LEGALLY ANONYMIZING PERSONAL DATA: AN APPLICATION TO SOUTH KOREAN COVID-19 LOCATION DATA

CAMERON D. BALE, JORDAN L. FISCHER, MATTHEW J. SCHNEIDER, AND STEVEN WEBER

Lebow College of Business, Drexel University, Philadelphia, PA
e-mail address: cdb327@drexel.edu

Thomas R. Kline School of Law, Drexel University, Philadelphia, PA
e-mail address: jlf324@drexel.edu

Lebow College of Business, Drexel University, Philadelphia, PA
e-mail address: mjs624@drexel.edu

Electrical and Computer Engineering, Drexel University, Philadelphia, PA
e-mail address: spw26@drexel.edu

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, variations in commonly used privacy terms create significant differences for legally anonymizing personal data. This complexity of privacy legislation necessitates a cohesive mathematical framework to assess compliance with privacy law. We propose a mathematical framework taken from regional privacy laws to convert personal data to non-personal data. We test our framework on two logical statistical solutions on geolocation data of COVID-19 patients in South Korea. We find that both solutions produced legally anonymized data, but with varied degrees of usefulness.

1. Introduction and Privacy Law Overview

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 (European Parliament and Council of European Union [2016]). Since then, a number of regions, including Japan, Brazil, Canada, and California, have followed suit with similar, and sometimes dissimilar, privacy-oriented laws.[1]

Key words and phrases: Privacy Law, Anonymization, Statistical Analysis, Location Data.

[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 Brazil [2018]; Japan's Act on the Protection of Personal Information, Act No 57 of 2003, as amended in 2016 Japan [2016]; see also Rustad and Koenig [2019], pp. 431–48, discussing the impact of the GDPR across numerous regions on the world.

1.1. **Legal Definitions of Personal Data and Non-Personal Data.** 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) European Parliament and Council of European Union [2016]; California Consumer Privacy Act (CCPA), 1798.140(o)(1) California State Legislature [2018]). 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.[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.). And, 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 European Parliament and Council of European Union [2016]). 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.

For example, in Case 582/14 – Patrick Breyer v Germany (Court of Justice of the European Union [2016]), the European Court of Justice (ECJ) held that Internet Protocol ("IP") addresses, in certain circumstances, are considered personal data.[3] 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, and often challenging, the definition of personal data. Within these definitions and approaches of the abstract concept of privacy, these laws all recognize, either expressly or inherently, that techniques exist to convert personal data to non-personal data; *i.e.*,

[2]See GDPR, Recital 26 European Parliament and Council of European Union [2016]; see also the California Consumer Privacy Act of 2018 ("CCPA"), § 1798.140(O(3) California State Legislature [2018]

[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, European Parliament and Council of European Union [1995]). However, its ruling still applies under the GDPR.

non-identifiable to the individual who is related to the data collected. And, this non-personal data is often exempt from the data security requirements. (GDPR, Recital 26 (European Parliament and Council of European Union [2016]); CCPA § 1798.140(O(3)) (California State Legislature [2018])).[4] 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.

Within this backdrop of abstract legal definitions and the application of those legal definitions of personal data, a growing challenge exists: the legal definition of non-personal data, and the practical consequences of those definitions. 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.

For example, the GDPR recognizes the term pseudonymization, defining it (GDPR, Art. 4(5) (European Parliament and Council of European Union [2016])) 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."

It is important to note that pseudonymized data does still fall under the GDPR.

And, 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 (European Parliament and Council of European Union [2016])). 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, EDPB [2020]) 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).

The April 2020 Guidelines further outline three criteria to determine the "robustness" of the anonymization technique used: "(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

[4] "De-identification or anonymization techniques can be used to remove the data set from the scope of regulations." (Achatz and Hubbard [2017])

about an individual)." ¶16. 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.

Comparatively, 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), California State Legislature [2018]) 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), European Parliament and Council of European Union [2016]).

The CCPA defines de-identification (§ 1798.140(h), California State Legislature [2018]) 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), California State Legislature [2018]) 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 short, these regulations sometimes use the same terms. But, these regulations often use different terms, and define those terms in various ways. Ultimately, these differences can often 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.

The GDPR also introduces the term anonymization to describe 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, European Parliament and Council of European Union [2016]).

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. 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. But, the language used to define each term can result in distinct mathematical differences.

1.2. **Technological Solutions for Personal Data.** Due to the disjointed landscape, each legal definition potentially overlaps with nuanced mathematical differences. Creating statistical or technological solutions to convert personal data to non-personal data requires the difficult navigation of these legal differences using a consistent mathematical framework. This may result in technological solutions with varying privacy and security impacts and corresponding risks. Given a consistent framework, the questions then become: What are the practical, statistical solutions that could be applied to convert personal data to non-personal data within these varying legal definitions? 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?

As an 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.

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), Japan [2016]) 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 (Garfinkel [2015]) 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.

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 as it appears to encapsulate many of the criteria that various legal definitions consider when determining if data is non-personal. From these criteria, we create a single mathematical framework that captures the intent of various privacy legislation. We apply this framework to location data from COVID-19 patients in South Korea, and demonstrate two possible statistical solutions to convert personal data to non-personal data, resulting in different impacts to the usefulness of the data.

