I) \in \mathbb{N}^I$ with $\hat{z}_i = \mu(\hat{\mathbf{y}}_i)$ and $\sum_i \hat{z}_i = M$. Singling-out occurs when $\hat{z}_i = 1$ for some $i \in [I]$ and is prevented when

$$\inf \hat{\mathbf{z}} \geq k, \ \forall k \geq 2. \tag{1}$$

Singling-out is prevented when the minimum number of occurrences of any row $\mathbf{y}_m$ in $\mathbf{Y}$ is at least two and no individual can be isolated. To illustrate, consider the following dataset $\mathbf{Y}$ with records containing location trajectories and an single categorical attribute.

	L1	L2	L3	Sex
1	60.123, 120.827	60.124, 120.831	60.438, 121.002	F
2	60.124, 120.829	60.128, 120.834	60.431, 121.004	F
3	61.879, 122.384	61.243, 121.939	61.001, 121.822	M
4	61.872, 122.386	61.241, 121.933	61.004, 121.823	F
5	58.847, 119.998	58.294, 120.353	59.559, 121.491	F
6	58.843, 119.991	58.295, 120.358	59.555, 121.496	M

Table 1: Dataset $\mathbf{Y}$.

Every individual in this dataset can be singled out based solely on one of the location attributes {`L1`, `L2`, `L3`}. Singling Out can be prevented by reducing the uniqueness of each record. This can be achieved by applying generalization or suppression to the attributes in $\mathbf{Y}$ with the goal of creating subsets of records, or equivalence classes, that share the same attribute values. An example of such an approach is truncating the decimal places in the latitude and longitude measurements in $\mathbf{Y}$ [46], which we illustrate here. Since `Sex` values cannot be generalized using a natural hierarchy, we will suppress non-matching `Sex` values in equivalence classes.

	L1	L2	L3	Sex
1	60.12, 120.82	60.12, 120.83	60.43, 121.00	F
2	60.12, 120.82	60.12, 120.83	60.43, 121.00	F
3	61.87, 122.38	61.24, 121.93	61.00, 121.82	*
4	61.87, 122.38	61.24, 121.93	61.00, 121.82	*
5	58.84, 119.99	58.29, 120.35	59.55, 121.49	*
6	58.84, 119.99	58.29, 120.35	59.55, 121.49	*

Table 2: Generalized dataset $\mathbf{Y}$ with Singling Out prevented.

By truncating the location measurements from three to two decimal places, we create three equivalence classes, separated by the thicker horizontal lines in Table 2. Each equivalence class contains only two records. Depending on the context and objectives of the anonymization process, this may be deemed sufficient, although privacy could be improved by choosing a larger value of k. For this example, however, it is not possible to create larger equivalence classes using the truncation technique even if all decimals are removed. This illustrates the difficulty of preventing Singling Out in location datasets especially with a small number of records. By our measurements, the geographic area represented by spatial points with two decimal places is approximately 1110 by 892 meters, meaning there is significant uncertainty about the precise locations of these individuals, and utility will be reduced significantly with further generalization.

The criterion in (1) is related to the concept of *k*-anonymity, with the key distinction being that records must match on *all* attributes in $\mathbf{Y}$ for (1) to be met, not just quasi-identifiers. In this sense, the prevention of Singling Out imposes a stricter requirement than traditional *k*-anonymity, and syntactic privacy methods may struggle to achieve (1) while maintaining data utility, especially in high dimensions [3]. Still, *k*-anonymity has been highlighted by the Working Party as an anonymization method that could be used to prevent Singling Out [7].

The Working party also provides several warnings on the use of *k*-anonymous algorithms to prevent Singling Out [7]. The first, which our criterion accounts for, is that it is inappropriate to exclude any quasi-identifiers from the *k*-anonymity calculation. The second is that equivalence classes should not contain uneven distributions of attributes. From a privacy perspective, this is a valid concern as it relates to the ability of an adversary to infer unknown attribute values. However, the prevention of singling out requires all records in an equivalence class to have the *same* attribute values. As such, this warning is more relevant to the Inference criterion, which we discuss later. The Working Party also warns against small values of k which may result in lower privacy. The value of k should be chosen to provide a reasonable balance between privacy and utility, and low values may be appropriate in contexts with high data security and other safety measures.

Whether an individual can be isolated based on their data (*i.e.*, singled out) is directly related to the concept of *unicity*, or the uniqueness of the location trajectories in a given dataset. Past research has demonstrated the relative ease with which individual's location traces can be singled out [40], and significant work has gone toward developing methods for reducing the unicity of location data, thereby improving privacy. A large body of work has been devoted to achieving variants of *k*-anonymity in the location data privacy literature, *e.g.*, [9, 41, 39]. See [28] for a more detailed review.

While (1) constitutes the prevention of Singling Out, it does not provide adequate overall privacy. It is well known that *k*-anonymity fails to prevent composition or inference attacks

[30]. In fact, the prevention of Singling Out enables these attacks to be *more* successful when attributes have the same value and are not suppressed. We will examine composition and inference attacks later as they relate to the risks of Linkability and Inference.

4.2 Linkability

Similar to [29], we define Linkability to mean that a one-to-one linkage occurs between records for some individual in $\mathbf{Y}$ and an external dataset $\mathbf{X}$, containing some or all of the same attributes for individuals in $\mathbf{Y}$. There are three types of linkage that do not constitute Linkability, and are acceptable: (1) a one-to-many linkage (one row in $\mathbf{X}$ to multiple rows $\mathbf{Y}$) (2) a many-to-many linkage (multiple rows in $\mathbf{X}$ to multiple rows $\mathbf{Y}$) and (3) a many-to-one linkage (multiple rows in $\mathbf{X}$ to one row in $\mathbf{Y}$).

Suppose the external linking dataset $\mathbf{X}$ contains P rows, each row identifying a unique individual, indexed by $p \in \{1, \ldots, P\}$, and Q columns indexed by $q \in \{1, \ldots, Q\}$ where row p is denoted $\mathbf{x}_p = (x_{p,1}, \ldots x_{p,Q})$. To be conservative in terms of privacy, we assume that the external dataset includes at least some data on all individuals in $\mathbf{Y}$ where $M \leq P$, i.e., the set of individuals with records in $\mathbf{Y}$ is a subset of the individuals with records in $\mathbf{X}$. This precludes the one-to-many linkage scenario as there will always be a row in $\mathbf{X}$ corresponding to each row in $\mathbf{Y}$.

There exists a nonempty subset of columns $\mathcal{K} \subseteq [N]$, with size $K \equiv |\mathcal{K}| \in [N]$ that are contained in both $\mathbf{Y}$ and $\mathbf{X}$, which we term common information (CI). For $\mathbf{y}_m$ a row in $\mathbf{Y}$, let $\bar{\mathbf{y}}_m = (y_{m,k}, k \in [K])$ denote a CI tuple which is the truncation of $\mathbf{y}_m$ leaving only the CI. The resulting matrix of unique CI tuples found in $\mathbf{Y}$ becomes the $J \times K$ matrix $\bar{\mathbf{Y}} = [\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_J]^{\mathbf{T}}$ with unique CI tuples indexed $j \in \{1, \ldots, J\}$ where $J \leq I$. Let the counting measure $\mu(\bar{\mathbf{y}}_j) = \#\{i \in [I] : \hat{\bar{\mathbf{y}}}_i = \bar{\mathbf{y}}_j\}$ denote the number of times the unique CI tuple $\bar{\mathbf{y}}_j$ is the result of shortening each of the I unique rows $\hat{\mathbf{y}}_i$ from $\hat{\mathbf{Y}}$, retaining only each row's CI, i.e., the N-vector $\hat{\mathbf{y}}_i$ is shortened to the K-vector $\hat{\bar{\mathbf{y}}}_i$. Finally, define the vector of counts $\bar{\mathbf{z}} = (\bar{z}_1, \ldots, \bar{z}_J) \in \mathbb{N}^J$, where $\bar{z}_j = \mu(\bar{\mathbf{y}}_j)$ is the count for unique CI tuple $j \in [J]$ and $\sum_j z_j = I$.

Since the external dataset includes at least some data on all individuals in $\mathbf{Y}$, the dataset $\mathbf{Y}$ can prevent Linkability when $\inf \bar{\mathbf{z}} \geq 2$ for all possible $\mathcal{K} \subseteq [N]$ which results in a many-to-many linkage. Note, $\inf \bar{\mathbf{z}} \geq 2$ for any $\mathcal{K} \subseteq [N]$ means there is no definition of CI under which any record from $\mathbf{Y}$ may be immediately linked with a public record from $\mathbf{X}$. However, $\bar{\mathbf{z}}$ will be difficult to determine as the organization may have no knowledge of which columns in $\mathbf{Y}$ are CI in the external dataset from the $2^N - 1$ possibilities. Conservatively, we suggest the organization assume a full CI with $K = N$ and prevent Singling-out by employing a statistical solution that ensures $\inf \hat{\mathbf{z}} \geq 2$. Once this is achieved, any row $\mathbf{y}_m$ is non-unique relative to $\mathbf{X}$. Therefore, any CI-truncation $\bar{\mathbf{y}}_j$ must also be non-unique, which implies that $\inf \bar{\mathbf{z}} \geq 2$, i.e., Linkability is prevented.

When Singling-out is not prevented, a many-to-one linkage (multiple rows in $\mathbf{X}$ to one row in $\mathbf{Y}$) or a one-to-one linkage is possible. Preventing Linkability occurs with a many-to-one linkage but it is difficult to show without knowledge of $\mathbf{X}$. In practice, we suggest removing unique rows or non-essential columns in $\mathbf{Y}$ to prevent both types of linkages.

To illustrate Linkability, we return to our example from Section 4.1. Suppose that an adversary has access to an external dataset $\mathbf{X}$. For simplicity, we show only a subset of $\mathbf{X}$. The common information columns $\mathcal{K}$ are the three location attributes {`L1`, `L2`, `L3`}. Within this subset of $\mathbf{X}$, records two and six correspond to individuals that also have records in $\mathbf{Y}$.

	L1	L2	L3
1	65.563, 125.824	65.229, 125.912	64.277, 124.800
2	60.124, 120.829	60.128, 120.834	60.431, 121.004
3	68.162, 123.688	68.253, 124.765	69.100, 123.134
4	58.245, 119.072	60.621, 118.834	61.356, 119.081
5	75.136, 130.364	73.753, 129.624	74.274, 130.421
6	58.843, 119.991	58.295, 120.358	59.555, 121.496

Table 3: External dataset $\mathbf{X}$.

