TRANSACTIONS ON DATA PRIVACY 17 (2024) 1–30

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, jordan@jordanfischer.me, mjs624@drexel.edu, spw26@drexel.edu, sc3887@drexel.edu

Received 20 June 2022; received in revised form 3 May 2023; accepted 19 September 2023

Abstract. In the last decade, different countries adopted data protection legislation to govern the collection and processing of personal data. Most of these legislative frameworks recognize that data can either be personal or non-personal. However, there is a lack of definitive criteria for when personal data has become non-personal data, as well as an understanding of the consequences of applying such criteria to the usefulness of personal data. This uncertainty creates confusion as to whether organizations can comply with privacy laws while retaining the usefulness of personal data. To address this problem, we use the existing data privacy literature to provide reasonable interpretations of legal anonymization criteria for location data under the GDPR. We apply these criteria to two reasonable anonymization solutions that produce protected person-level data. Using location data of COVID-19 patients in South Korea, we find that these solutions can produce legally anonymous location data or useful data, but not both. Further, we highlight examples of developing more sophisticated data protection solutions to better balance the tradeoff between privacy and usefulness for contextual data sets.

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 [22]. Since then, many regions 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 [8]; Japan's Act on the Protection of Personal Information, Act No 57 of 2003, as amended in 2016 [32]; see also [43], pp. 431–48, discussing the impact of the GDPR across numerous regions on the world, and [33].

Generally, under each privacy law exists a concept of personal data or personally identifiable information. Personal data encompasses data that directly identifies the individual, or direct identifiers (*e.g.*, name, address, social security number), and data that indirectly identifies the individual, or indirect identifiers (*e.g.*, gender, date of birth, physical characteristics) (GDPR Art. 4(1) [22]; California Consumer Privacy Act (CCPA), 1798.140(o)(1) [9]). These laws recognize that data exists on a spectrum, ranging from data that is clearly personal, to data that is reasonably likely to identify an individual, to data that is non-personal, *i.e.*, non-identifiable to the individual who is related to the data collected.[2] The distinction between personal and non-personal data ultimately depends on reasonableness (or proportionality): whether the identifiers provided are reasonably likely to identify an individual (*e.g.*, GDPR Recital 26 [22]). These identifiers can be considered personal data on their own or in combination with other identifiers, and often relate to inferences that can be made from the data.

Privacy laws also recognize that techniques exist to convert personal data to non-personal data. 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. We focus on the latter approach and assume that personal data will eventually be breached; preemptively converting personal data to non-personal data would reduce the damage from such a breach. The question is whether the conversion process maintains the usefulness of the data.

There are substantial benefits from converting personal data to non-personal data, also known as anonymized data. For example, none of the requirements within the GDPR apply to anonymized data (GDPR, Recital 26 [22] and [3, 31]). Anonymized data can be stored indefinitely, does not require user consent to processing, and 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 various definitions to distinguish personal from non-personal data, creating confusion and potential conflict between legal frameworks. A growing challenge exists within the backdrop of abstract legal definitions of personal data: the lack of definitive criteria for when personal data has become non-personal, and an understanding of the practical consequences of applying such criteria. This creates 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 ([12]), the European Court of Justice (ECJ) held that Internet Protocol ("IP") addresses, in certain circumstances, are considered personal data.[3] The ECJ determined that 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

[2] See GDPR, Recital 26 [22]; see also the California Consumer Privacy Act of 2018 ("CCPA"), § 1798.140(O(3) [9]).

[3] 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, [21]). However, its ruling still applies under the GDPR.

computer."¶38. However, it recognized that for 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. The ECJ recognized that the website owner alone did not maintain personal data via access to the dynamic IP address, but since the ISP could be legally required to turn over the information necessary to identify an individual based on her IP address, then the IP address constituted personal data for both parties (¶49).

This case, and subsequent interpretations of the boundaries of personal data, highlight the challenges in distinguishing personal data from non-personal data. Overcoming these challenges requires privacy tools that help data controllers reason through threats their data will be exposed to, what level of risk these threats pose to data subjects, and what data transformation is necessary to mitigate privacy risk and ensure that 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. This could result in anonymized data sets with widely varying privacy risks.

To help address these challenges, we use the existing data privacy literature to provide reasonable interpretations of legal criteria to evaluate the conversion of personal data to non-personal data under the GDPR. We interpret the criteria in the context of trajectory micro-data, hereafter referred to as location data, to complement the many anonymization techniques for this type of data [24, 35]. Two distinguishing features make the anonymization of location data especially challenging: (1) granularity, where a single location could identify an individual, and (2) longitudinal trajectories, where just two longitudinally linked granular locations are enough to uniquely identify 90% of individuals in our empirical application. This contrasts with standard data sets where three or four non-granular variables (*e.g.,* zip-code, gender, date of birth or age, ethnicity) are needed to identify most individuals [51, 45].

We evaluate the interpretations with two reasonable anonymization solutions applied to a location data set from COVID-19 patients in South Korea. To the best of our knowledge, this is the first attempt to interpret and apply legal anonymization criteria to location data. We assess whether the anonymized location data has low privacy risk and falls outside of the scope of the GDPR. Notably, there is an absence of case law showing when data meet legal anonymization criteria. We also examine how differences in the anonymized data affect the usefulness of the resulting data. We find that both solutions are viable for producing non-personal data but have deleterious effects on their usefulness. While practitioners will ultimately require case law to determine reasonable levels of anonymization, our work provides organizations with an example documenting a reasonable effort to anonymize location data [11].

This paper proceeds as follows. In Sections 2 and 3, we review the legal anonymization criteria and discuss how our work fits into the literature on protecting location data. In Section 4, we interpret the legal anonymization criteria based on the existing privacy literature. In Section 5, we use these interpretations to evaluate whether two anonymization solutions can legally anonymize the location trajectories from COVID-19 patients in South Korea while retaining the usefulness of personal data. 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 (GDPR, Art. 4(5) [22]), defining it 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."

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 [22]). 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) adopted additional guidance on anonymization. The April 2020 Guidelines (¶15-16, [18]) 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, [18]):

> (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 the EU's precursor to the GDPR known as the Article 29 Working Party Guidelines 05/2014 ("WP Guidelines" [6]). 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 [18], ¶¶ 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, [18]). Paul Ohm

[39] 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 [6] 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 [19]:

> "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 [6] 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.*, location data. 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 is beyond the scope of this paper. The EDPB's three criteria of anonymization in the context of the GDPR appear to encapsulate many criteria that various legal definitions consider when determining if data is non-personal[4]. Therefore, we consider the EDPB's definition

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

of anonymization criteria as exhaustive and focus on these criteria for the remainder of this paper.

3 Related Work

The privacy risks associated with location data are well documented (*e.g.*, [36, 15]). Two factors motivate our study of legally anonymizing location data: (1) Location data has many uses, from location based targeting [29] to the study of disease dynamics; and (2) it is difficult to anonymize due to the granularity and uniqueness of individuals' location trajectories [24]. Existing work has examined linkage, homogeneity or attribute linkage, and probabilistic attacks against location data, as well as the privacy principles that oppose these attacks, namely *indistinguishability* and *uninformativeness* [24, 35]. Privacy criteria, such as k-anonymity and differential privacy, have been applied to location data to quantify these privacy principles. What is lacking is a connection between the literature on anonymizing location data and the anonymization criteria dictated by privacy law, which we seek to address.

There has been relatively little work that interprets the EDPB's anonymization criteria. While [10] proposed a universal standard for Singling-out, they did not interpret Linkability or Inference, and did not incorporate the reasonability test or the context of the anonymization process. These authors assumed a worst case scenario where the data is subject to a computationally unbounded adversary with complete knowledge of the data generating distribution. This is unlikely to occur and protecting against such an adversary requires anonymization methods that can severely reduce data utility, *e.g.*, differential privacy [7, 42, 24], and likely goes beyond what would be determined as reasonable. Often, a data controller wishes to share personal data with processors who also have security controls for their data. Legal anonymization should incorporate context-specific factors such as the type of data being anonymized, the technical abilities of the data user, and the amount of time the anonymized data will be available, as recommended by the Working Party. The authors in [26] took such an approach and interpreted Singling-out, Linkability, and Inference in the context of protected queries submitted to a data base. We also take a context-specific approach, and interpret Singling-out, Linkability, and Inference in the context of location data.