1.3. **Legal Criteria for Converting Personal Data to Non-Personal Data.** As noted above, the EDPB 2020 April Guidelines outline three criteria used to determine the "robustness of anonymization" to convert personal data to non-personal data (¶ 16, EDPB [2020]):

> (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" Article 29 Data Protection Working Party [2014]). In WP Guidelines, the Working Party made clear that an inherent requirement of anonymization is that "the processing must be irreversible" (¶ 2.1). And, 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 the anonymization technique used is sufficiently robust is "whether identification has become reasonably impossible" (¶ 2.2.2; see also 2020 April Guidelines EDPB [2020], ¶¶ 15, 22).

Robustness of anonymization is key to ensuring that personal data becomes, and remains, non-identifiable. The risk of reidentification is a growing concern (¶19, EDPB [2020]). Paul Ohm (Ohm [2009]) 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 mathematical solution for anonymization.

Further complicating this analysis is the question: from whose perspective should reasonableness be viewed? The individual? The group of individuals relevant to the

data? The entity processing the data? The privacy engineer? When the perspective of reasonableness changes, as we explore later, it can create different statistical outcomes. These outcomes impact the usefulness of data, which can range from pure noise with strong privacy guarantees, to well-preserved statistics with weak privacy guarantees. Ultimately, this reasonableness concept increases the complexity in defining a single mathematical framework for anonymization within the opaque guidance of the law. The EU's definition of anonymization 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.

Since exploration of other regional data protection regulations provides little additional insight into the statistical interpretation of anonymization, the EDPB's definition of anonymization is exhaustive for the variety of different legal frameworks and regions and will be explored in further detail in the remainder of this paper. In Section 2, we take the criteria used to determine the robustness of anonymization to define a consistent mathematical framework. When considered jointly, these criteria provide a method of determining whether or not personal data is legally anonymized, and therefore it becomes non-personal data which is exempt from strict privacy and security controls. In Section 3, we apply two reasonable and straight-forward statistical solutions to location data from COVID-19 patients in South Korea in an attempt to convert personal data to non-personal data. This application is important to the legal privacy community because there is an absence of case law showing when data meets the criteria, and practitioners are likely to start with simple and well understood solutions for anonymization. We find that both solutions are viable, but result in different degrees of usefulness, and Section 4 concludes with legal recommendations and future areas of research.

2. Mathematical Framework of Anonymization

In this section, we translate the three criteria on the robustness of anonymization derived from EU guidance into mathematical criteria that constitute the prevention of these criteria and the successful conversion of personal data to non-personal data. Per EDPB guidelines, these mathematical criteria are intended to represent the most reasonable means used by an organization or third party.

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*

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

2.1. Singling-out.

Define $\hat{\mathbf{Y}}$ as the matrix formed from the unique rows found in the database $\mathbf{Y}$. 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 $i \in \{1, \dots, I\}$ 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

$$\inf \hat{\mathbf{z}} \geq 2. \tag{2.1}$$

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. Even though (2.1) is a direct translation of the law, it may not capture the full intention of the law and more generally, an organization can use k-anonymity to increase privacy (Sweeney [2002]) where

$$\inf \hat{\mathbf{z}} \geq k. \tag{2.2}$$

2.2. **Linkability.** We define Linkability to mean that a one-to-one linkage occurs for some individual between $\mathbf{Y}$ and an external database $\mathbf{X}$. There are three types of linkage that 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 there exists an external linking database $\mathbf{X}$ containing 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 database 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 database includes at least some data on all individuals in $\mathbf{Y}$, the pseudonymized database $\mathbf{Y}$ 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 database 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 $\inf \mathbf{z} \geq 2$. 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.

2.3. **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 "...re-identification performed by the most likely and reasonable means" (2020 April Guidance, ¶ 19-20, EDPB [2020]) an organization or third party will employ.

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 true value for individual m that the organization is trying to prevent Inference on. The adversary has a prior probability $p(s^* \,|\, \mathbf{X})$ of the individual having a value s^* based on external information only. Once $\mathbf{Y}$ is released, the adversary updates the probability to $p(s^* \,|\, \mathbf{x}, \mathbf{Y})$ based on the new information. The pseudonymized database $\mathbf{Y}$ prevents Inference for an individual m when

$$p(s^* \,|\, \mathbf{x}, \mathbf{Y}) - p(s^* \,|\, \mathbf{X}) \leq c. \tag{2.3}$$

In practice or not, the change in probability on the left-hand side (LHS) will always be between -1 and 1. Per the EDPB's guidelines, we view equation (2.3) 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 equation (2.3) is negative, the criterion is not violated as inference is worse on individual m.

For a fixed threshold of c, (2.3) is unlikely to hold for all M individuals unless $\mathbf{Y}$ is sufficiently protected. Thus, one reasonability test may average the left-hand side (LHS) of (2.3) over all individuals in cases where the inferences are improved, $p(s^* \,|\, \mathbf{x}, \mathbf{Y}) > p(s^* \,|\, \mathbf{X})$. Otherwise, we suggest setting the LHS to 0 in cases where inference is degraded, $p(s^* \,|\, \mathbf{x}, \mathbf{Y}) < p(s^* \,|\, \mathbf{X})$.