	L1	L2	L3	Sex
1	60.123, 120.827	60.124, 120.831	60.438, 121.002	F
2	60.124, 120.829	60.128, 120.834	60.431, 121.004	F
3	61.879, 122.384	61.243, 121.939	61.001, 121.822	M
4	61.872, 122.386	61.241, 121.933	61.004, 121.823	F
5	58.847, 119.998	58.294, 120.353	59.559, 121.491	F
6	58.843, 119.991	58.295, 120.358	59.555, 121.496	M

Table 4: Dataset $\mathbf{Y}$.

One-to-one linkage can occur between records two and six in the subset of $\mathbf{X}$ and their corresponding matches in $\mathbf{Y}$ using $\{\texttt{L1}, \texttt{L2}, \texttt{L3}\}$ when no data protection is applied. Applying generalization and suppression as before results in the new $\mathbf{Y}$ shown below.

	L1	L2	L3	Sex
1	60.12, 120.82	60.12, 120.83	60.43, 121.00	F
2	60.12, 120.82	60.12, 120.83	60.43, 121.00	F
3	61.87, 122.38	61.24, 121.93	61.00, 121.82	*
4	61.87, 122.38	61.24, 121.93	61.00, 121.82	*
5	58.84, 119.99	58.29, 120.35	59.55, 121.49	*
6	58.84, 119.99	58.29, 120.35	59.55, 121.49	*

Table 5: Dataset $\mathbf{Y}$ with Singling Out and Linkability prevented.

Now, only a one-to-many linkage can occur between record two in the subset of $\mathbf{X}$ and records one and two in $\mathbf{Y}$, and record six in the subset of $\mathbf{X}$ and records five and six in $\mathbf{Y}$. If all records in $\mathbf{X}$ were examined, only many-to-many linkages could occur. Hence, Singling Out and Linkability are prevented in $\mathbf{Y}$.

While one-to-one linkage is prevented when Singling Out is prevented, the WP Guidelines point out that the probability that an external record and a record with matching attributes in an equivalence class of size k correspond to the same individual is $1/k$ [7] (assuming that the external record only matches to one equivalence class). An upper bound on this probability for all records in $\mathbf{Y}$ is given by $1/\inf \hat{\mathbf{z}}$. Choosing a higher value for k in (1) reduces this upper bound, thereby improving privacy.

Record linkage attacks have been widely studied in the privacy literature for location data, and most of the proposed protection methods focus on preventing these attacks [28]. However, preventing singling out and linkability does not prevent composition or inference attacks [30]. This means that an adversary can deduce additional information pertaining

to an individual using the records in $\mathbf{Y}$ that match their record in $\mathbf{X}$, even if Singling Out and Linkability are prevented. In the location data privacy literature, this is known as an attribute linkage attack, or homogeneity attack, where an adversary cannot perform a one-to-one linkage between records, but the matching records in $\mathbf{Y}$ correspond to a single attribute value (or one attribute value with high probability), which the adversary can link to a record of interest [28]. This leads to the examination of the final legal privacy risk, Inference.

4.3 Inference

Inference is deducing, with significant probability, unknown information about an individual. Recall that evaluating whether $\mathbf{Y}$ prevents Inference depends on a reasonability test where "a solution against [Inference] would be robust against re-identification performed by the most likely and reasonable means the data controller and any third party may employ," (pp. 12, [7]). In the context of location data, unknown information could take many forms, such as location measurements, personal attributes, or sensitive information such as home or work addresses. As described in [29], we consider an adversary's goal with Inference to be able to correctly make statements of the form, "the value of this attribute is s" based on the information in $\mathbf{Y}$, for some individual(s) of interest. According to the legal definition of Inference, we need to limit the probability that such statements are correctly made.

We define Inference in the context of our dataset $\mathbf{Y}$, and suppose an adversary seeks to use some external information on individual m, $\mathbf{x}$ (a row in $\mathbf{X}$ corresponding to individual m), to identify the value of a sensitive categorical (or intervaled) variable for individual m in $\mathbf{Y}$. Let S denote the discrete random variable and s denote the value of the random variable. Let s^* be the true value for individual m that the organization is trying to prevent Inference on. The adversary has a prior probability $p(s^* \mid \mathbf{X})$ of the individual having a value s^* based on external information only. Once $\mathbf{Y}$ is released, the adversary obtains the probability $p(s^* \mid \mathbf{x}, \mathbf{Y})$ based on the new information. The pseudonymized dataset $\mathbf{Y}$ prevents Inference for an individual m when

$$p(s^* \mid \mathbf{x}, \mathbf{Y}) - p(s^* \mid \mathbf{X}) \leq c. \tag{2}$$

The change in probability on the left-hand side (LHS) will always be between -1 and 1. Per the EDPB's guidelines, we view (2) as the most likely and reasonable means that a third party would use with the value of c set between 0 and 1 based on their definition of significant probability. Note that if the LHS of (2) is negative, the criterion is not violated as Inference is worse on individual m. For a fixed threshold of c, (2) is unlikely to hold for all M individuals unless $\mathbf{Y}$ is strongly protected. Thus, one reasonability test may average the LHS of (2) over all individuals in cases where Inference is improved, $p(s^* \mid \mathbf{x}, \mathbf{Y}) > p(s^* \mid \mathbf{X})$. Otherwise, we suggest setting the LHS to 0 in cases where Inference is degraded, $p(s^* \mid \mathbf{x}, \mathbf{Y}) < p(s^* \mid \mathbf{X})$.

We revisit our example from Sections 4.1 and 4.2 to illustrate (2). Suppose an adversary seeks to deduce the `Sex` of an individual m in $\mathbf{Y}$ using only $\mathbf{X}$, which contains location data on individuals from the United States population. An adversary may base their prior belief over the `Sex` of individual m using the population probability mass function shown below [52]

$$p(s \mid \mathbf{X}) = \begin{cases} 0.505, & s = \text{F} \\ 0.495, & s = \text{M}, \end{cases}$$

which consists of the relative frequencies in the population for male (M) and female (F) `Sex` categories. In this case, the prior probabilities are approximately equal to randomly guessing a `Sex` for individual m. Prior knowledge should be reasonable to serve as a baseline for assessing whether Inference is prevented in a dataset $\mathbf{Y}$. Of course, we cannot limit the amount of external information available to an adversary. For example, the adversary may know the exact `Sex` of individual m without obtaining $\mathbf{Y}$ which implies that the probability of correct Inference can never be improved and the criterion is not violated. This highlights the external risks of re-identification and attribution, and the increasing emphasis that the law does not adequately address the privacy risks, even when data reaches the legally defined scope of anonymization.

Suppose the adversary's individual of interest corresponds to the second record in $\mathbf{X}$. For brevity, we examine Inference in $\mathbf{Y}$ after Singling Out and Linkability have been prevented.[6] In this example, $\mathbf{x}$ links to the first and second records in $\mathbf{Y}$ based on the CI.

	L1	L2	L3
1	65.563, 125.824	65.229, 125.912	64.277, 124.800
2	60.124, 120.829	60.128, 120.834	60.431, 121.004
3	68.162, 123.688	68.253, 124.765	69.100, 123.134
4	58.245, 119.072	60.621, 118.834	61.356, 119.081
5	75.136, 130.364	73.753, 129.624	74.274, 130.421
6	58.843, 119.991	58.295, 120.358	59.555, 121.496

Table 6: External dataset $\mathbf{X}$.

	L1	L2	L3	Sex
1	60.12, 120.82	60.12, 120.83	60.43, 121.00	F
2	60.12, 120.82	60.12, 120.83	60.43, 121.00	F
3	61.87, 122.38	61.24, 121.93	61.00, 121.82	*
4	61.87, 122.38	61.24, 121.93	61.00, 121.82	*
5	58.84, 119.99	58.29, 120.35	59.55, 121.49	*
6	58.84, 119.99	58.29, 120.35	59.55, 121.49	*

Table 7: Dataset $\mathbf{Y}$ with Singling Out and Linkability prevented.

This results in the following new probability mass function for `Sex` based on $\mathbf{Y}$ and $\mathbf{x}$

$$p(s \mid \mathbf{x}, \mathbf{Y}) = \begin{cases} 1.00, & s = \text{F} \\ 0.00, & s = \text{M}. \end{cases}$$

Under this scenario, the adversary deduces with one hundred percent probability that the `Sex` of individual m is F. Using (2),

[6] Inference is trivial when Singling Out and Linkability are not prevented. The adversary simply has to find the record $\mathbf{y}_m$ which links to $\mathbf{x}$ and they can infer the values of any attributes not in $\mathbf{x}$.

$$p(s^* \mid \mathbf{x}, \mathbf{Y}) - p(s^* \mid \mathbf{X}) = 1.00 - 0.505 = 0.495$$

which is almost certainly in violation of (2). The Inference condition would be similarly violated for the other individual in the same equivalence class. However, for the individuals corresponding to records three through six in $\mathbf{Y}$, (2) would likely not be violated. Since these records are contained in equivalence classes of two records with suppressed `Sex` values, the adversary knows that the values are different, and the updated probability mass function becomes

$$p(s \mid \mathbf{x}, \mathbf{Y}) = \begin{cases} 0.50, & s = \text{F} \\ 0.50, & s = \text{M}, \end{cases}$$

such that $p(s^* \mid \mathbf{x}, \mathbf{Y}) - p(s^* \mid \mathbf{X}) \in \{-0.005, 0.005\}$ for records three through six. If a reasonability test required measuring the average of the LHS of (2), we observe the following:

$$\frac{1}{3} * 0.495 + \frac{1}{3} * 0.005 = 0.167.$$

Another reasonability test may use t-closeness to measure the variational distance [15, 36] between $p(s \mid \mathbf{x}, \mathbf{Y})$ and $p(s \mid \mathbf{X})$. For the individuals in the first equivalence class, t-closeness is

$$\frac{1}{2}\Big[\big|1.00 - 0.505\big| + \big|0.00 - 0.495\big|\Big] = 0.495,$$

which is poor privacy. For the second and third equivalence classes, t-closeness is

$$\frac{1}{2}\Big[\big|0.50 - 0.505\big| + \big|0.50 - 0.495\big|\Big] = 0.005,$$

which is significantly lower and illustrates much better privacy. Preventing Inference for the individuals in the first equivalence class would require further generalizing the location data to create equivalence classes with non-homogenous sex values, or suppressing all values of `Sex`. This would severely degrade the utility of the data, highlighting the difficulty of legally anonymizing location data. Overall, whether the previous examples constitute violations of (2) depends on the value c and the choice of the sensitive variable, which must be reasonable and most likely, even if it does not actually result in truly non-identified information.

The Working Party has pointed out that k-anonymity provides no protection against Inference, and that extensions such as l-diversity and t-closeness are needed to limit the probability of correctly deducing unknown information about an individual [7]. This is reflected in our proposed criteria and highlighted by our example: once (1) is achieved, additional protection is needed to achieve (2) and prevent an adversary from performing Inference.