Other examples of context based anonymization include the Five Safes [5, 4], which is a framework for reasonably assessing the safety of different portions of the anonymization process, and the European Medicines Agency (EMA) guidance on anonymizing clinical data [20]. The EMA stated that either Singling-out, Linkability, and Inference must be prevented, or the identification risks must be deemed acceptably low. Ultimately, a data set can be anonymized in different ways, each of which may be adequate depending on the risk of re-identification in the context in which anonymized data is disclosed. In any context, the EMA supported maximizing the utility of the data for scientific study with adequate privacy. Similarly, in our empirical application, we evaluate two anonymization solutions for location data and examine the level of protection that maximizes data utility and achieves legal anonymization.

4 Interpretation of Legal Anonymization Criteria

In this section, we interpret the three legal criteria used to evaluate the robustness of anonymization as described in the EDPB Guidelines [18]. These criteria define when Singling-out, Linkability, and Inference are prevented, and 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 consider a data set $\mathbf{Y}$ which has been pseudonymized through the removal of any direct identifiers. The data set $\mathbf{Y}$ contains N rows, each row assumed to correspond to a unique individual, labeled $[N] = \{1, \dots, N\}$ and indexed $n \in [N]$, and P columns, labeled $[P] = \{1, \dots, P\}$ and indexed $p \in [P]$. Note that the data set $\mathbf{Y}$ is a collection of rows; $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^\mathbf{T}$ where a row $\mathbf{y}_n = (y_{n,1}, \dots y_{n,P})$ and $y_{n,p}$ denotes the value in row n and column p.

4.1 Singling-out

Singling-out is "isolat[ing] some or all records which identify an individual in [a] data set," [10] 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 [26]. We assume each individual has one record in $\mathbf{Y}$. Therefore, Singling-out amounts to isolating an individual's record from the rest of the records in $\mathbf{Y}$. We assume isolation occurs if an adversary identifies a record $\mathbf{y}_n$ that differs from all other records on at least one value $y_{n,p} \in \mathbf{y}_n$.

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

$$\min \hat{\mathbf{z}} \geq 2. \tag{1}$$

Singling-out is prevented when the minimum number of occurrences of any row $\mathbf{y}_n$ in $\mathbf{Y}$ is at least two and no individual can be isolated. To illustrate, consider the data set $\mathbf{Y}$ in Table 1 with records containing location trajectories and two categorical attributes.

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

Table 1: data set $\mathbf{Y}$.

Every individual can be singled out based solely on one of the location attributes {`L1`, `L2`, `L3`}. Singling-out can be prevented by applying generalization or suppression to the attributes in $\mathbf{Y}$ to create subsets of records, or equivalence classes, that share the same attribute values [44]. We create a protected version of $\mathbf{Y}$ in Table 2 by truncating the decimal places in the latitude and longitude measurements [41] and suppressing non-matching `Sex` and `Disease` values in equivalence classes.

	L1	L2	L3	Sex	Disease
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	*	Hypertension
4	61.87, 122.38	61.24, 121.93	61.00, 121.82	*	Hypertension
5	58.84, 119.99	58.29, 120.35	59.55, 121.49	*	Dementia
6	58.84, 119.99	58.29, 120.35	59.55, 121.49	*	Dementia

Table 2: Generalized data set $\mathbf{Y}$ with Singling-out prevented.

This level of protection against Singling-out may be sufficient, depending on the context and objectives of the anonymization process. Larger equivalence classes must be created if better privay is required. However, for this data set 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 data sets 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. We note that preventing Singling-out alone does not provide adequate overall privacy, *e.g.*, attribute disclosure occurs since the value of `Sex` is disclosed for the first two rows, and the value of `Disease` is disclosed for the last four rows. We discuss the prevention of this type of disclosure in Section 4.4.

Whether an individual can be singled out is related to the concept of *unicity*, or the uniqueness of the location trajectories in a given data set. Significant work has gone toward developing methods for reducing the unicity of location data [24, 35] since individuals' location trajectories can be easily singled out [36]. To measure the risk associated with a unique location trajectory in $\mathbf{Y}$, one could consider the probability that a trajectory is also unique in the population [48]. This probability affects whether a match between a population trajectory and a unique trajectory in $\mathbf{Y}$ is correct, which we discuss in the next section and in our empirical application.

4.2 Linkability

We define Linkability to mean that a one-to-one linkage occurs between records for some individual in $\mathbf{Y}$ and an external data set $\mathbf{X}$, containing some or all of the same attributes in $\mathbf{Y}$ [26]. To be conservative, we assume it is known that the one-to-one linkage is correct. 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 data set $\mathbf{X}$ contains N' rows, each row identifying a unique

individual, and P' columns. To be conservative in terms of privacy, we assume that the external data set includes at least some data on all individuals in $\mathbf{Y}$ where $N \leq N'$, *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 $\mathbf{Y}$.

Define common information (CI) as the nonempty subset of columns $\mathcal{K} \subseteq [P]$ of size $K \equiv |\mathcal{K}|$ that is contained in both $\mathbf{Y}$ and $\mathbf{X}$.[5] For a row $\mathbf{y}_n$, let $\bar{\mathbf{y}}_n = (y_{n,k}, k \in [K])$ denote a CI tuple which is the truncation of $\mathbf{y}_n$ leaving only the CI. The matrix of unique CI tuples found in $\mathbf{Y}$ is 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 $\mu(\bar{\mathbf{y}}_j) = \#\{i \in [I] : \bar{\hat{\mathbf{y}}}_i = \bar{\mathbf{y}}_j\}$ 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 $\hat{\mathbf{Y}}$, *i.e.*, the P-vector $\hat{\mathbf{y}}_i$ is shortened to the K-vector $\bar{\hat{\mathbf{y}}}_i$. Define the vector of counts $\bar{\mathbf{z}} = (\bar{z}_1, \ldots, \bar{z}_J)$, where $\bar{z}_j = \mu(\bar{\mathbf{y}}_j)$ is the count for unique CI tuple $j \in [J]$ and $\sum_j z_j = I$. We focus on a direct one-to-one linkage attack, *i.e.*, an external record is linked to a record in $\mathbf{Y}$ when the CI in both records is identical. Linkability occurs when $\bar{z}_j = 1$ for some $j \in [J]$ and is prevented when

$$\min \bar{\mathbf{z}} \geq k, \forall k \geq 2. \tag{2}$$

The criteria in (2) is synonymous with *k*-anonymity [51] where one-to-one linkage is prevented when there are at least two records in $\mathbf{Y}$ corresponding to each CI tuple $\bar{\mathbf{y}}_j$ based on $\mathcal{K}$. This is not as strong of a requirement as the prevention of Singling-out and we discuss the difference in Section 4.3. To illustrate, we return to our example from Section 4.1. Suppose an adversary has access to an external data set $\mathbf{X}$. For simplicity, we show a subset of $\mathbf{X}$ in Table 3. The common information $\mathcal{K}$ are the three location attributes and sex, *i.e.*, {`L1`, `L2`, `L3`, `Sex`}. One-to-one linkage occurs between records two and six in $\mathbf{X}$ and their matches in $\mathbf{Y}$ using {`L1`, `L2`, `L3`, `Sex`} when no data protection is applied.[6]

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

Table 3: External data set $\mathbf{X}$.

[5] Common information is sometimes referred to as 'key' variables in the literature [47].

[6] In fact, one-to-one linkage can occur between records two and six and their matches using any one of the location variables.

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

Table 4: Data set $\mathbf{Y}$.