To illustrate (2.3), suppose an adversary seeks to deduce the income of an individual m in $\mathbf{Y}$ using only $\mathbf{X}$, which consists of the relative frequencies in the population for standard income brackets. In this case, the prior probabilities are equal and equate to randomly guessing an income bracket.

$$p(s \,|\, \mathbf{X}) = \begin{cases} 0.20, & s \leq 24,999 \\ 0.20, & 25,000 \leq s \leq 49,999 \\ 0.20, & 50,000 \leq s \leq 74,999 \\ 0.20, & 75,000 \leq s \leq 99,999 \\ 0.20, & 100,000 \leq s \end{cases}$$

Prior knowledge should be reasonable to serve as a baseline for assessing whether Inference is prevented in a database $\mathbf{Y}$. Of course, we cannot limit the amount of external information available to an adversary - *i.e.*, the adversary may know the exact income of individual m without obtaining $\mathbf{Y}$ which implies that 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.

2.3.1. *Example with Two Individuals.* Suppose the adversary observes only two individuals in $\mathbf{Y}$ with the same CI as targeted individual m, which results in the following updated probabilities:

$$p(s \,|\, \mathbf{x}, \mathbf{Y}) = \begin{cases} 0.00, & s \leq 24,999 \\ 0.00, & 25,000 \leq s \leq 49,999 \\ 0.00, & 50,000 \leq s \leq 74,999 \\ 0.00, & 75,000 \leq s \leq 99,999 \\ 1.00, & 100,000 \leq s \end{cases}$$

Under this scenario, the adversary deduces with one hundred percent probability that the income of individual m is greater than 100,000, and

$$p(s^* \,|\, \mathbf{x}, \mathbf{Y}) - p(s^* \,|\, \mathbf{X}) = 1.00 - 0.20 = 0.80$$

which is almost certainly in violation of (2.3). Averaging the left-hand side of (2.3) over the individuals yields the same result since both individuals are in the same income category.

Another reasonability test may use t-closeness to measure the variational distance (Dosselmann et al. [2019]; Li et al. [2007]) between $p(s \,|\, \mathbf{x}, \mathbf{Y})$ and $p(s \,|\, \mathbf{X})$. In this case, t-closeness is

$$\frac{1}{2}\Big[\big|0.00 - 0.20\big| + \big|0.00 - 0.20\big| + \big|0.00 - 0.20\big| + \big|0.00 - 0.20\big| + \big|1.00 - 0.20\big|\Big] = 0.80$$

which is poor privacy.

2.3.2. *Example with More Than Two Individuals.* Suppose that the adversary observes 100 indiviuals in $\mathbf{Y}$ with the same CI as the targeted individual in $\mathbf{x}$, and the updated probabilities are the following:

$$p(s \,|\, \mathbf{x}, \mathbf{Y}) = \begin{cases} 0.19, & s \leq 24,999 \\ 0.10, & 25,000 \leq s \leq 49,999 \\ 0.21, & 50,000 \leq s \leq 74,999 \\ 0.10, & 75,000 \leq s \leq 99,999 \\ 0.40, & 100,000 \leq s \end{cases}$$

.

The probability of the adversary deducing unknown information about individual m has changed, no matter which bin s^* falls into. If the income of the targeted individual m is at least 100,000, then

$$p(s^* \,|\, \mathbf{x}, \mathbf{Y}) - p(s^* \,|\, \mathbf{X}) = 0.40 - 0.20 = 0.20$$

If $50,000 \leq s^* \leq 74,999$, then the LHS evaluates to 0.01, which would most likely not be considered a violation of the law. Finally, if s^* falls into any of the other bins, then the LHS of (2.3) is negative, which does not constitute a violation.

The LHS of (2.3) may be averaged over the 100 individuals only when inference is improved, $((0.21 - 0.20) \times 0.21 + (0.40 - 0.20) \times 0.40 + (0) \times 0.39) = 0.082$.

In this case, t-closenes is

$$\frac{1}{2}\Big[\big|0.19 - 0.20\big| + \big|0.10 - 0.20\big| + \big|0.21 - 0.20\big| + \big|0.10 - 0.20\big| + \big|0.40 - 0.20\big|\Big] = 0.21$$

which is also significantly lower and results in much better privacy than the previous example.

Overall, whether the previous examples constitute violations of (2.3) 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. However, more strongly,

differential privacy could be used to bound increased inference on all individuals (Dwork [2006]), but may currently be beyond the technical abilities or interests of most organizations. Although we do not address differentially private mechanisms in this paper, we refer the reader to recent research using PrivTree (Zhang et al. [2016]). 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 (2.3), it is a bound in the odds ratio (instead of a subtraction) of databases differing on at most one record. Furthermore, it would be more similar to measuring the worst-case scenario for any individual in database $\mathbf{Y}$ for (2.3) which we explore in the next section.

3. Application to South Korea COVID-19 Location Data

3.1. **Data Description.** We use a location database 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 (DS4C [2020]). 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 (Douriez et al. [2016]). Furthermore, we build on the notation introduced in Section 2 and define a released longitudinal location database $\mathbf{Y}^{\ell,d}$. 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 database $\mathbf{Y}^{\ell,d}$ 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 database 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.

3.2. **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..."(EDPB [2020]). 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.