As highlighted by [28], very little work has examined attribute linkage attacks against location data (see [57, 56] for examples). Reducing the probability of the success of these attacks is essential in legally anonymized data, so we examine such an attack in our empirical application.

4.4 Relating Legal Anonymization Criteria to Privacy Principles

The work in [28] discusses two privacy principles which anonymization techniques are designed to achieve. The first is *indistinguishability*, where any record cannot be distinguished

from a larger group of records in the same dataset. Preventing Singling Out and Linkability achieves indistinguishability for location trajectories since individuals' records are no longer unique. Preventing Inference achieves indistinguishability at the attribute level by ensuring that an anonymized dataset does not improve an adversary's probability of deducing unknown information about an individual by more than c. Achieving our proposed criteria requires assumptions about the baseline knowledge of an adversary, and this baseline knowledge should be determined as part of the reasonability test based on the context in which anonymized data is released.

The second privacy principle discussed in [28] is *uninformativeness*, which requires that the difference in knowledge of the adversary before and after accessing the dataset $\mathbf{Y}$ must be small without making assumptions on an adversary's knowledge. This provides stronger protection than indistinguishability and is typically achieved using differential privacy, which can be used to bound increased Inference on all individuals [19]. However, this may currently be beyond the technical abilities or interests of most organizations (*e.g.*, [29] only cite two known operational deployments of differential privacy which both operate with unbounded privacy loss). Although we do not address differentially private mechanisms in this paper, we refer the interested reader to recent research in this area [60, 33, 47]. For non-technical readers and the legal community, differential privacy is a theoretical guarantee of privacy and prevents increased Inference on all individuals by a bound of $\exp(\epsilon)$. Unlike (2), it is a bound in the odds ratio (instead of a subtraction) of datasets differing on at most one record. Furthermore, it would be more similar to measuring the worst-case scenario for any individual in the dataset $\mathbf{Y}$ for (2) which we explore in the next section.

5 Application to South Korean COVID-19 Location Data

5.1 Data Description

We use a location dataset collected by the Korean Centers for Disease Control and Prevention (KCDC) containing several Global Positioning System (GPS) coordinates for COVID-19 positive individuals in South Korea [18]. This longitudinal data is useful for epidemiologists and public health officials in tracking and mitigating the spread of COVID-19. However, location trajectories are also associated with severe privacy issues especially when a GPS coordinate can be linked to a home address or sensitive locations [16]. We build on the notation introduced in Section 4 and define a released longitudinal location dataset $\mathbf{Y}^{\ell,d}$. The superscript ℓ denotes that rows have longitudinal data and d denotes the number of decimal places used for latitude and longitude coordinates.

The original dataset $\mathbf{Y}^{\ell,d}$ with $d = 5$ is a collection of $M = 1,472$ location trajectories of COVID-19 positive patients tracked from January 20, 2020 to June 01, 2020; denoted $\mathbf{Y}^{\ell,d} = (\mathbf{y}^{\ell}_m,\, m \in [M])$ where $\mathbf{y}^{\ell}_m = (y_{m,1}, y_{m,2}, \ldots, y_{m,N_m})$ is the trajectory for individual m and $y_{m,n}$ denotes the n-th location tuple for individual m. There are N_m location tuples in the location trajectory for individual m, where N_m ranges from one to forty-five tuples for all individual. We delete trajectories with $N_m < 5$, resulting in 595 trajectories across all of South Korea, and 247 trajectories in Seoul. Additionally, we delete the timestamps associated with the trajectories because nearly all of the trajectories are unique with them. After this redaction, 97.31% of individuals have a unique location trajectory and 63.73% of location tuples in this dataset are unique. Hence, without further disclosure limitation approaches, the legal criterion of preventing Singling-out (and hence the definition of non-

personal data) is not met and would require the deletion of nearly all of the trajectories.

5.2 Data Use Cases

We show how our framework can be used to assess context specific anonymization by considering two separate use cases for the dataset $\mathbf{Y}^{\ell,d}$. The first use case we consider is the analysis of individuals' location trajectories to study the movement patterns of diseased individuals. The second use case is releasing aggregated counts data, such as Facebook's aggregated daily movement data during the COVID-19 pandemic [17], which is publicly available. This gives aggregate information on the number of individuals present in a set of relatively large geographical areas. Note that this data disregards the longitudinal relationship between individuals' locations, and comes with lower privacy risks. These use cases require very different anonymized datasets (individual location trajectories vs. aggregate counts) and thus will require different anonymization techniques to "...achieve the targeted anonymization while producing some useful data," [7]. Our framework is used to assess whether the target level of anonymization is met, i.e., the framework criteria are met, while the strength of privacy protection is balanced against the utility of the data.

5.3 Converting Personal Data to Non-Personal Data

To fulfill the legal criteria to convert personal data ($\mathbf{Y}^{\ell,d}$ with $d = 5$) to non-personal data, the EDPB guidelines on anonymization within the context of geolocation tracking state that the reasonability test for geolocation data must "take into account both objective aspects...and contextual elements...including population density, nature and volume of data..."[22]. In accordance with these guidelines we separate our analysis into two geographical regions with varying population densities - the city of Seoul (approximately 605 km^2) and the country of South Korea (approximately 100,210 km^2). Figure 1 illustrates the difference in location densities between South Korea and Seoul. The number of observed location tuples across South Korea is approximately 0.05 per square kilometer, whereas in Seoul there are approximately four observed location tuples per square kilometer. In practice, additional contextual factors should be considered to find the proper balance between privacy and utility in legally anonymized data, such as the technical abilities of the recipients of anonymized data, and the length of time the anonymized data will be available for analysis.

The previously mentioned use cases drive the selection of two logical and relatively simple solutions to protect the original location trajectory dataset $\mathbf{Y}^{\ell,5}$: (1) Location Coarsening and (2) Aggregating to Counts. Both approaches fall under the scope of generalization which is an anonymization technique to protect against re-identification [22]. Additionally, both solutions alter the data with no randomness, producing a released location tuple (or count) consistent with the original data and palatable across the organization. We describe the solutions briefly first, then investigate whether they produce data that meets the legal criteria of preventing Singling-out, Linkability, and Inference, and examine the resulting data utility.

Before proceeding with the analysis, we remark on two common uses of geolocation data not studied in this paper: mandatory quarantines of COVID-19 positive individuals (*e.g.*, as seen in Hong Kong and Poland) and contact tracing (*e.g.*, as seen in Singapore, Israel, and South Korea) [43, 54]. Mandatory quarantines enforced using geolocation data do not meet the criteria of non-personal data because by definition, they require individuals to be

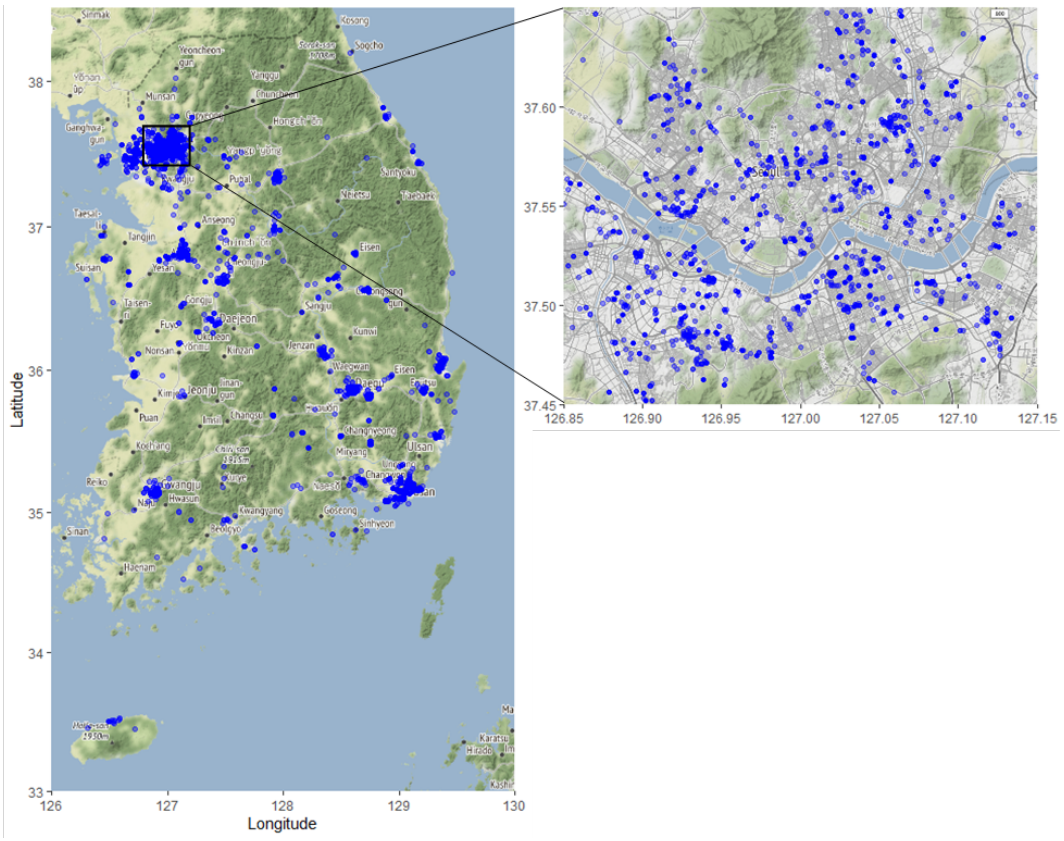

Figure 1: Location Tuples Mapped Over South Korea (left) and Seoul (right) for $(d = 5)$

Singled-out within a small geographical area. Contact tracing requires GPS data or Bluetooth data, but more commonly uses Bluetooth data and indicates whether an individual m had contact with another individual at a specific time (or duration) within a close proximity (*i.e.*, Bluetooth connection or a few meters with GPS data). Contact tracing is designed with the goal of contacting all individuals that have a first (or second) degree connection with a COVID-19 positive individual. This networked data mimics the properties of social network data (*e.g.*, Facebook or LinkedIn data) and fundamental privacy issues are discussed at length by [27] and [51].