We apply generalization and suppression to the variables in $\mathcal{K}$ to produce the $\mathbf{Y}$ shown in Table 5.

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

Table 5: Data set $\mathbf{Y}$ with Linkability prevented.

Now, only a one-to-many linkage can occur between record two in $\mathbf{X}$ and records one and two in $\mathbf{Y}$, and record six in $\mathbf{X}$ and records five and six in $\mathbf{Y}$. If all records in $\mathbf{X}$ were examined, only many-to-many linkages could occur. 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 $1/k$ [6] (assuming that the external record only matches to one equivalence class). While k-anonymity is one of the most commonly applied protection approaches for location data [35], it has known shortcomings, such as failing to prevent composition attacks [27], and the Working Party provides several warnings on its use [6].

First, small values of k result in low privacy, but may be appropriate in contexts with high data security and other safety measures. The value of k should always be chosen to provide a reasonable balance between privacy and utility. Second, equivalence classes should not contain uneven distributions of attributes. This relates to the ability of an adversary to perform attribute disclosure, *i.e.*, deduce information pertaining to a target individual when the linked records in $\mathbf{Y}$ correspond to a single attribute value (or one attribute value with high probability) [24]. For example, if an adversary links a target record to the second equivalence class in Table 5, they learn the target individual has Hypertension. We examine attribute disclosure in creater detail in Section 4.4. Third, it is inappropriate to exclude any quasi-identifiers from the k-anonymity calculation. However, the line between quasi-identifiers and sensitive attributes (non-identifying variables, *e.g.*, `Disease` in Table 5) is blurred for location data, which we discuss more in Section 4.3.

We note that there are other approaches to record linkage such as principal component score rankings or probabilistic, distance, or cluster-based matching [13, 47, 59, 14]. We also assumed that an adversary knew a one-to-one linkage was correct. In practice, a unique record in a data set may not be unique in the population, especially if the size of the data set is small relative to the size of the population [48]. However, for location data, the probability that a unique sample record is a unique population record is likely very high: four

spatio-temporal points are enough to Single-out 95% of 1.5 million individuals at a relatively low spatial resolution [36].

4.3 Singling-out vs. Linkability

Depending on the external data set, some locations may be CI while others are sensitive such as the location of an individual's home or work [35]. This means $\bar{\mathbf{z}}$ will be difficult to determine since the organization may have no knowledge of which columns in $\mathbf{Y}$ are CI from the $2^N - 1$ possibilities. Since we assume the external data set contains a record for each individual in $\mathbf{Y}$, one-to-one linkage is prevented for all possible $\mathcal{K} \subseteq [P]$ when Singling-out is prevented, *i.e.*, $\min \hat{\mathbf{z}} \geq 2$. In this case, $\mathbf{y}_n$ is non-unique relative to $\mathbf{X}$. Therefore, any CI-truncation $\bar{\mathbf{y}}_j$ is also non-unique, which implies that $\min \bar{\mathbf{z}} \geq 2$, i.e., Linkability is prevented. Conservatively, we suggest organizations take this approach and prevent Singling-out such that no record from $\mathbf{Y}$ may be directly linked with a public record from $\mathbf{X}$ under any $\mathcal{K}$.

Preventing Linkability does not necessarily prevent Singling-Out. For example, rows one and two in Table 5 can be Singled-out based on `Disease`. However, Singling-out (and hence Linkability) are prevented in Table 2 since no row can be isolated and one-to-one linkage is prevented based on any CI. The distinction is that any record must match at least one other record on *all variables*, not just CI, for Singling-out to be prevented. This is a stricter requirement than traditional *k*-anonymity, and syntactic privacy methods may struggle to achieve (1) while maintaining data utility, especially in high dimensions [2]. However, most of the proposed protection methods for location data focus on preventing linkage attacks [24] so organizations can choose the method that yields the highest utility while meeting the legal criteria. We note that many-to-one and many-to-many linkages can still allow attribute disclosure to occur (*e.g.*, `Sex` for the first equivalence class and `Disease` for the second and third equivalence classes in Table 5, respectively) which we discuss in Section 4.4.

4.4 Inference

Inference is deducing, with significant probability, unknown information about an individual. 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, [6]). In the context of location data, unknown information could take many forms, such as personal attributes or home or work locations. 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" for some individual of interest, based on the information in $\mathbf{Y}$, any publicly available data $\mathbf{X}$, and background knowledge $\mathbf{b}$ [26]. We need a reasonable interpretation of Inference that allows us to limit the increase in probability that such statements are correctly made by accessing the anonymized data set.

We define Inference using an attribute disclosure attack against our data set $\mathbf{Y}$. An adversary seeks to identify the value of a sensitive categorical (or intervaled) variable for individual n using some external information on individual n, such as $\mathbf{x}$ (a row in $\mathbf{X}$ corresponding to individual n) and/or $\mathbf{b}$ (background knowledge about the individual or the sensitive variable of interest). Let S denote the discrete random variable and s denote the value of the random variable. Let s^* be the true value for individual n that the organization is trying to prevent Inference on. The adversary has a prior probability $p(s^* \mid \mathbf{X}, \mathbf{b})$ of the

individual having a value s^* based on external information and background knowledge. Once $\mathbf{Y}$ is released, the adversary updates the probability $p(s^* \mid \mathbf{x}, \mathbf{b}, \mathbf{Y})$ based on the new information. The pseudonymized data set $\mathbf{Y}$ prevents Inference for individual n when

$$p(s^* \mid \mathbf{x}, \mathbf{b}, \mathbf{Y}) - p(s^* \mid \mathbf{X}, \mathbf{b}) \leq c. \tag{3}$$

The change in probability on the left-hand side (LHS) is always between -1 and 1. Per the EDPB's guidelines, we view (3) 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 (3) is negative, the criterion is not violated as Inference is worse on individual n and the probability that the adversary infers the value s^* is lower than it was prior to accessing $\mathbf{Y}$. Exploring the ramifications for individual n if the adversary infers an incorrect value s is outside the scope of this paper. We suggest setting the LHS to 0 in cases where Inference is degraded, $p(s^* \mid \mathbf{x}, \mathbf{b}, \mathbf{Y}) < p(s^* \mid \mathbf{X}, \mathbf{b})$, and limiting the maximum increase in probability across all individuals in $\mathbf{Y}$.

We revisit our example from Sections 4.1 and 4.2 to illustrate (3). Suppose an adversary seeks to deduce the `Disease` of individual n in $\mathbf{Y}$ using the corresponding record in $\mathbf{X}$. The adversary bases their prior belief over the `Disease` of individual n using the following probability mass function (PMF)

$$p(s \mid \mathbf{X}, \mathbf{b}) = \begin{cases} 0.6, & s = \text{Hypertension}, \\ 0.2, & s = \text{Dementia}, \\ 0.2, & s = \text{Cancer}, \end{cases}$$

which could be based on prior beliefs $\mathbf{b}$, the population PMF in $\mathbf{X}$, or both. For simplicity, we assume the `Disease` categories are mutually exclusive and collectively exhaustive. In practice, we do not know the amount of external information available to an adversary, and the prior PMF should be chosen to be both reasonable and conservative to serve as a baseline for assessing whether Inference is prevented.

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.[7] In this example, $\mathbf{x}$ links to the first and second records in $\mathbf{Y}$ based on the CI.

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

Table 6: External data set $\mathbf{X}$.

[7] 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 $\mathbf{x}$.

	L1	L2	L3	Sex	Disease
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	*	Hypertension
4	61.87, 122.38	61.24, 121.93	61.00, 121.82	*	Hypertension
5	58.84, 119.99	58.29, 120.35	59.55, 121.49	*	Dementia
6	58.84, 119.99	58.29, 120.35	59.55, 121.49	*	Dementia

Table 7: data set $\mathbf{Y}$ with Singling-out and Linkability prevented.