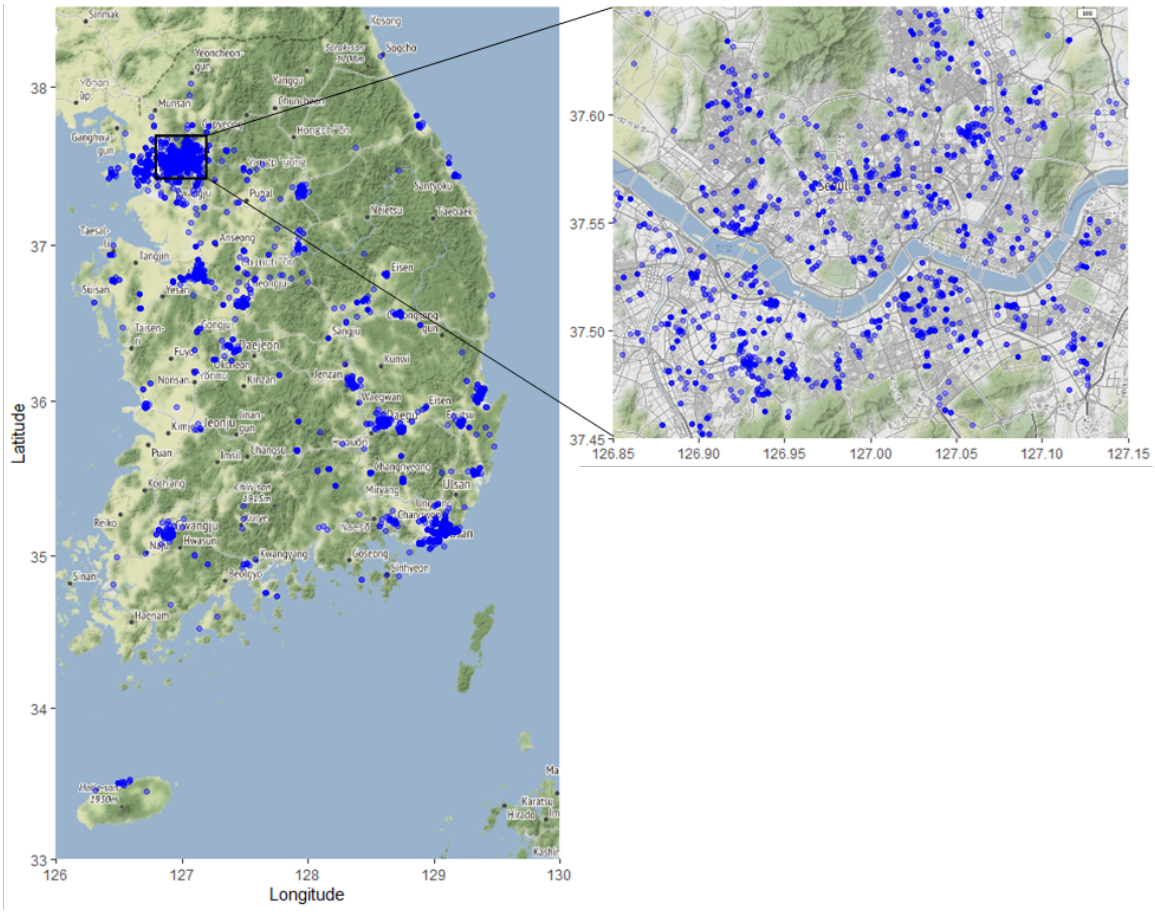

Figure 1: Location Tuples Mapped Over South Korea (left) and Seoul (right) for ($d = 5$)

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) (NRF [2020]; Tsang [2020]). 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 Fienberg [2013] and Stutzman et al. [2013].

We use two logical and relatively simple solutions to protect the original location trajectory database $\mathbf{Y}^{\ell,5}$: (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 (2020 April Guidelines, para. 15-16 EDPB [2020]).

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 them briefly first, and then investigate how well they meet the legal criteria of preventing Singling-out, Linkability, and Inference.

3.2.1. *Location Coarsening.* Location Coarsening rounds latitude and longitude coordinates in the original data to a lower number of decimal places, $d < 5$. 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 database $\mathbf{Y}^{\ell,d}$. 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 use the geosphere package in R to measure the geographical distance (in meters) that points are shifted by Location Coarsening. We perform the Location Coarsening manually by rounding to d decimal places and Figure 2 illustrates the location tuples when $d = 1$.

3.2.2. *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 database $\mathbf{Y}^{\ell,5}$ is transformed to a database of counts which removes the longitudinal nature of the location trajectories. Define an Aggregating to Counts database $\mathbf{C}^d$ 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 (Dwork et al. [2017]; Tu et al. [2018]).[5]

3.3. **Preventing Singling-out.**

3.3.1. *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 databases for each value of d. These 100 simulated databases are created by sampling (without replacement) one to five location tuples from each trajectory in the original database. We calculate the percentage of trajectories that are unique in each simulated database. We perform this analysis twice; once using the less dense South Korea data and once using the dense Seoul data. We measure the resulting usefulness of the data by computing the geographical distance shifted between coarsened location tuples and original location tuples.