5.3.1 Location Coarsening

To produce anonymized location trajectories for the first use case, we use Location Coarsening to round latitude and longitude coordinates in the original data to a lower number of decimal places, $d < 5$. Similar to truncation [46], this coarsens the location measurements in the trajectories. The superscript d will be used to denote the number of decimal places at which (latitude, longitude) GPS coordinates are coarsened to within the protected dataset $\mathbf{Y}^{\ell,d}$. For example, a coordinate released at $d = 3$ would be (37.593, 127.017). Under the law, coarsening is a form of generalization since there are many specific location tuples, *e.g.*, (37.59266, 127.0174), (37.59290, 127.0169), (37.59282, 127.0171), that correspond to the coarsened location tuple (37.593, 127.017). As a result, the number of coarsened location tuples when $d < 5$ is always at least as big as the number of location tuples when $d = 5$ for a given tuple. The same is also true for coarsened trajectories. We perform the Location Coarsening manually by rounding to d decimal places and Figure 2 illustrates the location tuples when $d = 1$. We use the geosphere package in R to measure the geographical distance (in meters) that points are shifted by Location Coarsening and use this distance to represent the utility of the coarsened data.

5.3.2 Aggregating to Counts

Aggregating to Counts sums the number of location tuples within a geographical distance of a pre-specified latitude and longitude coordinate. The longitudinal dataset $\mathbf{Y}^{\ell,5}$ is transformed to a dataset of counts which removes the longitudinal nature of the location trajectories. Define an Aggregating to Counts dataset $\mathbf{C}^d$ which counts how many of the $\sum_{m=1}^{M} N_m$ tuples are inside of a geographical region with a center defined by every latitude and longitude tuple with d decimal places in South Korea or Seoul. Note that as d decreases, the number of counts in $\mathbf{C}^d$ decreases. Additionally, we note that although generalization methods, such as Aggregating to Counts, are capable of meeting the criteria of anonymization, it has been shown that aggregated data, location-based and otherwise, are still susceptible to other privacy attacks [20, 55].[7] To balance the privacy and utility of the aggregated counts data, we let the size of the geographical regions for which counts are released represent the utility of the data, the idea being that counts are more informative for smaller geographical regions.

[7] Case law will determine whether these aggregation attacks are "reasonable" or not.

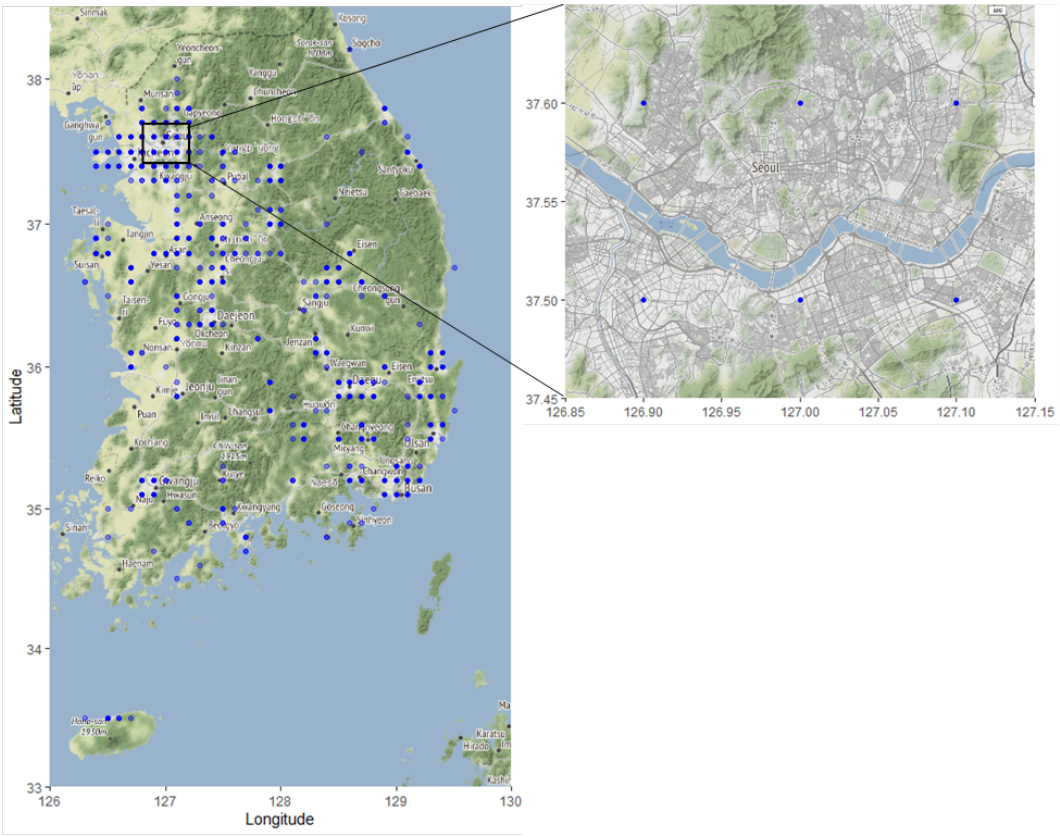

Figure 2: Location Tuples Mapped Over South Korea (left) and Seoul (right) for $d = 1$

5.4 Preventing Singling-out

5.4.1 Location Coarsening

To prevent Singling-out, no individual can have a unique location trajectory. A simple solution would be to delete all unique location trajectories in the original data, however this requires the deletion of 97.31% of the trajectories.

One issue for evaluating the effects of Location Coarsening for $d < 5$ is that the original location trajectories are of different lengths N_m from 5 to 45. To address this issue, we simulate 100 protected datasets for each value of d. These 100 simulated datasets are created by sampling (without replacement) one to five location tuples from each trajectory in the original dataset. We calculate the percentage of trajectories that are unique in each simulated dataset. We perform this analysis twice; once using the less dense South Korea data and once using the dense Seoul data.

Table 8 reports quantiles of distance shifted for each value of d and averages the number of unique trajectories. The results show that the usefulness of the coarsened data is not very good for low values of d. In order to prevent Singling-out, over 88% of the trajectories must be deleted when $d \geq 2$ and over 50% of the trajectories must be deleted in the entirety of South Korea when $d = 1$ where the median distance shifted is approximately 4 kilometers.

	Quantiles of Distances (in meters) Shifted			Mean % of Unique Trajectories	
d	2.5%	50%	97.5%	Full Data	Seoul Data
5	0.08	0.39	0.65	94.46%	94.38%
4	1.00	3.94	6.36	94.46%	94.33%
3	10.61	40.36	64.04	94.23%	94.03%
2	82.62	394.35	642.58	89.42%	88.03%
1	987.36	3,990.97	6,378.20	50.93%	20.19%
0	15,223.54	47,203.55	63,859.24	12.92%	1.62%

Table 8: *Location Coarsening*: Quantiles of distances shifted between coarsened tuples and original tuples (in meters) and the mean percent of unique trajectories

Figure 3 expands the detail of Table 8 by breaking down the statistics by sampled trajectory length. Each boxplot shows the percent of unique trajectories for each of 100 simulations given a sampled trajectory length and value of d. Without deletion, Location Coarsening does not prevent Singling-out at any value of d in Seoul or South Korea. However, the required number of deletions for trajectories is considerably less for the high density region of Seoul.[8] Overall, coarsening to $d = 1$ appears to provide the best balance between privacy and utility for this particular context, since most of the data must be deleted for $d \geq 2$, and the median distance shifted when $d = 0$ is over 47 kilometers.

5.4.2 Aggregating to Counts

Recall that Aggregating to Counts converts the original dataset $\mathbf{Y}^{\ell,5}$ to a protected dataset of counts $\mathbf{C}^d$ by removing both location tuples and trajectories. The magnitude of a count within a region is inversely related to the value of d; lower values of d produce higher counts. Singling-out is prevented in $\mathbf{C}^d$ if the minimum count value is two.

[8] Our results were obtained from location data without considering the temporal aspect that often accompanies this type of data. Including temporal information in location traces further raises the identifiability of individuals [46] and would greatly increase the difficulty of preventing Singling-out.

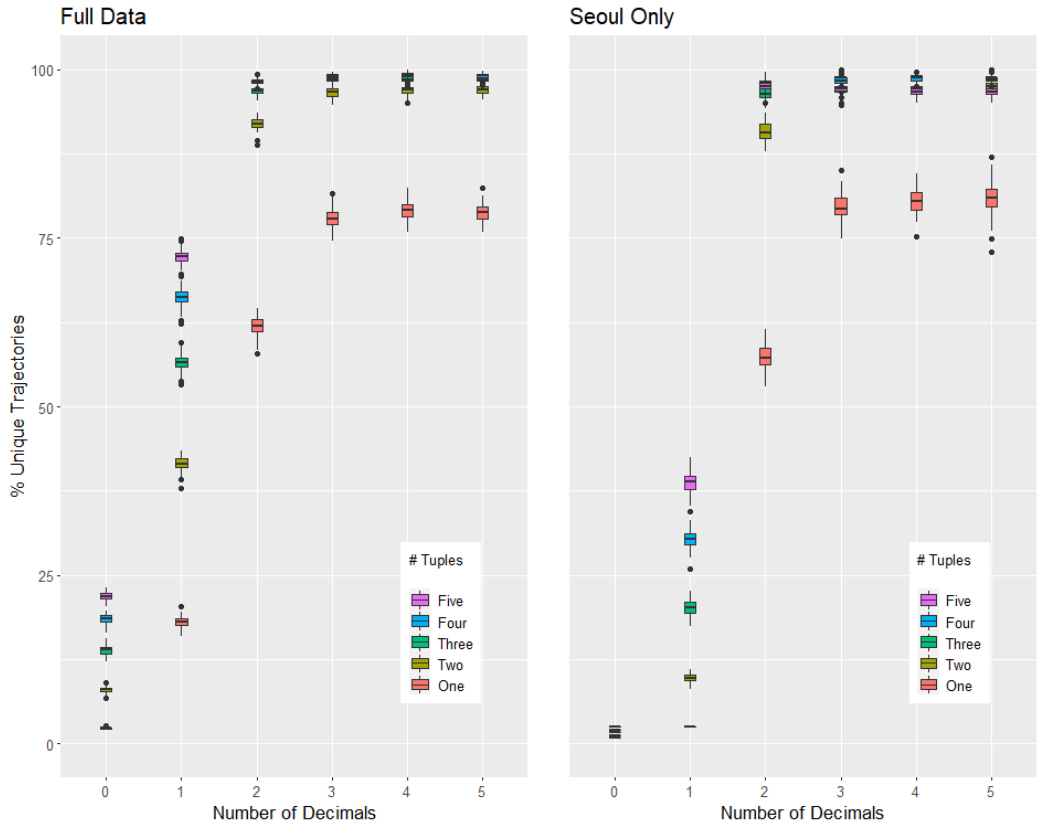

Figure 3: Percentage of unique trajectories for each value of $d = 0, 1, \ldots, 5$ and sampled trajectory lengths from one to five tuples in South Korea (left) and Seoul (right)