This results in the following updated `Disease` PMF for the first two rows based on $\mathbf{x}$, $\mathbf{b}$, and $\mathbf{Y}$,

$$p(s \mid \mathbf{x}, \mathbf{b}, \mathbf{Y}) = \begin{cases} 0.42, & s = \text{Hypertension} \\ 0.29, & s = \text{Dementia} \\ 0.29, & s = \text{Cancer}, \end{cases}$$

which is conditioned on the values of `Disease` being different for the first two rows. Under this scenario, there is an increase of 0.09 in the probability of the adversary deducing that the `Disease` of individual n is "Cancer". Using (3),

$$p(s^* \mid \mathbf{x}, \mathbf{b}, \mathbf{Y}) - p(s^* \mid \mathbf{X}, \mathbf{b}) = 0.29 - 0.20 = 0.09,$$

which is a relatively low increase. However, for the individuals corresponding to records three through six in $\mathbf{Y}$, (3) would be violated since the adversary can deduce their `Disease` values with 100% certainty *without considering any external information*. Thus, the maximum of the LHS of the Inference condition is equal to

$$\max\{0.09, 0.09, 0.40, 0.40, 0.80, 0.80\} = 0.80.$$

Reducing the LHS for records three through six would require further generalization to create equivalence classes with non-homogenous `Disease` values (which is not possible through additional truncation), or suppressing all values of `Disease` which would severely degrade the utility of the data. The examples in this section have shown that k-anonymity does not prevent Singling-out and that extensions such as l-diversity and t-closeness would be needed to prevent Inference via attribute disclosure [6, 57, 55]. Overall, whether (3) is violated depends on c and the choice of the sensitive variable, which must be reasonable and most likely, even if it does not actually result in non-identified information. We note that other methods of evaluating attribute disclosure such as interval disclosure [52] could be used. Overall, very little work has examined attribute disclosure in location data [24] and we examine such an attack in our empirical application.

5 Application to South Korean COVID-19 Location Data

5.1 Data Description

Our data set was collected by the Korean Centers for Disease Control and Prevention (KCDC) and contains latitude/longitude coordinates for COVID-19 positive individuals in South Korea [16]. We build on the notation introduced in Section 4 and define a longitudinal location data set $\mathbf{Y}^\ell$. The superscript ℓ denotes that rows have longitudinal data. The

original data set $\mathbf{Y}^{\ell}$ is a collection of $N^{\ell} = 1,472$ location trajectories of COVID-19 positive patients tracked from January 20, 2020 to June 01, 2020, where $\mathbf{y}_n^{\ell} = (y_{n,1}, y_{n,2}, \ldots, y_{n,P_n})$ is the trajectory for individual n and $y_{n,p}$ denotes the p-th location tuple for individual n. There are P_n location tuples in the location trajectory for individual n, where P_n ranges from one to forty-five tuples for all individuals. We delete trajectories with $P_n < 5$, 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 data set are unique. Without further disclosure limitation, the prevention of Singling-out would require the deletion of nearly all trajectories.

5.2 Use Cases of Location Data

We assess reasonable anonymization solutions of $\mathbf{Y}^{\ell,d}$ for two use cases involving individual location trajectories and individual location tuples. The anonymization solutions intend to "...achieve the targeted anonymization while producing some useful data," [6]. We use the criteria from Section 4 to assess how the usefulness of the data changes at the point in which the legal criteria are met.[8]

The first use case stores the data for each individual in a location trajectory which is useful for tracking and mitigating the spread of disease, predicting consumers' path-to-purchase [50], providing accurate point-of-interest recommendations and context-aware marketing messages [34], and influencing taxi drivers' decision-making processes to improve driver incomes and market efficiency [61].

The second use case drops the longitudinal links and treats each location as an independent observation which is useful for disease mapping[9] and geo-targeting/conquesting [25]. This removes information on individuals' movements and we denote this non-longitudinal location data set as $\mathbf{Y}$ which has $N = \sum_{n=1}^{N^{\ell}} P_n$ observations.

Furthermore, the EDPB states that the reasonability test for location data must "take into account both objective aspects...and contextual elements...including population density, nature and volume of data..."[18]. Thus, 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) - shown in Figure 1. The number of locations observed across South Korea and Seoul are approximately 0.05 and 4 per square kilometer, respectively. In practice, additional contextual factors should be considered when balancing privacy and utility, such as the technical abilities of the data user, and the length of time the anonymized data will be available.

5.3 Anonymization Solutions for Location Data

For the first use case, we perform Location Coarsening by manually rounding the latitude and longitude coordinates in $\mathbf{Y}^{\ell}$ to a lower number of decimal places, denoted d, which is similar to truncation [41]. For example, there are many specific location tuples, *e.g.*, (37.5926645, 127.0174081), (37.5929072, 127.0169073), and (37.5928234, 127.017167) at $d = 7$, that correspond to the coarsened tuple (37.593, 127.017) at $d = 3$. As a result, the number

[8] We discuss additional use cases in the appendix.

[9] For example, see coronamap.site.

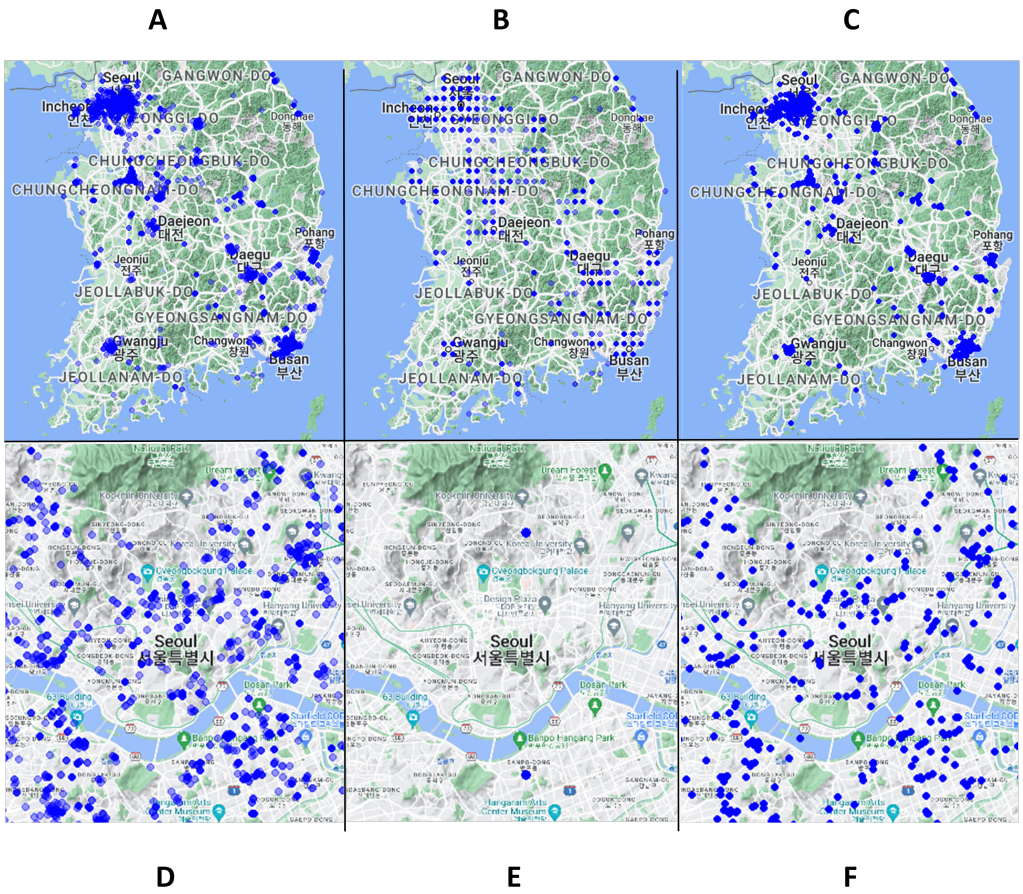

Figure 1: Location tuples mapped over South Korea (top row) and Seoul (bottom row). **A** and **D** contain original locations, **B** and **E** contain coarsened locations for $d = 1$, and **C** and **F** contain microaggregated locations for $k = 5$.

of coarsened location tuples when $d \leq 7$ is always at least as big as the number of location tuples when $d = 7$ for a given tuple. The same is also true for coarsened trajectories.