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

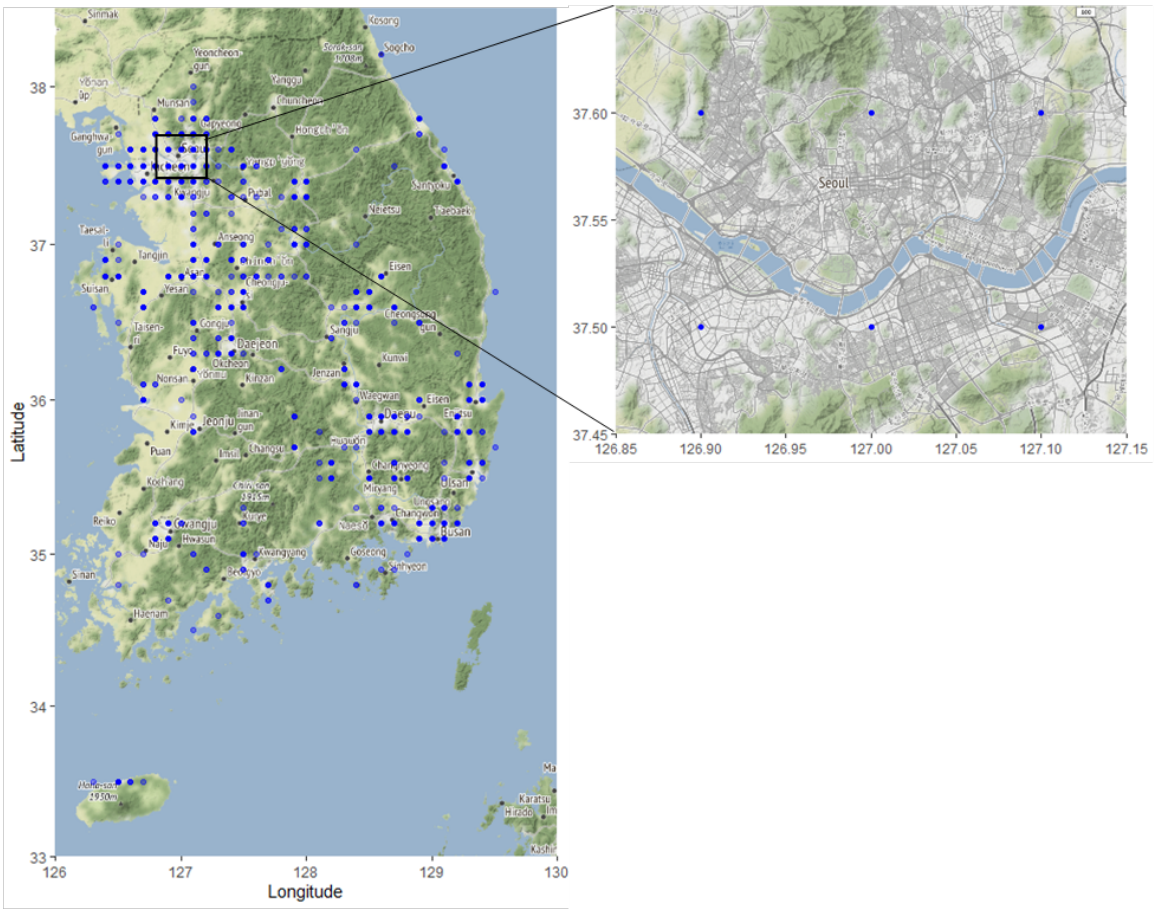

Figure 2: Location Tuples Mapped Over South Korea (left) and Seoul (right) for $d = 1$

Table 1 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.

Figure 3 expands the detail of Table 1 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.[6]

[6]Our results were obtained from location data without considering the temporal aspect that often accompanies this type of data. Spatio-temporal data is likely even more difficult to prevent Singling-out since time is another variable on which to differentiate individuals.