Table 9 shows that Singling-out is prevented when $d = \{0, 1\}$ in Seoul, *i.e.*, for large geographies (approximately 100 square kilometers for $d = 1$ and 10,000 square kilometers for $d = 0$) in a high density region. In order to prevent Singling-out for $d = 5$, approximately 63% of regions in $\mathbf{C}^5$ must be deleted which is an advantage over Location Coarsening. When $d = 2$ in Seoul, less than 20% of regions must be deleted to prevent Singling-out, which means counts can be kept on 233 geographies (approximately one square kilometer each), which is an improvement in usefulness over Location Coarsening at similar levels of protection.[9] For this use case, $d = \{1, 2\}$ provides the best trade-off between utility and privacy, depending on whether more granular statistics or more geographical coverage is desired. Choosing $d = 1$ ensures that counts are provided for more geographical area, while the counts themselves are drawn from large areas of 100 square kilometers. On the other hand, $d = 1$ provides counts on much smaller geographical areas, but covers less geographical area overall. An interesting note is that the Facebook implementation of aggregated mobility data omitted any region with fewer than three hundred individuals - a much higher value than the minimum necessary to prevent Singling Out [17]. Overall, when preventing Singling-out, Aggregating to Counts results in less deleted regions compared to Location Coarsening. However, this comes at the cost of removing the longitudinal information, *e.g.*, destroying travel patterns of COVID-19 positive individuals.

d	Longitudinal Length (m)	Latitudinal Length (m)	Region Area (km^2)	% (number) Regions with a count of 2 or more (Full)	% (number) Regions with a count of 2 or more (Seoul)
5	0.88	1.10	0.000001	36.40% (919)	37.23% (347)
4	8.91	11.09	0.0001	36.86% (924)	37.73% (349)
3	89.14	110.97	0.01	41.78% (963)	42.67% (361)
2	891.52	1,109.74	0.99	62.98% (711)	81.18% (233)
1	8,915.16	11,097.52	98.94	71.86% (166)	100% (6)
0	89,042.96	110,976.95	9,881.97	94.44% (17)	100 % (2)

Table 9: *Aggregating to Counts*: Areas of regions and percentage of regions preventing Singling-out

5.5 Preventing Linkability

When not preventing Singling out, preventing Linkability using Location Coarsening and Aggregating to Counts requires the knowledge of external linking datasets. In practice, we believe that an organization would just meet the Singling-out criterion to prevent Linkablity since both are required for the conversion to non-personal data. However, even when one-to-one linkage is prevented, the size of equivalence classes is directly related to the probability that an external record and a matching record in $\mathbf{Y}^{\ell,d}$ correspond to the same individual. We examine the distribution of these probabilities under all simulations for both protection methods. We also detail a reasonable process by which an organization could prevent Linkability when Singling-out is not prevented. To represent the worst-case

[9] Regions containing no location tuples with a count of 0 were not included in this analysis since no individuals existed in these locations. Also, the sample size of our dataset was much lower than the population of South Korea (about 51 million) which implies this analysis is a conservative lower bound of privacy.

scenario, we assume an adversary has a location dataset of the entire population, $\mathbf{X}^{\ell,d}$ with $d \geq 5$.

5.5.1 Location Coarsening

Using the same simulation described in Section 5.4.1, we calculate the inverse of the equivalence class size k for each *non-unique* trajectory for all simulated protected datasets. Figure 4 plots the distribution of $1/k$ across all simulations for each combination of d and sampled trajectory length.

In both the full South Korea and the Seoul data, we see that most trajectories are in equivalence classes of size two for $d \geq 2$, which is the minimum size for legally anonymized data. This may be an acceptable level of privacy in contexts with other privacy or security measures, although privacy can be improved by moving to $d = 1$, such that large proportions of records are placed into larger equivalence classes. Overall, the dense Seoul data exhibits larger equivalence classes for $d = 1$ than the full South Korea data. The best privacy occurs in the Seoul data when $d = 0$, and most trajectories are contained in equivalence classes of at least ten records. However, due to the massive reductions in data utility that occur at $d = 0$, $d = 1$ represents a better tradeoff between utility and privacy for both datasets (note that unique trajectories must still be deleted to legally anonymize the data).

When Singling-out is not prevented, there exists some individual m in $\mathbf{Y}^{\ell,d}$ with a completely unique trajectory. Suppose an adversary does not know the true identity of individual m and the adversary's external linking dataset $\mathbf{X}^{\ell,d}$ contains one or more individuals with trajectories that could be linked to individual m. As a result, the adversary may have a one-to-one linkage or a many-to-one linkage, which prevents Linkability.

To assess reasonableness, past research using mobile phone records [40] indicates that four spatio-temporal points are enough to uniquely identify 95 percent of individuals from a population of 1.5 million. These points correspond to geographical areas ranging between 0.15 to 15 square kilometers. In our data, Table 9 shows that location tuples coarsened to $d = 2$ fall within this square kilometer range. For, $d = 2$, Figure 3 shows that over 90% of trajectories are unique with only two location tuples and over 95% of trajectories are unique with three or more location tuples. Thus, an adversary with a dataset of the entire population and CI consisting of four or more location tuples would likely be able to perform one-to-one linkage for over 95% of the individuals in $\mathbf{Y}^{\ell,2}$. This implies that preventing Linkability when Singling-out is not prevented using Location Coarsening or Aggregating to Counts is very difficult unless the organization uses a value of $d \leq 1$ in both high and low density regions.

We suggest organizations attempting to convert personal data to non-personal data using Location Coarsening first prevent Singling-out; otherwise consider different approaches such as noise infusion in order to release a less linkable dataset of protected latitude and longitude coordinates.

5.5.2 Aggregating to Counts

Using the same simulation described in Section 5.4.2, we calculate the inverse of the equivalence class size (count value) k for each region with a count of at least two for all simulated protected datasets. Figure 5 plots the distribution of $1/k$ across all simulations for each value of d.

The distributions for Aggregating to Counts give a better picture of privacy than Location Coarsening since the majority of counts are larger than two. In both the full South Korea

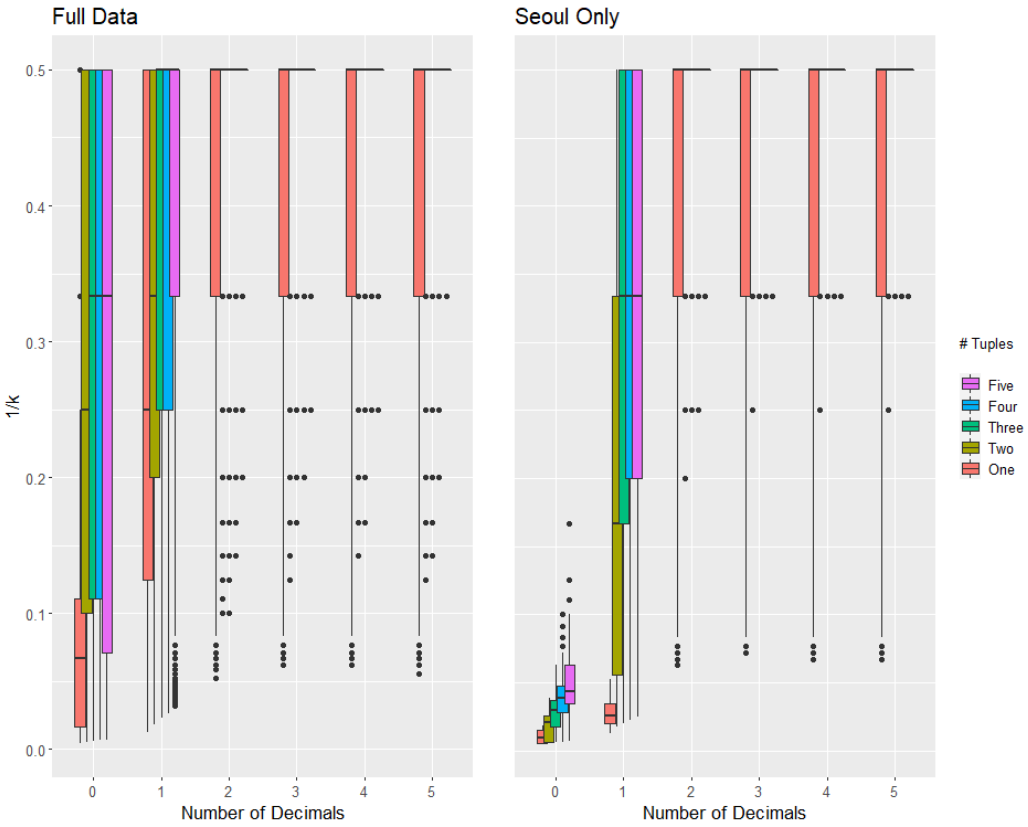

Figure 4: Distribution of $1/k$ for each value of $d = 0, 1, \ldots, 5$ and sampled trajectory lengths from one to five tuples in South Korea (left) and Seoul (right).

and the Seoul data, we see that most trajectories are in equivalence classes of at least three for any value of d. The Seoul data exhibits much larger equivalence classes for $d = \{0, 1\}$ than the full South Korea data, which represents the best privacy. For the full data, $d = 1$ provides a reasonable balance between the size of equivalence classes and the size of the geographical regions. For the Seoul data, a higher value of $d = 2$ provides counts for smaller regions with similar count magnitudes.

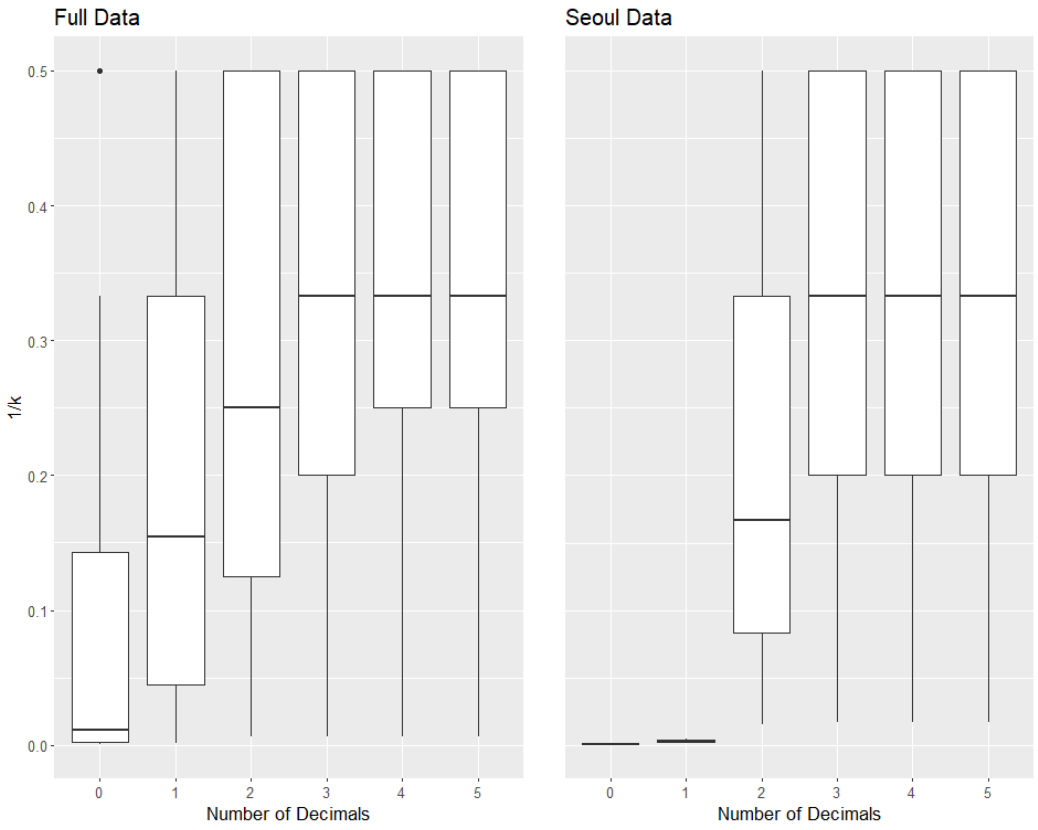

Figure 5: Distribution of $1/k$ for each value of $d = 0, 1, \ldots, 5$ in South Korea (left) and Seoul (right).