For the second use case, we use the sdcMicro package in R [52] to perform microaggregation to replace the values of location tuples in clusters of size k with the centroid of the cluster. For example, the clustered tuples (37.5926645, 127.0174081), (37.5929072, 127.0169073), and (37.5928234, 127.017167) for $k = 3$ would all be replaced with their centroid (37.5927984, 127.0171608). Figure 1 illustrates the coarsened location tuples when $d = 1$ and the microaggregated tuples for $k = 5$, respectively.

Both approaches alter the data with no randomness and produce location trajectories and tuples consistent with the original data and palatable across the organization.

5.4 Preventing Singling-out

To prevent Singling-out, no individual can have a unique location trajectory (for use case one) or location tuple (for use case two). One issue with evaluating the effects of Location Coarsening is that the original trajectories have different lengths P_n from 5 to 45. To address this issue, we simulate 500 protected data sets for each value of $d \in \{0, 1, 2, 3, 4, 5\}$.[10] These 500 simulated data sets are created by sampling (without replacement) one to five location tuples from each trajectory in the original data set. We calculate the percentage of trajectories that are unique in each simulated data set. For the second use case, we perform microaggregation for cluster sizes $k \in \{2, 5, 10, 15, 20, 25\}$. We apply both anonymization solutions to both the South Korea and Seoul data sets.

Each boxplot in Figure 2 shows the percent of unique trajectories for each of 100 simulations given a sampled trajectory length and value of d. One weakness of Location Coarsening is that it is possible to reduce the granularity of the locations but still have a unique trajectory. 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.[11] The advantage of microaggregation in this case is that it protects against Singling-out for any value of k since it replaces the locations in each cluster with the cluster centroid.

Table 8 reports quantiles of the distance shifted (in meters) for locations under each anonymization method, calculated using the geosphere package in R. For the first use case, the usefulness of the coarsened data is poor for low values of d. 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 South Korea when $d = 1$ where the median distance shifted is approximately 4 kilometers. The utility results for coarsened data are better in Seoul for $d = 1$, but worse for other values of d. The reason is that location coarsening places points on a grid determined by the decimal values. The distance shifted depends on how close the original locations are to that grid. The locations in Seoul are close to the grid points for $d = 1$ (see Figure 1), which results in low quantiles.

For the second use case with microaggregation, the distances shifted depend on the proximity (density) of the original locations. Hence, the utility of the Seoul data is significantly better at nearly every quantile and aggregation level. For location coarsening, $d = 1$ provides the best balance between privacy and utility since most of the data must be deleted

[10] We only consider decimal places $d \leq 5$ because there is little effect from coarsening to values of d larger than five.

[11] 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 [41] and would greatly increase the difficulty of preventing Singling-out.

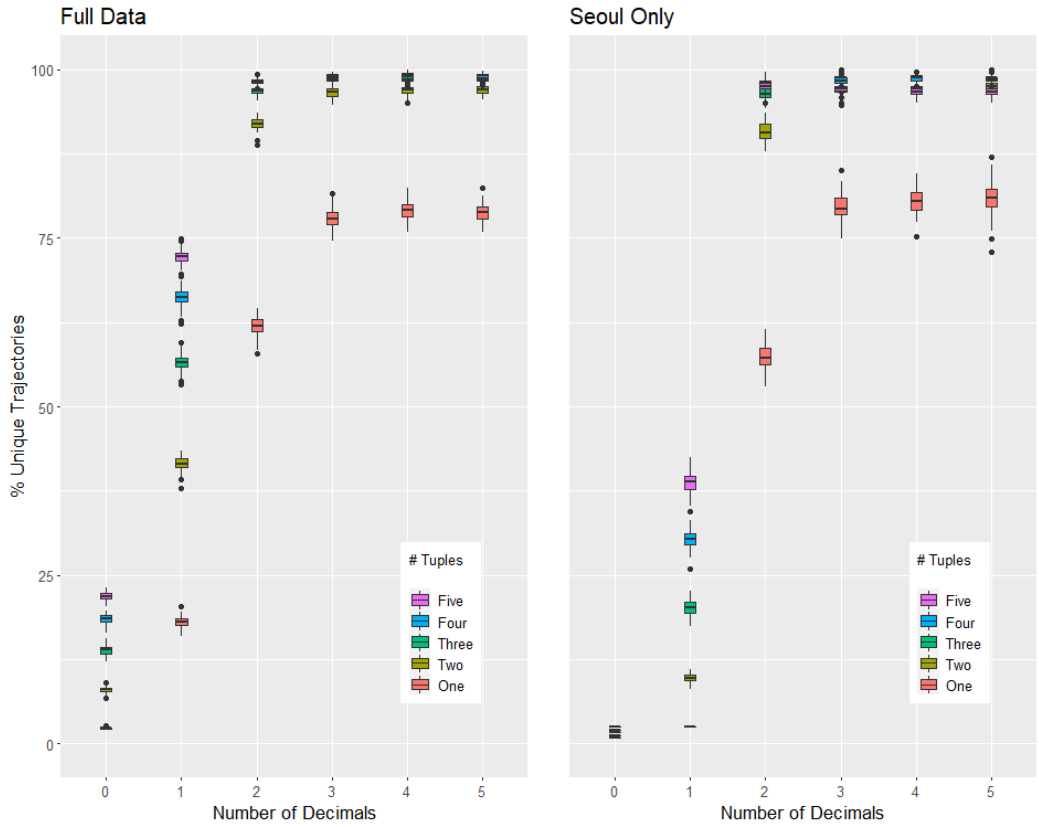

Figure 2: Percentage of unique trajectories for each value of $d = 0, 1, \ldots, 5$ and sampled trajectory lengths from one to five tuples in South Korea (left) and Seoul (right)

Anonymization Method		Full Data			Seoul Data		
	d	*2.5%*	*50%*	*97.5%*	*2.5%*	*50%*	*97.5%*
	5	0.08	0.39	0.65	0.08	0.39	0.63
	4	1.00	3.94	6.36	1.00	3.96	6.48
Location Coarsening	3	10.61	40.36	64.04	10.71	39.94	64.14
	2	82.62	394.35	642.58	123.86	413.19	645.01
	1	987.36	3,990.97	6,378.20	861.47	3730.77	6050.11
	0	15,223.54	47,203.55	63,859.24	40,978.59	51,748.75	55,882.45
	k	*2.5%*	*50%*	*97.5%*	*2.5%*	*50%*	*97.5%*
	2	0.00	0.00	3,455.16	0.00	0.00	576.80
	5	0.00	140.81	8,449.03	0.00	106.53	983.13
Microaggregation	10	0.00	352.33	13,752.10	0.00	246.27	1,382.91
	15	0.00	484.22	18,314.74	0.00	352.74	1,724.26
	20	0.00	680.27	18,846.68	0.00	482.35	1,818,04
	25	0.00	776.56	26,569.35	0.97	587.22	2,130.31

Table 8: Quantiles of distances shifted (in meters) between coarsened tuples and original tuples and microaggregated tuples and original tuples for the full and Seoul data sets.

for $d \geq 2$, and the median distance shifted when $d = 0$ is over 47 kilometers. Any choice of k for microaggregation prevents Singling-out, so organizations can choose the value that best balances utility and privacy in their specific context.

5.5 Preventing Linkability

In practice, an organization would meet the stronger Singling-out criterion to prevent Linkablity since both are required to convert to non-personal data. Even when Linkability is prevented, the size of equivalence classes is directly related to the probability that an external record and a matching record in an anonymized data set correspond to the same individual. Based on the same simulation described in Section 5.4, Figure 3 plots the distribution of the inverse equivalence class size $1/k$ for all *non-unique* trajectories across all simulations for each combination of d and sampled trajectory length, and for each value of k under microaggregation.