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

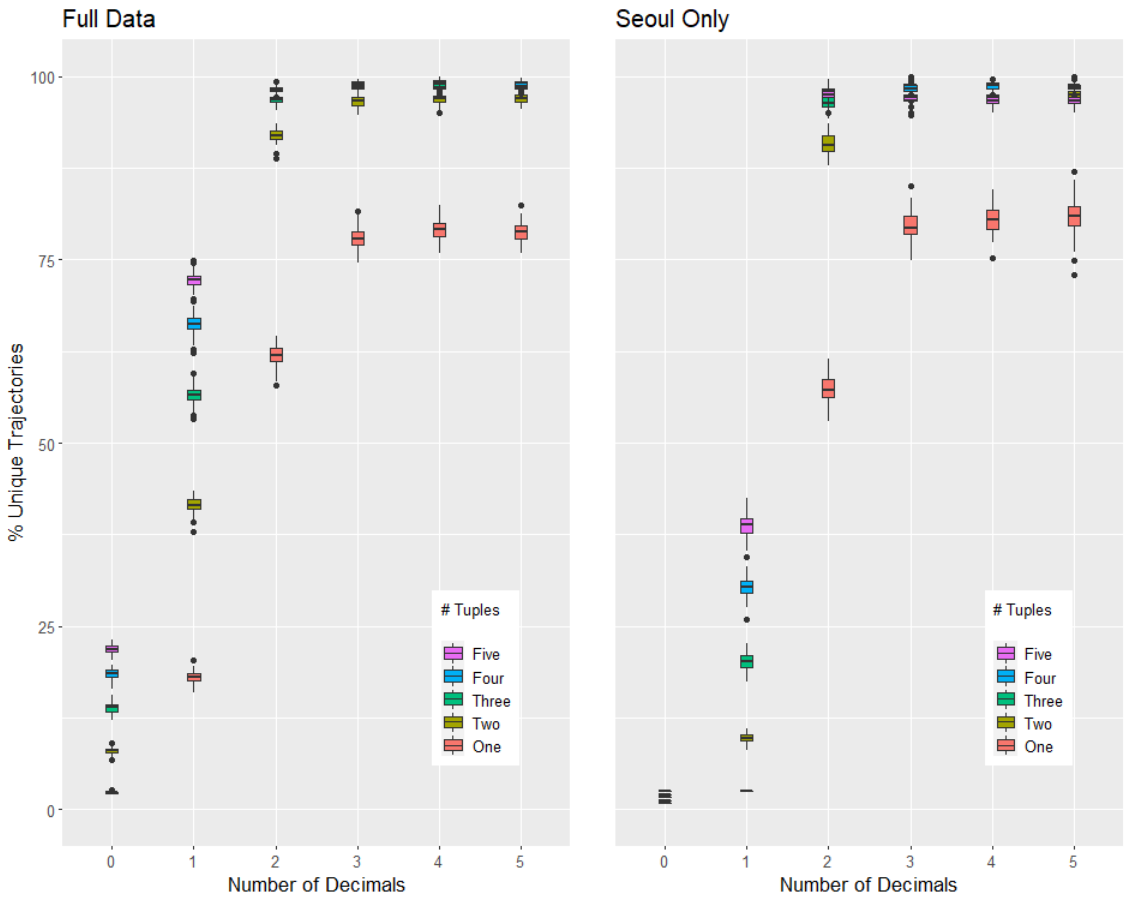

Figure 3: Percentage of unique trajectories for each value of $d = 0, 1, \ldots, 5$ and sampled trajectory lengths from one to five tuples in South Korea (left) and Seoul (right)

3.3.2. *Aggregating to Counts.* Recall that Aggregating to Counts converts the original database $\mathbf{Y}^{\ell,5}$ to a protected database 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.

Table 2 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, but comes at a cost to the longitudinal information. 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), which is an improvement in usefulness over Location Coarsening at similar levels of protection.[7] 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	0.000099	36.86% (924)	37.73% (349)
3	89.14	110.967	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 2: *Aggregating to Counts*: Areas of regions and percentage of regions preventing Singling-out

3.4. **Preventing Linkability.** When not preventing Singling-out, preventing Linkability for Location Coarsening and Aggregating to Counts requires the knowledge of external linking databases. 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, in this subsection we detail a reasonable process by which an organization could prevent Linkability when Singling-out is not prevented. To represent the worst-case scenario, we assume an adversary has a location database of the entire population, $\mathbf{X}^{\ell,d}$ with $d \geq 5$.

[7]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 database was much lower than the population of South Korea (about 51 million) which implies this analysis is a conservative lower bound of privacy.

3.4.1. *Location Coarsening.* 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 database $\mathbf{X}^{\ell,d}$ 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 (de Montjoye et al. [2013]) 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 2 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 database 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 database of protected latitude and longitude coordinates.

3.4.2. *Aggregating to Counts.* When Singling-out is not prevented, there exists some count equal to one. Table 2 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 database $\mathbf{X}^{\ell,d}$ 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.

3.5. **Preventing Inference.** There is a great deal of sensitive information that could be deduced from location data. For example, Macha et al. [2020] 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 database is the probability that an individual has COVID-19.

To setup this scenario, we randomly assign 1% of individuals to have COVID-19 in our original database $\mathbf{Y}^{\ell,5}$. 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 database are $p(1\,|\,\mathbf{x},\,\mathbf{Y}^{\ell,d})$ for Location Coarsening and $p(1\,|\,\mathbf{x},\,\mathbf{C}^{d})$ for Aggregating to Counts.

3.5.1. *Location Coarsening.* To assess whether Location Coarsening prevents Inference, we assume the adversary matches the larger population database $\mathbf{X}$ to the location trajectories released in $\mathbf{Y}^{\ell,d}$. 100 simulated databases 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, $p(1 \mid \mathbf{x}, \mathbf{Y}^{\ell,d})$ for each positive individual. The left-hand side of equation (2.3) 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.