When Singling-out is not prevented, there exists some count equal to one. Table 9 shows that when $d = 2$, 27.02% and 18.82% of regions have a count of one in South Korea and Seoul, respectively. Accordingly, the adversary with an external linking dataset $\mathbf{X}^{\ell,d}$ of the entire population will have a one-to-one or many-to-one linkage in these areas. Practically, we suggest organizations attempting to prevent Linkability using Aggregating to Counts should carefully design their geographical regions to ensure that essential regions (which cannot be deleted) have count of at least two. At a reasonable level of aggregation, $d = 2$, deletion is required for approximately 20% of the counts.

5.6 Preventing Inference

There is a great deal of sensitive information that could be deduced from location data. For example, [38] find that in the absence of data obfuscation, individuals' location trajectories can be used to accurately predict individuals' home addresses within an average radius of 2.5 miles, and two randomly sampled locations are enough to fully identify 49% of individuals' entire location trajectories. We assume that the most likely Inference of interest in our dataset is the probability that an individual has COVID-19.

To setup this scenario and study the probability of a successful attribute linkage attack, we randomly assign 1% of individuals to have COVID-19 in our original dataset $\mathbf{Y}^{\ell,5}$. Define, $s_m = \{0, 1\}$ as the COVID-19 status of individual m with the prior probabilities, $p(1|\,\mathbf{X}) = 0.01$ and $p(0|\,\mathbf{X}) = 0.99$. The updated probabilities when the adversary has the protected dataset are $p(1\,|\,\mathbf{x},\,\mathbf{Y}^{\ell,d})$ for Location Coarsening and $p(1\,|\,\mathbf{x},\,\mathbf{C}^{d})$ for Aggregating to Counts.

5.6.1 Location Coarsening

To assess whether Location Coarsening prevents Inference, we assume the adversary matches the larger population dataset $\mathbf{X}$ to the location trajectories released in $\mathbf{Y}^{\ell,d}$. 100 simulated datasets are generated for each value of d and number of external location tuples (one to five). For each simulation, one percent of individuals are randomly assigned a positive COVID-19 status. Then, for each positive individual, we randomly sample the location tuples (without replacement) once and these tuples are treated as CI. We then calculate the number of times these sampled trajectories in $\mathbf{x}$ match individuals' trajectories in $\mathbf{Y}^{\ell,d}$ and take the mean of the matching COVID-19 statuses to produce the updated probability, $p(1\,|\,\mathbf{x},\,\mathbf{Y}^{\ell,d})$ for each positive individual. The left-hand side of equation (2) follows and we average across all COVID-19 positive individuals in each simulation. This process is performed once using the full data and once using the Seoul data.

Figure 6 shows boxplots of the average of the left-hand side of equation (2) for each simulation. A boxplot is shown for each value of d and one to five external information tuples in South Korea and Seoul, respectively. The figure shows that preventing Inference is difficult in low density regions of South Korea. Locations must be coarsened to $d = 0$ for half of the simulations to have less than a 5% increase in probability in (2). Preventing Inference is more likely in high density regions like Seoul for $d \leq 1$ with most simulations having less than a 5% increase in probability. However, there are still a couple of simulations in the city of Seoul that have over an average 25% increase in probability with three or more tuples at $d = 1$, which may be a violation of the law. The reason is that some location tuples are unique to simulated COVID-19 positive individuals and are shared by very few other individuals in $\mathbf{Y}^{\ell,d}$. To prevent Inference for Location Coarsening, we suggest suppressing location tuples that are nearly unique to one individual to reduce the ability of an adversary to deduce unknown information about these individuals. Overall, in a low density area, $d = 0$ is required to prevent inference with a small amount of location tuples or trajectories requiring deletion. In a higher density region such as Seoul, $d = 1$ provides a similar trade-off, with higher utility than $d = 0$ in a low density region.

5.6.2 Aggregating to Counts

For Aggregating to Counts, we perform 100 simulations for each value of d and consider all tuples for each randomly assigned COVID-19 positive individual. In each simulation, we calculate the updated probability $p(1\,|\,\mathbf{x},\,\mathbf{C}^{d}) = \frac{\mathbf{p^d}}{\mathbf{c^d}}$ for each tuple corresponding to

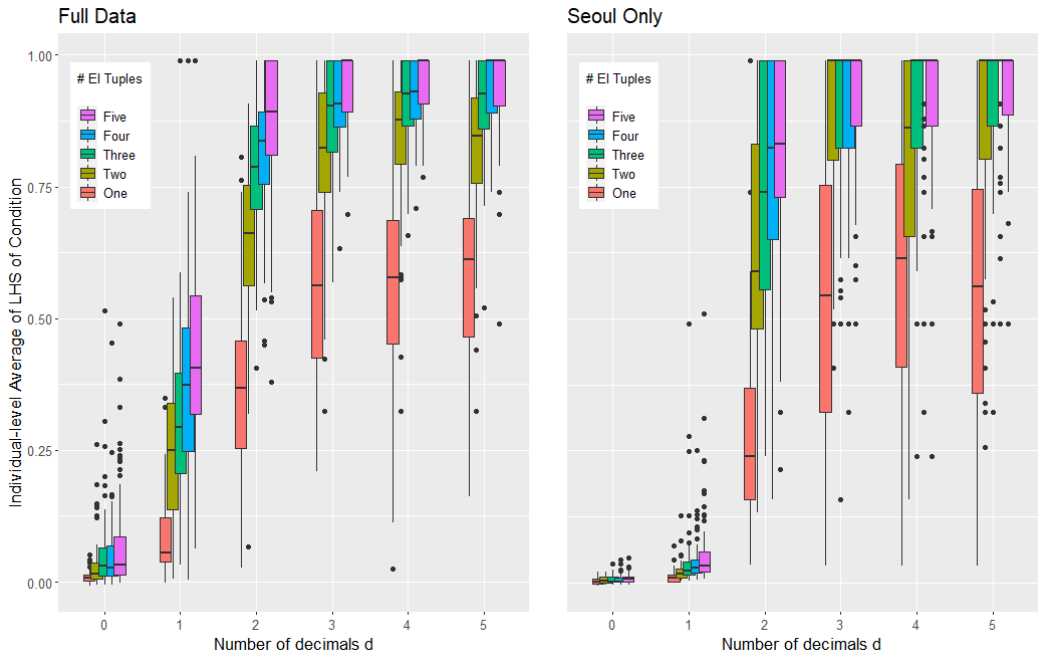

Figure 6: Location Coarsening: Boxplots of the increases in Inference probabilities for COVID-19 positive individuals

individual m where $\mathbf{c^d}$ is the count of individuals and $\mathbf{p^d}$ is the count of COVID-19 positive individuals corresponding to that individual's tuple (region). We then subtract the prior probability $p(1 \mid \mathbf{X}) = 0.01$ to compute the LHS of (2), and average across all tuples for individual m.

Figure 7 shows boxplots of these individual averages for all simulations. Since Aggregating to Counts removes the longitudinal trajectories, the results are generally better for preventing Inference. In Seoul, most of the individuals with COVID-19 are limited to an increase of 25% for $d = 2$ and every individual is limited to a 8% increase in Inference with $d \leq 1$. For Aggregating to Counts, preventing Inference is feasible for $c = .08$ (*i.e.*, with a prior probability of 1% and an updated probability of 9%) in relatively large ($d \leq 1$) and highly dense (Seoul) regions. Although this change may seem reasonable, we should note that the odds ratio is still quite large,

$$\frac{p(1 \mid \mathbf{x}, \mathbf{C}^d)}{p(0 \mid \mathbf{x}, \mathbf{C}^d)} \Big/ \frac{p(1 \mid \mathbf{X})}{p(0 \mid \mathbf{X})} = \frac{.09}{.91} \Big/ \frac{.01}{.99} = 9.8 \tag{3}$$

for the individual with the worst-case privacy. For a low density region, $d = 0$ provides the best balance between privacy and utility, and may still allow for significant increases in inference probability. High density regions present a more optimistic picture, as $d = 1$ renders small increases in inference probability with higher utility than $d = 0$ in a low density region.