For use case one, most trajectories are in equivalence classes of size two for $d \geq 2$, the minimum size for legally anonymized data. This requirement may be an acceptable level of privacy in contexts with other privacy or security measures. However, privacy improves by moving to $d = 1$ to place more trajectories into larger equivalence classes. The dense Seoul data exhibits larger equivalence classes under location coarsening for $d = 1$ than the South Korea data, where the best privacy occurs when $d = 0$ and most trajectories are contained in equivalence classes of at least ten records. $d = 1$ represents a better tradeoff between utility and privacy for both coarsened data sets due to the massive reductions in data utility at $d = 0$ (note that unique trajectories must still be deleted to legally anonymize the data).

The results are better for use case two, where microaggregation produces equivalence class sizes bounded by our choice of k, and there are outlying clusters with larger numbers of location tuples. Between the two methods, comparing equivalence classes of at least ten locations shows that utility is orders of magnitude better under microaggregation than Location Coarsening, *e.g.*, a median shift of 246.27 meters for $k = 10$ in the microaggregated Seoul data vs. 3,730.77 meters for $d = 1$ in the coarsened Seoul data.

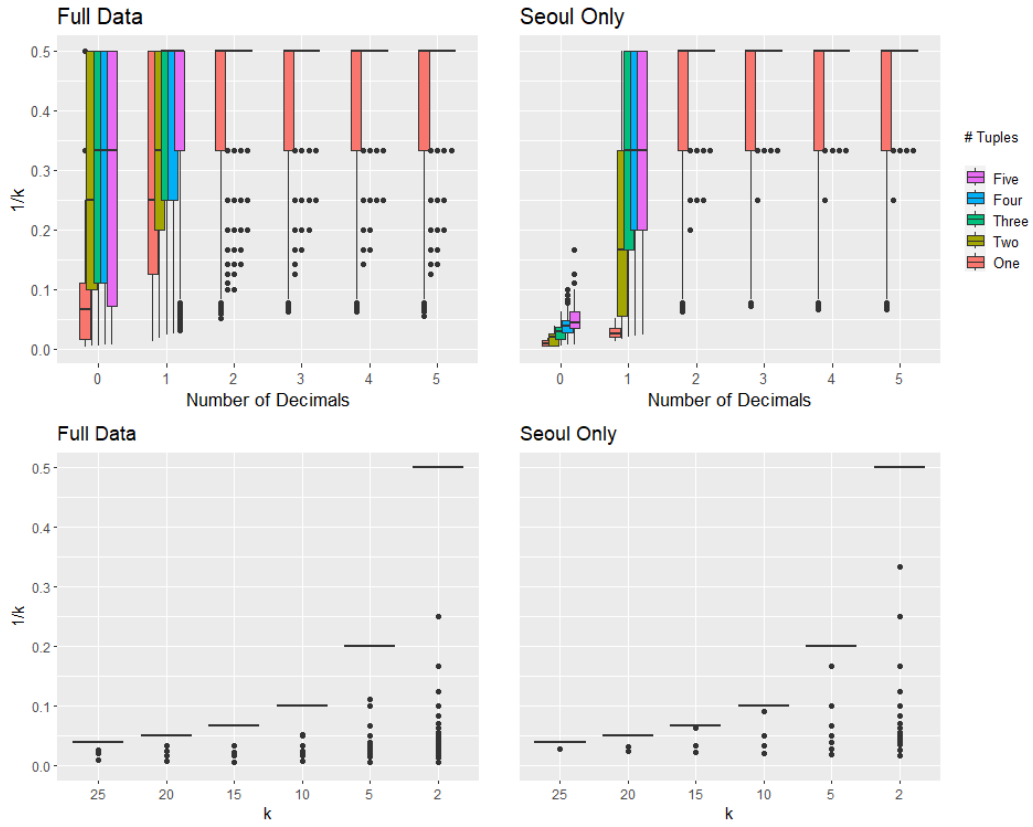

Figure 3: Distribution of $1/k$ for each value of $d = 0, 1, \ldots, 5$ and sampled trajectory lengths from one to five tuples in South Korea (left) and Seoul (right).

Given that preventing Singling-out is a stronger requirement than preventing Linkability, we also consider the case where a data controller decides to only prevent Linkability. When Singling-out is not prevented in $\mathbf{Y}^\ell$, some individual with a completely unique trajectory exists. Suppose an adversary does not know the true identity of individual n and the adversary's external linking data set $\mathbf{X}^\ell$ contains one or more individuals with trajectories that could be linked to individual n. As a result, the adversary may have a one-to-one linkage, which constitutes Linkability, or a many-to-one linkage, which prevents Linkability. Past research indicates that four spatio-temporal points corresponding to geographical areas ranging between 0.15 to 15 square kilometers are enough to uniquely identify 95 percent of individuals from a population of 1.5 million [36]. By our calculation, location tuples coarsened to $d = 2$ fall within the range of one square kilometer. For $d = 2$, Figure 2 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 data set of the entire population and CI consisting of four or more location tuples would likely have a one-to-one linkage for over 95% of the individuals in $\mathbf{Y}^\ell$. This implies that preventing Linkability by coarsening locations when Singling-out is not prevented is difficult unless the organization uses a value of $d \leq 1$ in both high and low density regions. We suggest organizations attempting to anonymize location trajectories using Location Coarsening first prevent Singling-out; otherwise, consider different approaches, such as noise infusion, to release a less linkable data set of protected latitude and longitude coordinates.

5.6 Preventing Inference

There is a great deal of sensitive information that could be deduced from location data. For example, [34] 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. For our application, the most likely inference of interest in our data set is attribute disclosure to determine whether an individual has COVID-19.

To study the probability of a successful attribute disclosure, we randomly assign 1% of individuals to have COVID-19 in our original data set $\mathbf{Y}^\ell$. All location tuples from these individuals' trajectories are marked as having COVID-19 in the data set $\mathbf{Y}$. Using Equation (3), we define $s_n = \{0, 1\}$ as the COVID-19 status of individual or location n with the prior probabilities $p(1|\,\mathbf{X}, \mathbf{b}) = 0.01$ and $p(0|\,\mathbf{X}, \mathbf{b}) = 0.99$. The updated probabilities when the adversary has the protected data set are $p(1\,|\,\mathbf{x},\, \mathbf{b},\, \mathbf{Y}^\ell)$ for Location Coarsening and $p(1\,|\,\mathbf{x},\, \mathbf{b},\, \mathbf{Y})$ for microaggregation.

For a reasonable attribute disclosure attack in use case one, we assume the adversary matches the larger population data set $\mathbf{X}^\ell$ to the location trajectories released in a version of $\mathbf{Y}^\ell$ with Singling-out and Linkability prevented. To produce the updated probability in Equation (3), 100 simulated data sets are generated for each value of d and trajectory length (one to five). We only consider values of $d \in \{0, 1, 2\}$ since the majority of trajectories had to be deleted to prevent Singling-out and Linkability when $d > 2$. For each simulation, one percent of individuals are randomly assigned a positive COVID-19 status. Then, for each positive individual, we randomly sample one location tuple to treat as CI. We identify the trajectories that match each CI tuple $\mathbf{x}$ and take the mean of the matching COVID-19 statuses to compute the updated probability, $p(1\,|\,\mathbf{x},\, \mathbf{b}, \mathbf{Y}^\ell)$ for each positive individual. The left-hand side of equation (3) follows and we compute the maximum across all COVID-19 positive individuals in each simulation. This process is performed once using the full data

and once using the Seoul data.