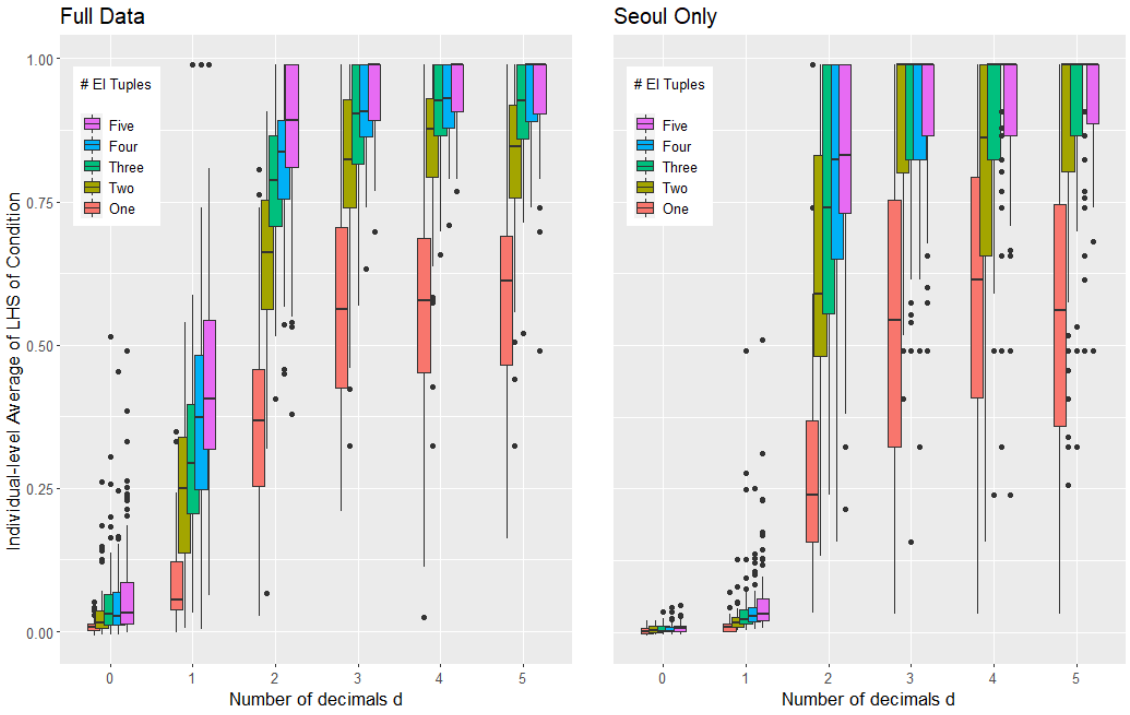

Figure 4: Location Coarsening: Boxplots of the increases in inference probabilities for COVID-19 positive individuals

Figure 4 shows boxplots of the average of the left-hand side of equation (2.3) 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 (2.3). 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.

3.5.2. *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 $p(1 \mid \mathbf{x},\, \mathbf{C}^d) = \frac{\mathbf{p^d}}{\mathbf{c^d}}$ for each tuple corresponding to 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 equation (2.3), and average across all tuples for individual m.

Figure 5 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,

$$\frac{p(1 \mid \mathbf{x},\, \mathbf{C}^d)}{p(0 \mid \mathbf{x},\, \mathbf{C}^d)} \Big/ \frac{p(1 \mid \mathbf{X})}{p(0 \mid \mathbf{X})} = \frac{.09}{.91} \Big/ \frac{.01}{.99} = 9.8 \tag{3.1}$$

for the individual with the worst-case privacy.

Aggregating to Counts is unlikely to bound the odds ratio for all individuals at high levels 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 5.

3.6. **Summary of Results.** Overall, this application illustrates the difficulty of preventing Singling-out, Linkability, and Inference from a technical perspective using relatively simple methods that are commonly used in practice and referenced in legislation. 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, but greatly reduce the epidemiological value of the data by destroying its longitudinal properties.

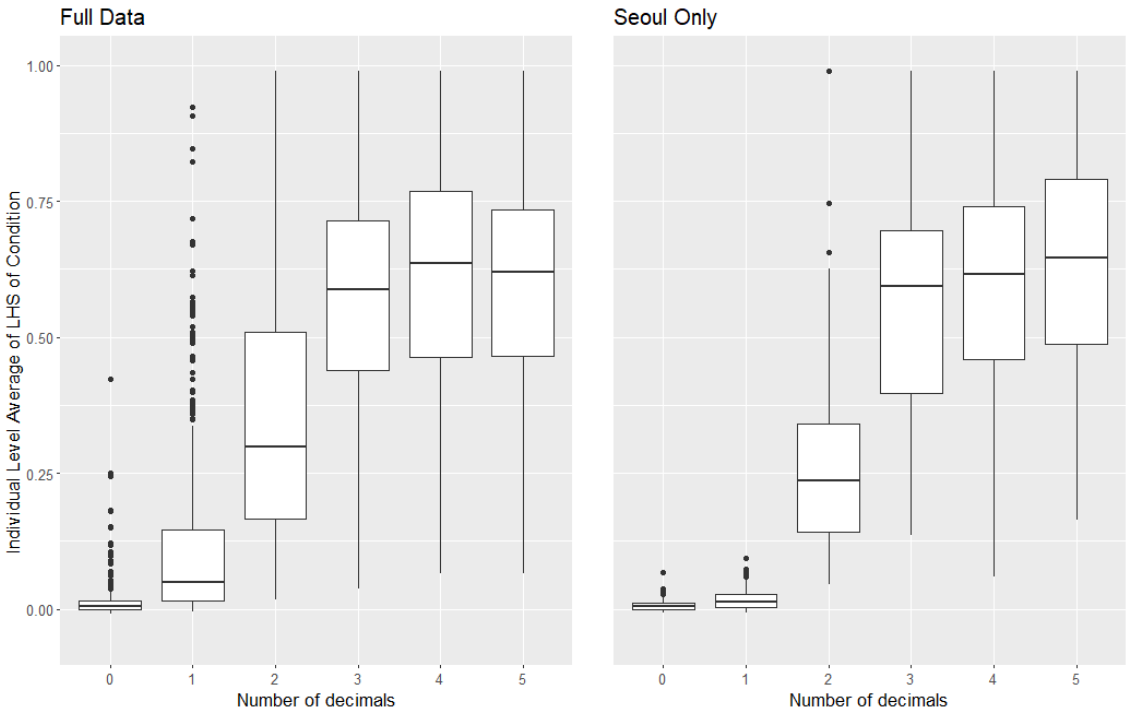

Figure 5: Aggregated Counts: Distributions of individual-level averages of left-hand side of equation (2.3) across 100 simulations for $d = 0, 1, \ldots, 5$ for Full Data (left) and Seoul (right)

4. Discussion

Using a variety of regional privacy laws, we defined a single mathematical framework to evaluate the conversion of personal data to non-personal data under the law. If protected data met all of the criteria in our framework, it was no longer considered personal data nor under the scope of the data protection law. If not, then organizations using personal data were required to invest in additional privacy and security measures (and bear the risks of that personal data) or cease using the personal data. We investigated technological solutions and showed that two logical and relatively simple statistical approaches can convert personal data to non-personal data, but came at different costs to the usefulness of the underlying data. This establishes that there needs to be a tighter marriage between legal definitions and usage, and mathematical definitions and usage.