Aggregating to Counts is unlikely to bound the odds ratio for all individuals at high levels

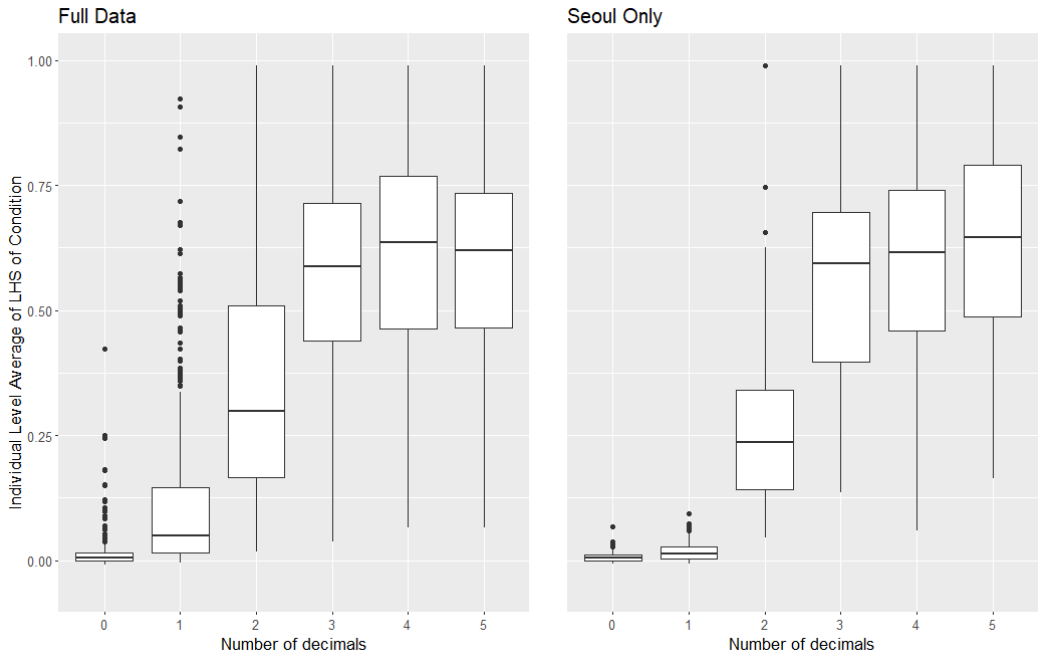

Figure 7: Aggregated Counts: Distributions of individual-level averages of left-hand side of equation (2) across 100 simulations for $d = 0, 1, \ldots, 5$ for Full Data (left) and Seoul (right).

of usefulness ($d \geq 3$). Thus, it will be challenging for organizations demanding a high level of usefulness to prevent Inference for all individuals in even the simplest of cases, such as using a binary variable that indicates disease status. However, it may be more reasonable to consider the average effects across all individuals, which would provide a more optimistic picture of privacy than shown in Figure 7.

5.7 Summary of Results

This application illustrates how our framework can be used to assess whether an anonymization method produces legally anonymous location data. For location coarsening, a low value of $d = 0$ is necessary to anonymize location data for low density regions without significant data deletion. However, this has a large negative impact on data utility. Data from high density regions can be anonymized for a slightly higher value of $d = 1$ without significant data deletion, and has higher utility. Aggregating to counts can produce legally anonymized data at a higher value of $d = 2$ in high density regions, but will likely require a significant amount of data deletion to limit Inference to a legally acceptable level.

Overall, the simple methods which we assessed had difficulty preventing Singling-out, Linkability, and Inference, and an assessment of more sophisticated techniques for anonymizing location trajectories should be a topic of future work. Using Location Coarsening, we found that individual-level location trajectories are not very private and require a large degree of coarsening to meet legal criteria. Using Aggregating to Counts, we found that count data are more likely to meet legal criteria, although this data does not provide the

same level of granularity as trajectories. We also showed the importance of considering contextual elements in the anonymization process - relative to low density areas, high density anonymous location data can preserve a higher level of utility while meeting legal anonymization requirements.

6 Discussion

We proposed a framework of legal anonymization criteria to evaluate the conversion of personal location data to non-personal location data under the GDPR. If protected data met all of the criteria in our framework, it was no longer considered personal location data nor under the scope of the data protection law. If not, then organizations using personal data are required to invest in additional privacy and security measures (and bear the risks of that personal data) or cease using the personal data. We tested our framework on two logical and relatively simple statistical anonymization solutions which converted personal location data to non-personal location data, but came at different costs to the usefulness of the underlying data. We found that location tuples that correspond to very few individuals represent high privacy risk and increase the difficulty of legal anonymization. Deleting these location tuples is necessary to achieve legal anonymization and preserve overall data utility. We found that context-specific factors, such as population density, lead to differences in the privacy and utility of the anonymized data. Overall, there needs to be a tighter marriage between legal definitions and usage, and mathematical definitions and usage.

The EDPB has also provided commentary [22] on the ability of various anonymization techniques, broadly categorized as either randomization or generalization, to protect against re-identification. The EDPB concluded that none of the techniques considered on their own are guaranteed to anonymize data based on Singling-out, Linkability, and Inference. While based on past research, the EDPB's opinions were not specific to geolocation data and they ultimately concluded that the optimal solution for anonymizing a dataset should be decided on a case-by-case basis, as demonstrated in this paper. Furthermore, the absence of legislature or case law defining a "significant probability" or a "reasonability test" leaves practitioners guessing and was a motivation for testing the framework on our application in this paper. Ultimately, any method used to convert personal data to non-personal data must meet legal criteria, but there is no guarantee that meeting these criteria provides adequate overall privacy [55] (*e.g.*, 2-anonymity may comply with the law but provide little privacy). Further, while our framework is intended to encapsulate anonymization requirements from multiple privacy laws, no one solution may fit all legal requirements impacting a data set. This is a challenge that requires a joint legislative and mathematical solution, and we encourage further research combining privacy law and disclosure limitation. An important next step will be to determine whether more complex anonymization methods for location data are capable of meeting legal anonymization requirements, and whether they provide a better privacy-utility tradeoff than the methods examined in this paper. One promising area is synthetic data models, although more research is needed to fully understand the privacy offered by synthetic data and whether it meets legal anonymity requirements [59].

We also note that the anonymization techniques we used for location data are not likely to meet the three criteria using other data sources such as retail point-of-sale transactions [50], time series data [42], textual data stored from chatbots, search history logs [53], facial images [61, 2], social network data [27], or even the related contact tracing data [21] derived from geolocation data. Large organizations or services (*e.g.*, Google, Amazon Web Services,

etc.) may eventually invest in and sell technological solutions that convert contextually-dependent personal data to non-personal data; however, the costs are likely prohibitive for small- to medium-sized organizations. In the short term, these smaller organizations will either use straight-forward statistical solutions (as shown in our application) that may only satisfy certain regulatory definitions or incur the regulatory costs and corresponding risks to continue using personal data.[10] Even with additional security and legal measures designed to protect personal data, personal data still gets out.[11] With the increasing emphasis on data minimization and the heightened privacy risks of certain personal data, it is arguable if retaining data in its identifiable form even complies with the laws themselves.[12] This leads to the question of whether it is more important to comply with the law or attempt to meet the privacy criteria set forth in this paper by permanently altering the personal data.

This paper is a demonstration of the value of the bringing together multidisciplinary stakeholders together to develop privacy-promoting methods to address personal and non-personal data. And, we hope for both and encourage lawyers and statisticians to sit at the table together to continue to engineer privacy-oriented solutions.

References

[1] Chris Achatz and Susan Hubbard. "US vs. EU Guidelines for De-Identification, Anonymization, and Pseudonymization". In: *J. INTERNET L.* 20 (2017), p. 11.

[2] Alessandro Acquisti, Ralph Gross, and Frederic D Stutzman. "Face recognition and privacy in the age of augmented reality". In: *Journal of Privacy and Confidentiality* 6.2 (2014), p. 1.

[3] Charu C Aggarwal. "On k-anonymity and the curse of dimensionality". In: *VLDB*. Vol. 5. 2005, pp. 901–909.

[4] Luk Arbuckle and Khaled El Emam. *Building an anonymization pipeline: Creating safe data*. O'Reilly Media, 2020.

[5] Luk Arbuckle and Muhammad Oneeb Rehman Mian. "Engineering risk-based anonymisation solutions for complex data environments". In: *Journal of Data Protection & Privacy* 3.3 (2020), pp. 334–343.

[6] Luk Arbuckle and Felix Ritchie. "The five safes of risk-based anonymization". In: *IEEE Security & Privacy* 17.5 (2019), pp. 84–89.

[7] Article 29 Data Protection Working Party. *Opinion 05/2014 on Anonymisation Techniques*. 2014. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[8] Jane Bambauer, Krishnamurty Muralidhar, and Rathindra Sarathy. "Fool's gold: an illustrated critique of differential privacy". In: *Vand. J. Ent. & Tech. L.* 16 (2013), p. 701.

[10] A study commissioned by the Attorney General's Office of California's Department of Justice stated that small-sized firms with less than 20 employees would incur $50,000 in initial costs for CCPA compliance and medium-sized firms with 20-100 employees would incur $100,000 in initial costs for CCPA compliance based on data from a TrustArc survey [45].

[11] A report by Verizon found that 34% of the 2,013 data breaches in 2019 involved internal actors such as employees [58].

[12] The GDPR and the recently approved California ballot initiative, Proposition 24, the California Privacy Rights Act, both emphasize the implementation of data minimization procedures to limit overall personal data collection.

[9] Claudio Bettini, X Sean Wang, and Sushil Jajodia. "Protecting privacy against location-based personal identification". In: *Workshop on Secure Data Management*. Springer. 2005, pp. 185–199.

[10] Brazil. *Lei Geral de Proteçao de Dados (LGPD)*. 2018. URL: https://www.lgpdbrasil.com.br/wp-content/uploads/2019/06/LGPD-english-version.pdf.

[11] California State Legislature. *California Consumer Privacy Act of 2018*. 2018. URL: https://oag.ca.gov/privacy/ccpa.

[12] Aloni Cohen and Kobbi Nissim. "Towards Formalizing the GDPR's Notion of Singling Out". In: *Proceedings of the National Academy of Sciences* 117 (2020), pp. 8344–8352. DOI: 10.1073/pnas.1914598117.

[13] Giulio Coraggio and Giulia Zappaterra. "The risk-based approach to privacy: Risk or protection for business?" In: *Journal of Data Protection & Privacy* 1.4 (2018), pp. 339–344.

[14] Court of Justice of the European Union. *Judgment in Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland*. 2016. URL: http://curia.europa.eu/juris/documents.jsf?num=C-582/14.

[15] Richard Dosselmann, Mehdi Sadeqi, and Howard J. Hamilton. *A Tutorial on Computing t-Closeness*. 2019. arXiv: 1911.11212 [cs.CR].

[16] Marie Douriez et al. "Anonymizing nyc taxi data: Does it matter?" In: *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE. 2016, pp. 140–148.

[17] Alex Dow et al. *Protecting privacy in Facebook mobility data during the COVID-19 response*. 2020. URL: https://research.facebook.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/.

[18] DS4C. *Data Science for COVID-19 (DS4C)*. 2020. URL: https://www.kaggle.com/kimjihoo/coronavirusdataset.

[19] Cynthia Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 978-3-540-35908-1.

[20] Cynthia Dwork et al. "Exposed! A Survey of Attacks on Private Data". In: *Annual Review of Statistics and Its Application* 4.1 (2017), pp. 61–84. DOI: 10.1146/annurev-statistics-060116-054123.

[21] Cynthia Dwork et al. "On Privacy in the Age of COVID-19". In: *Journal of Privacy and Confidentiality* 10.2 (2020). DOI: 10.29012/jpc.749.

[22] EDPB. *Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak*. 2020. URL: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf.

[23] Khaled El Emam and Luk Arbuckle. *Anonymizing health data: case studies and methods to get you started*. " O'Reilly Media, Inc.", 2013.