For use case two, we also perform 100 simulations to assess a reasonable attribute disclosure attack against the microaggregated data sets with Singling-out and Linkability prevented. We sample one location tuple from the trajectory of each COVID-19 positive individual and treat these tuples as CI. We assume the adversary performs a nearest-neighbor based matching such that each CI tuple is linked to its nearest equivalence class based on Euclidean distance. We compute $p(1 \mid \mathbf{x}, \mathbf{b}, \mathbf{Y})$ as the mean of the COVID-19 status within the linked equivalence class, and compute the maximum of the LHS of the Inference condition across all COVID-19 positive individuals in each simulation.

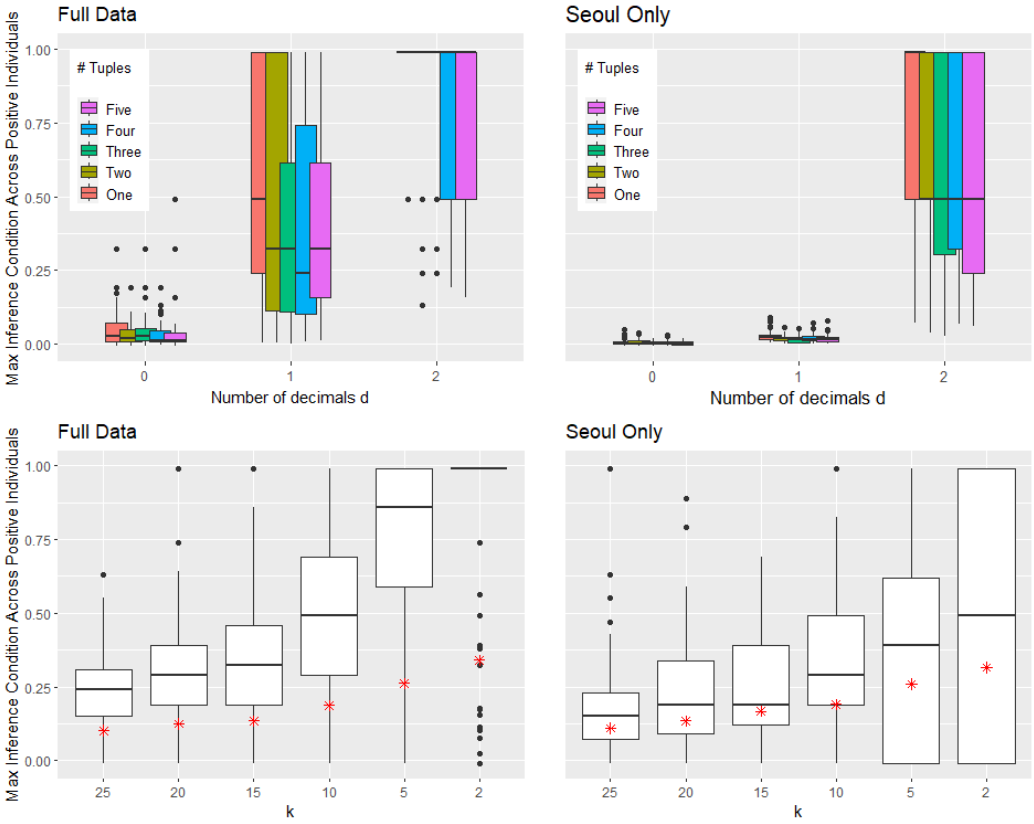

Figure 4: Boxplots of the maximum increase in Inference probabilities for COVID-19 positive individuals under Location Coarsening (top row) and Microaggregation (bottom row).

Figure 4 plots boxplots of the maximum increase in Inference for each use case. For Location Coarsening in the top row, a boxplot is shown for each value of d and trajectory lengths of one to five tuples in South Korea and Seoul, respectively. For microaggregation in the bottom row, a boxplot is shown for each value of k in South Korea and Seoul.

For use case one with Location Coarsening, preventing Inference is extremely difficult in low density regions of South Korea. Locations must be coarsened to $d = 0$ for most of the simulations to have a maximum increase in probability of 12% or less. Preventing

Inference is more likely in high density regions like Seoul for $d \leq 1$ with most simulations having 5% maximum increase in probability. We also see that the maximum increase in Inference tends to be slightly lower for longer trajectories since each trajectory contains a higher number of locations and it is more likely that a single CI tuple matches to multiple trajectories. To prevent Inference for Location Coarsening at higher values of d, 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.

For use case two with microaggregation, the maximums are lower in Seoul than South Korea. Interestingly, the maximum increases in Inference are significant for large equivalence class sizes, *e.g.*, the maximum increase is over 12% in about 50% of simulations when $k = 25$. To explain this result, we plot red stars in Figure 4 that denote the average proportion of locations corresponding to the targeted COVID-19 individual in the linked equivalence class. The targeted individuals tend to have multiple locations in the targeted equivalence class, *e.g.*, about three out of twenty-five locations correspond to the target individual when $k = 25$, leading to larger increases in Inference.

For comparison between the two use cases, Location Coarsening in the Seoul data for $d = 0$ produces equivalence classes of approximately twenty-five trajectories, and most of the maximum increases are less than 5%, since each positive individual can only increase the probability of Inference by $1/25$. We also see cases where the coarsened trajectories provide better protection against Inference than Microaggregation for significantly smaller equivalence class sizes. For example, the distributions in the Seoul data for $d = 1$ are significantly lower than for microaggregation for $k \in \{10, 15, 20, 25\}$ even though most of the coarsened tuples are contained in equivalence classes of six or fewer individuals. Overall, it will be challenging for organizations demanding a high level of usefulness to prevent Inference for all individuals in the simplest of cases, such as using a binary variable that indicates disease status. Another reasonable approach may be to delete the data of individuals who have a maximum probability of Inference above an acceptable level.

5.7 Summary of Results

This application illustrated a case study of whether anonymization methods can produce legally anonymous location data while retaining usefulness.

For use case one with location coarsening, a low value of $d = 0$ was necessary to anonymize location data for low density regions. However, this had an enormous negative impact on usefulness, with most locations shifting by over 47 kilometers. Data from high density regions was anonymized for a slightly higher value of $d = 1$, but most locations still shifted by 3.7 kilometers or more. In sum, coarsened individual-level location trajectories were not private and required too much coarsening to meet legal criteria.

For use case two, microaggregation improved upon Location Coarsening at preventing Singling-out and Linkability, which resulted in most locations being shifted by zero meters when $k = 2$ and 140 meters when $k = 5$. However, microaggregation removed the longitudinal links between tuples and prevented most applications such as predicting consumers' path-to-purchase or tracking the spread of diseases. Microaggregation also struggled to prevent Inference and required significant data deletion to limit Inference to a legally acceptable level. On the other hand, Location Coarsening produced better protection against Inference at $d = 1$ than Microaggregation at any value of k.

We assessed two reasonable anonymization solutions for specific use cases with location data. We found that the use cases were severely degraded when preventing all three criteria simultaneously.

We also showed the importance of incorporating contextual factors in the anonymization process, such as using location data in low and high density areas. Specifically, in high density areas, anonymized location data can meet legal anonymization requirements while retaining a higher degree of usefulness.

6 Discussion

Location data are essential for tracking and mitigating the spread of disease, predicting consumers' path-to-purchase , providing accurate point-of-interest recommendations and context-aware marketing messages, influencing taxi drivers' decision-making processes to improve driver incomes and market efficiency, disease mapping, and geo-targeting and conquesting. In this paper, we proposed an interpretation of legal anonymization criteria to evaluate converting personal location data to non-personal location data under the GDPR. Using two reasonable anonymization solutions for the above use cases, we found that it was not possible to produce legally anonymous location data while retaining usefulness. We also found that context-specific factors, such as population density, influence whether data can be legally anonymized.

This study was the first attempt to interpret and apply legal anonymization criteria to location data. Without case law on this challenging problem, our study provides organizations with an example to document a reasonable effort at anonymizing location data. We note that the anonymization techniques we used are not likely to meet anonymization criteria for other complex data sets, including retail point-of-sale transactions [46], time series [37], textual data stored from chatbots, search history logs [53], facial images [62, 1], social network data [23], or even contact tracing data [17] derived from location data. As a result, we encourage government agencies and privacy-minded organizations to invest in more sophisticated anonymization techniques that retain the usefulness of complex data sets.