The EDPB has also provided commentary (2020 April Guidelines, ¶ 15-16, EDPB [2020]) 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 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 database should be decided on a case-by-case basis. 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 (*i.e.*, 2-anonymity may comply with the law but provide little privacy). Further, since those criteria may vary based on the applicable data protection regulation, 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.

We also found that the protection of personal data is contextually dependent - *i.e.*, the statistical approaches we used for geolocation data are not likely to meet the three criteria using other data sources such as retail point-of-sale transactions (Schneider et al. [2018]), time series data (Nin and Torra [2009]), textual data stored from chatbots, search history logs (Toubiana and Nissenbaum [2011]), facial images (Zhou et al. [2020]; Acquisti et al. [2014]), social network data (Fienberg [2013]), or even the related contact tracing data (Dwork et al. [2020]) 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.[8] And, even with additional security and legal measures designed to protect personal data, personal data still gets out.[9] And, 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.[10] This leads to the question of whether it's 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

C. Achatz and S. Hubbard. Us vs. eu guidelines for de-identification, anonymization, and pseudonymization. *J. INTERNET L.*, 20:11, 2017.

A. Acquisti, R. Gross, and F. D. Stutzman. Face recognition and privacy in the age of augmented reality. *Journal of Privacy and Confidentiality*, 6(2):1, 2014.

[8] 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 (Roland-Holst et al. [2020])

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

[10] 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.

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.

Brazil. Lei geral de proteçao de dados (lgpd). 2018. URL https://www.lgpdbrasil.com.br/wp-content/uploads/2019/06/LGPD-english-version.pdf.

California State Legislature. California consumer privacy act of 2018. 2018. URL https://oag.ca.gov/privacy/ccpa.

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.

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: the privacy bounds of human mobility. *Sci Rep*, 3, 2013. doi: https://doi.org/10.1038/srep01376.

R. Dosselmann, M. Sadeqi, and H. J. Hamilton. A tutorial on computing t-closeness, 2019.

M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva. Anonymizing nyc taxi data: Does it matter? In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 140–148. IEEE, 2016.

DS4C. Data science for covid-19 (ds4c), 2020. URL https://www.kaggle.com/kimjihoo/coronavirusdataset.

C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, pages 1–12. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-35908-1.

C. Dwork, A. Smith, T. Steinke, and J. Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84, 2017. doi: 10.1146/annurev-statistics-060116-054123.

C. Dwork, A. Karr, K. Nissim, and L. Vilhuber. On privacy in the age of covid-19. *Journal of Privacy and Confidentiality*, 10(2), 2020. doi: 10.29012/jpc.749.

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.

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.

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.

S. E. Fienberg. Is the privacy of network data an oxymoron? *Journal of Privacy and Confidentiality*, 4(2), 2013.

S. L. Garfinkel. De-identification of personal information. *National institute of standards and technology*, 2015.

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.

N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. *2007 IEEE 23rd International Conference on Data Engineering*, 2007. doi: 10.1109/ICDE.2007.367856.

M. Macha, B. Lei, N. Z. Foutz, and A. Ghose. Perils of location tracking? personalized and interpretable privacy preservation in consumer mobile trajectories. *Working paper, Carnegie Mellon*, 2020. URL https://mmacha.github.io/resources/GeoTargetingPrivacy.pdf.

J. Nin and V. Torra. Towards the evaluation of time series protection methods. *Information Sciences*, 179(11):1663–1677, 2009.

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.

P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, 57:1701, 2009.

D. Roland-Holst, S. Evans, D. Behnke, S. Neal, L. Frölund, and Y. Xiao. Standardized regulatory impact assessment: California consumer privacy act of 2018 regulations 2019. *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.

M. L. Rustad and T. H. Koenig. Towards a global data privacy standard. *Fla. L. Rev.*, 71: 365, 2019.

M. J. Schneider, S. Jagpal, S. Gupta, S. Li, and Y. Yu. A flexible method for protecting marketing data: An application to point-of-sale data. *Marketing Science*, 37(1):153–171, 2018.

F. D. Stutzman, R. Gross, and A. Acquisti. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of privacy and confidentiality*, 4(2):2, 2013.

L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 2002.

V. Toubiana and H. Nissenbaum. An analysis of google log retention policies. 2011.

S. 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/.

Z. Tu, F. Xu, Y. Li, P. Zhang, and D. Jin. A new privacy breach: User trajectory recovery from aggregated mobility data. *IEEE/ACM Transactions on Networking*, 26(3):1446–1459, 2018.

Verizon. Verizon 2019 data breach investigations report (dbir), 2019.

J. Zhang, X. Xiao, and X. Xie. Privtree: A differentially private algorithm for hierarchical decompositions. In *Proceedings of the 2016 International Conference on Management of Data*, pages 155–170, 2016.

Y. Zhou, S. Lu, and M. Ding. Contour-as-face framework: A method to preserve privacy and perception. *Journal of Marketing Research*, page 0022243720920256, 2020.