[24] European Medicines Agency. *External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use*. 2018. URL: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data.

[25] European Parliament and Council of European Union. *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. 1995. URL: https://eur-lex.europa.eu/eli/dir/1995/46/oj.

[26] European Parliament and Council of European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[27] Stephen E Fienberg. "Is the Privacy of Network Data an Oxymoron?" In: *Journal of Privacy and Confidentiality* 4.2 (2013).

[28] Marco Fiore et al. "Privacy in trajectory micro-data publishing: a survey". In: *Transactions on Data Privacy* 13 (2020), pp. 91–149.

[29] Paul Francis et al. "Extended Diffix". In: *arXiv preprint arXiv:1806.02075* (2018).

[30] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. "Composition attacks and auxiliary information in data privacy". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, pp. 265–273.

[31] Simson L Garfinkel. "De-identification of personal information". In: *National institute of standards and technology* (2015).

[32] Anindya Ghose, Beibei Li, and Siyuan Liu. "Mobile targeting using customer trajectory patterns". In: *Management Science* 65.11 (2019), pp. 5027–5049.

[33] Mehmet Emre Gursoy et al. "Differentially private and utility preserving publication of trajectory data". In: *IEEE Transactions on Mobile Computing* 18.10 (2018), pp. 2315–2329.

[34] Mike Hintze and Khaled El Emam. "Comparing the benefits of pseudonymisation and anonymisation under the GDPR". In: *Journal of Data Protection & Privacy* 2.2 (2018), pp. 145–158.

[35] Japan. *Japan's Act on the Protection of Personal Information*. 2016. URL: www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf.

[36] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". In: *2007 IEEE 23rd International Conference on Data Engineering* (2007). DOI: 10.1109/ICDE.2007.367856.

[37] Taylor Lively. *US State Privacy Legislation Tracker*. 2022. URL: https://iapp.org/resources/article/us-state-privacy-legislation-tracker/.

[38] Meghanath Macha et al. "Personalized Privacy Preservation in Consumer Mobile Trajectories". In: *Working paper, Carnegie Mellon* (2022). URL: https://ssrn.com/abstract=3628640.

[39] Anna Monreale et al. "Movement data anonymity through generalization." In: *Trans. Data Priv.* 3.2 (2010), pp. 91–121.

[40] Yves-Alexandre de Montjoye et al. "Unique in the crowd: the privacy bounds of human mobility". In: *Sci Rep* 3 (2013). DOI: https://doi.org/10.1038/srep01376.

[41] Mehmet Ercan Nergiz et al. "Towards Trajectory Anonymization: a Generalization-Based Approach". In: *Transactions on Data Privacy* 2(1) (1 2009), pp. 47–75.

[42] Jordi Nin and Vicenç Torra. "Towards the evaluation of time series protection methods". In: *Information Sciences* 179.11 (2009), pp. 1663–1677.

[43] NRF. "Contact tracing apps: A new world for data privacy". In: (2020). URL: https://www.nortonrosefulbright.com/en/knowledge/publications/d7a9a296/contact-tracing-apps-a-new-world-for-data-privacy.

[44] Paul Ohm. "Broken promises of privacy: Responding to the surprising failure of anonymization". In: *UCLA l. Rev.* 57 (2009), p. 1701.

[45] David Roland-Holst et al. "Standardized Regulatory Impact Assessment: California Consumer Privacy Act of 2018 Regulations 2019". In: *Berkeley Economic Advising and Research, LLC* (2020). URL: http://www.dof.ca.gov/Forecasting/Economics/Major_Regulations/Major_Regulations_Table/documents/CCPA_Regulations-SRIA-DOF.pdf.

[46] Luca Rossi, James Walker, and Mirco Musolesi. "Spatio-temporal techniques for user identification by means of GPS mobility data". In: *EPJ Data Science* 4.1 (2015), p. 11.

[47] Harichandan Roy, Murat Kantarcioglu, and Latanya Sweeney. "Practical differentially private modeling of human movement data". In: *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer. 2016, pp. 170–178.

[48] Steven Ruggles et al. "Differential privacy and census data: Implications for social and economic research". In: *AEA papers and proceedings*. Vol. 109. 2019, pp. 403–08.

[49] Michael L Rustad and Thomas H Koenig. "Towards a global data privacy standard". In: *Fla. L. Rev.* 71 (2019), p. 365.

[50] Matthew J Schneider et al. "A flexible method for protecting marketing data: An application to point-of-sale data". In: *Marketing Science* 37.1 (2018), pp. 153–171.

[51] Frederic D Stutzman, Ralph Gross, and Alessandro Acquisti. "Silent listeners: The evolution of privacy and disclosure on Facebook". In: *Journal of privacy and confidentiality* 4.2 (2013), p. 2.

[52] The World Bank. *Population, female (% of total population) - United States*. 2020. URL: https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS?locations=US.

[53] Vincent Toubiana and Helen Nissenbaum. "An analysis of google log retention policies". In: (2011).

[54] Samantha Tsang. "Here are the contact tracing apps being deployed around the world". In: (2020). URL: https://iapp.org/news/a/here-are-the-contact-tracing-apps-being-employed-around-the-world/.

[55] Z. Tu et al. "A New Privacy Breach: User Trajectory Recovery From aggregated Mobility Data". In: *IEEE/ACM Transactions on Networking* 26.3 (2018), pp. 1446–1459.

[56] Zhen Tu et al. "Beyond k-anonymity: protect your trajectory from semantic attack". In: *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE. 2017, pp. 1–9.

[57] Zhen Tu et al. "Protecting Trajectory From Semantic Attack Considering *k*-Anonymity, *l*-Diversity, and *t*-Closeness". In: *IEEE Transactions on Network and Service Management* 16.1 (2018), pp. 264–278.

[58] Verizon. *Verizon 2019 data breach investigations report (DBIR)*. 2019.

[59] Thomas Zerdick. *Is the future of privacy synthetic?* 2021. URL: `https://edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en`.

[60] Jun Zhang, Xiaokui Xiao, and Xing Xie. "Privtree: A differentially private algorithm for hierarchical decompositions". In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, pp. 155–170.

[61] Yinghui Zhou, Shasha Lu, and Min Ding. "Contour-as-Face Framework: A Method to Preserve Privacy and Perception". In: *Journal of Marketing Research* (2020), p. 0022243720920256.

Appendix

Here, we describe the legal analysis of the anonymization criteria in various legal frameworks. Our conclusion was that the three anonymization criteria described by the EDPB, namely, Singling Out, Linkability, and Inference, appear to encapsulate many of the criteria that various legal definitions consider when determining if data is non-personal. Exploration of other regional data protection regulations also provides little additional insight into the statistical interpretation of anonymization, so we consider the EDPB's definition of anonymization as exhaustive and was the focus of our paper.

Assessing Anonymization Criteria Across Legal Frameworks

The CCPA uses three terms relevant to determining if personal information is non-identifiable:

(1) "Aggregate consumer information";

(2) "Deidentified"; and

(3) "Pseudonymization".

The CCPA defines pseudonymization (§ 1798.140(r), [11]) as:

> "The processing of personal information in a manner that renders the personal information no longer attributable to a specific consumer without the use of additional information, provided that the additional information is kept separately and is subject to technical and organizational measures to ensure that the personal information is not attributed to an identified or identifiable consumer."

This definition is almost verbatim the definition found in the GDPR (Art. 4(5), [26]). The CCPA defines de-identification (§ 1798.140(h), [11]) as:

> "Information that cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer, provided that a business that uses deidentified information:

> (1) Has implemented technical safeguards that prohibit reidentification of the consumer to whom the information may pertain.
>
> (2) Has implemented business processes that specifically prohibit reidentification of the information.
>
> (3) Has implemented business processes to prevent inadvertent release of deidentified information.
>
> (4) Makes no attempt to reidentify the information."

And, finally, the CCPA recognizes the idea of aggregation as a method to decrease the identifiability of personal information, defining "aggregate consumer information" (§ 1798.140(a), [11]) as:

> "Information that relates to a group or category of consumers, from which individual consumer identities have been removed, that is not linked or reasonably linkable to any consumer or household, including via a device. "Aggregate consumer information" does not mean one or more individual consumer records that have been deidentified."

In essence, under the CCPA, there is a legal distinction between the terms pseudonymization and de-identification, whereas under the GDPR, there is a legal distinction between the terms pseudonymization and anonymization. The terms "anonymized data" and "deidentified data" both refer to protected data and are sometimes used interchangeably. These terms do have shared characteristics. For example, both the EDPB's criteria for the term anonymization and the CCPA's definition of the term deidentification focus on the concepts of Linkability of data to an individual and Inference. Further, the EDPB's criteria for anonymization and the CCPA's definition of the term aggregate consumer information include the concept of singling-out. However, these terms ultimately have different meanings under the CCPA and the GDPR. These differences can result in varying practical and quantitative implications for privacy as well as corresponding privacy impacts that are not clearly outlined, or understood, within the legal community. Arguably, both regulations have the same intent: to recognize that there are technical means to remove, at least substantially, identifiers from personal data such that it is no longer attributable to an individual.

Looking beyond the EU and the CCPA, there are additional definitions of anonymization that can play a role in generating a mathematical equivalent to the legal definition of anonymization. For example, Japan's Act on the Protection of Personal Information defines anonymization (Ch. 1, 9(i), [35]) as:

> "Information relating to an individual that can be produced from processing personal information so as neither to be able to identify a specific individual by taking action prescribed in each following item in accordance with the devisions [sic] of personal information set forth in each said item nor to be able to restore the personal information. (i) personal information falling under paragraph (1), idem (i) ; Deleting a part of descriptions etc. contained in the said personal information (including replacing the said part of descriptions etc. with other descriptions etc. using a method with no regularity that can restore the said part of descriptions etc.) (ii) personal information falling under paragraph (1), idem (ii) ; Deleting all individual identification codes contained in the said personal information (including replacing the said individual identification codes with

> other descriptions etc. using a method with no regularity that can restore the said personal identification codes)."

Japan's definition focuses heavily on the concept of deletion of data, which presumes some permanent method of removing the identifiers from personal data, unlike prior definitions that focus on the unlinkability of the data.

Further, there are standards, which are not laws or regulations, that also use the term anonymization. For example, the National Institute of Standards and Technology ("NIST") defines this term [31] as:

> "The process that removes the association between the identifying dataset and the data subject."

NIST's definition focuses on the association between datasets and individuals, but provides little insight into the criteria to be taken into consideration when creating anonymization techniques.

From a broad review of these varying definitions, it appears that the EDPB's definition of anonymization covers a wider variety of mechanisms to convert personal data to non-personal data, including the removal of identifying information, aggregation, and other statistical approaches such as noise infusion.