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 straightforward solutions (as shown in our application) that may only satisfy certain regulatory definitions, incur the regulatory costs and corresponding risks to continue using personal data, or, more likely, implement privacy and security controls using personal data.[12] However, even with additional security and legal measures designed to protect personal data, personal data still gets out. A report by Verizon found that 34% of the 2,013 data breaches in 2019 involved internal actors such as employees [58], and the number of data breaches more than doubled in the following years. Assuming a data breach will eventually occur, this leads to the question of whether it is more important to comply with the law or anonymize the data at its source.

For limitations of this paper, we focused on generalization techniques in our empirical application since these do not introduce randomness into the data and are more palatable to the user. However, applying randomization techniques such as differential privacy approaches to location data is an active area of research [35], and future work should compare randomization vs. generalization-based approaches in achieving legal anonymization for

[12] 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 [40].

location data. In Section 4, we noted that the literature supports the possibility of other interpretations of Singling-out, Linkability, and Inference. As opposed to [10], we excluded differential privacy from our interpretation as not being reasonable for organizations. Future work could examine the implications of these interpretations on the resulting balance of privacy and utility in the anonymized data. It would also be valuable to analyze whether more advanced anonymization techniques, such as applying varying levels of coarsening on a per-trajectory basis [30] or creating synthetic location data using generative models [35, 60] can produce more useful legally anonymized data.

This paper is a demonstration of the value of bringing multidisciplinary stakeholders together to develop privacy-promoting methods to address personal and non-personal data. And, we hope for lawyers, statisticians, and data users to sit at the table together to continue to engineer privacy-oriented solutions. The EDPB has also provided commentary [18] on the ability of various anonymization techniques, broadly categorized as either randomization or generalization, to protect against re-identification. The EDPB concluded that no anonymization technique on its own is guaranteed to anonymize data based on Singling-out, Linkability, and Inference. While based on past research, the EDPB's opinions were not specific to location data and they ultimately concluded that the optimal solution for anonymizing a data set 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 the 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 [56] (*e.g.*, 2-anonymity may comply with the law but provide little privacy). Further, while our interpretations were 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.

References

[1] Alessandro Acquisti, Ralph Gross, and Frederic D. Stutzman. Face recognition and privacy in the age of augmented reality. *Journal of Privacy and Confidentiality*, 6.2, (2014), pp. 1–20.

[2] Charu C. Aggarwal. On *k*-anonymity and the curse of dimensionality. *VLDB*, 5, (2005), pp. 901–909.

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

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

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

[6] 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.

[7] Jane Bambauer, Krishnamurty Muralidhar, and Rathindra Sarathy. Fool's gold: an illustrated critique of differential privacy. *Vand. J. Ent. & Tech. L.*, 16, (2013), pp. 701–755.

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

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

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

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

[12] 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.

[13] Josep Domingo-Ferrer and Vicenç Torra. Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, 13, (2003), pp. 343–354.

[14] Josep Domingo-Ferrer and Vicenç Torra. Disclosure risk assessment in statistical data protection. *Journal of Computational and Applied Mathematics*, 164, (2004), pp. 285–293.

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

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

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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] Stephen E. Fienberg. Is the Privacy of Network Data an Oxymoron? *Journal of Privacy and Confidentiality,* 4.2, (2013), pp. 1–5.

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

[25] Nathan M. Fong, Zheng Fang, and Xueming Luo. Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research,* 52.5, (2015), pp. 726–735.

[26] Paul Francis et al. Extended Diffix. *arXiv 1806.02075* (2018).

[27] Srivatsava R. Ganta, Shiva P. Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* (2008), pp. 265–273.

[28] Simson L. Garfinkel. De-identification of Personal Information. *US Department of Commerce, National Institute of Standards and Technology,* (2015).

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

[30] Marco Gramaglia and Marco Fiore. Hiding mobile traffic fingerprints with glove. *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies,* (2015), pp. 1–13.

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

[32] 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.

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

[34] Meghanath Macha et al. Personalized Privacy Preservation in Consumer Mobile Trajectories. *Forthcoming in Information Systems Research,* (2023). DOI: https://doi.org/10.1287/isre.2023.1227.

[35] Anna Monreale and Roberto Pellungrini. A Survey on Privacy in Human Mobility. *Transactions on Data Privacy,* 16.1, (2023), pp. 51–82.

[36] Yves-Alexandre de Montjoye et al. Unique in the crowd: the privacy bounds of human mobility. *Scientific Reports,* 3, (2013). DOI: https://doi.org/10.1038/srep01376.

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

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

[39] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.,* 57, (2009), pp. 1701–1777.

[40] David Roland-Holst et al. Standardized Regulatory Impact Assessment: California Consumer Privacy Act of 2018 Regulations. *State of California Department of Justice Office of the Attorney General*, (2019). URL: https://dof.ca.gov/wp-content/uploads/sites/352/Forecasting/Economics/Documents/CCPA_Regulations-SRIA-DOF.pdf.

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

[42] Steven Ruggles et al. Differential privacy and census data: Implications for social and economic research. *AEA papers and proceedings*, 109, (2019), pp. 403–408.

[43] Michael L. Rustad and Thomas H. Koenig. Towards a global data privacy standard. *Fla. L. Rev.*, 71, (2019), pp. 365–453.

[44] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA.* (1998), pp. 1–19.

[45] Matthew J. Schneider and Dawn Iacobucci. Protecting survey data on a consumer level. *Journal of Marketing Analytics*, 8, (2020), pp. 3–17.

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

[47] Chris Skinner. Assessing disclosure risk for record linkage. *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference, Istanbul, Turkey,* (2008), pp. 166–176.

[48] Chris J. Skinner and Mark J. Elliot. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 64.4, (2002), pp. 855–867.

[49] Frederic D. Stutzman, Ralph Gross, and Alessandro Acquisti. Silent listeners: The evolution of privacy and disclosure on Facebook. *Journal of privacy and confidentiality*, 4.2, (2013), pp. 7–41.

[50] Chenshuo Sun et al. Predicting Stages in Omnichannel Path to Purchase: A Deep Learning Model. *Information Systems Research*, 33.2, (2022), pp. 429–445.

[51] Latanya Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10.5, (2002), pp. 557–570.

[52] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67, (2015), pp. 1–36.

[53] Vincent Toubiana and Helen Nissenbaum. An analysis of google log retention policies. *Journal of Privacy and Confidentiality*, 3.1, (2011), pp. 3–26.

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

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

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

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

[58] Verizon. Verizon 2019 data breach investigations report (DBIR). 2019. URL: https://www.verizon.com/business/resources/reports/2019-data-breach-investigations-report.pdf.

[59] William E. Winkler. Matching and record linkage. *Wiley interdisciplinary reviews: Computational statistics*, 6.5, (2014), pp. 313–325.

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

[61] Yingjie Zhang, Beibei Li, and Ramayya Krishnan. Learning individual behavior using sensor data: The case of global positioning system traces and taxi drivers. *Information Systems Research*, 31.4, (2020), pp. 1301–1321.

[62] Yinghui Zhou, Shasha Lu, and Min Ding. Contour-as-Face Framework: A Method to Preserve Privacy and Perception. *Journal of Marketing Research*, 57.4, (2020), pp. 617–639.

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 the EDPB's definition of anonymization was considered 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), [9]) 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), [22]).
The CCPA defines de-identification (§ 1798.140(h), [9]) 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), [9]) 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), [32]) 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 [28] as:

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

NIST's definition focuses on the association between data sets 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.

Additional Use Cases for COVID-19 Location Data

There are other 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) [38, 54]. Mandatory quarantines enforced using geolocation data do not meet the criteria of non-personal data because by definition, they require individuals to be 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 [23] and [49].