Balancing Utility and Privacy of High-Dimensional Archives
Datasets: Mobile Phone Metadata

by

Alejandro Noriega Campero

Submitted to the Institute for Data, Systems, and Society in partial fulfillment of the requirements for the degree of Masters of Science in Technology and Policy at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY September 2015

© Massachusetts Institute of Technology 2015. All rights reserved.
Balancing Utility and Privacy of High-Dimensional Datasets: Mobile Phone Metadata

by

Alejandro Noriega Campero

Submitted to the Institute for Data, Systems, and Society on Aug 7, 2015, in partial fulfillment of the requirements for the degree of Masters of Science in Technology and Policy

Abstract

Large-scale datasets of human behavior have the potential to fundamentally transform the way we develop cities, fight disease and crime, and respond to natural disasters. However, understanding the privacy of these data sets is key to their broad use and potential impact, for these consist of sensitive information such as citizens' geo-location. Moreover, recent research has shown adversarial methods that successfully associate sensitive information in the datasets to individuals, even under pseudonymization of all personal identifiers.

This thesis conceptualizes, relates, and generalizes salient methodologies for disclosure analysis of pseudonymized data that have been developed in the last two decades, such as: k-anonymity, t-closeness, and unicity. Data at the core of the so-called "big data" revolution is fundamentally high-dimensional. We show implications of high-dimensionality as paradigmatic to modern disclosure analysis. Consequently, we propose and analyze a methodological framework that couples information-theoretic concepts from t-closeness and δ-disclosure with the partial adversarial knowledge model introduced by unicity [1] [2], as well as its possible extensions. The various methodologies were applied and compared on a large dataset of mobile phone records (CDRs), where results empirically showed ordinal equivalence among unicity measures and information distance measures EM-disclosure and KL-disclosure. Advantages of the proposed framework are highlighted, and future research avenues identified.

We also investigate the tradeoff between data privacy and data usefulness related to mobile phone metadata (CDRs) and its real-world applications. On the disclosure side, four spatio-temporal points were enough to identify uniquely +95% of individuals, at a [ZIP code, 1 hour] spatiotemporal granularity – consistent with main results in the literature. As the dataset was coarsened in space and time, the ratio (unicity) decreased to values below 0.2% for data specified at [District, 1 week] granularity or lower. We confirmed the existence of a utility-privacy tradeoff for the
10 experts surveyed for this study, i.e., a positive relationship between reidentification risk and data utility. However, Pareto analysis revealed that several granularity levels (generalization profiles) are Pareto-suboptimal, thus the tradeoff is not strict. Non-strictness implies that not all privacy gains entail utility loss, and conversely, not all utility gains entail privacy loss. Results thus suggest that data policy decisions should rest on an understanding of the underlying privacy-utility tradeoff, as inefficient policies can otherwise be implemented, unnecessarily incurring in privacy or utility losses. Lastly we show that, due to ordinal equivalence tested on the CDR dataset, Pareto properties are preserved and thus these results on the utility-privacy tradeoff are invariant to assessing disclosure by information distance measures such as EM-disclosure and KL-disclosure.

This work contributes to shed light on the privacy and utility tradeoff inherent to high-dimensional datasets of large societal systems. Its results and methodology are relevant for actors in both academic and policy domains, and germane as society engages in debate over technological and legal frameworks for potentially ubiquitous data generation and use.

Thesis Supervisor: Alex 'Sandy' Pentland
Title: Professor
Acknowledgments

This thesis would not have been possible without the thoughtful inputs and collaboration of Yves de Montjoye, Oren Lederman, Alex Rutherford, and Mila Romanoff; nor without the stimulating intellectual environment centered at Sandy Pentland’s Human Dynamics Laboratory.

A Pamen y Carlos, que por su cariño, ejemplo y educación soy quien soy.
## Contents

1 Introduction ................................. 13
   1.1 Usefulness of Mobile Phone Metadata ............................................. 14
   1.2 Privacy Risk in Pseudonymized Mobile Phone Metadata ..................... 15
   1.3 Research Questions and Structure of this Thesis .............................. 16

2 Disclosure Analysis of Pseudonymized Datasets ................................. 19
   2.1 Basic Concepts and Vocabulary ..................................................... 19
   2.2 Disclosure Guarantees and Measures .............................................. 22
   2.3 Privacy-Preserving Techniques ..................................................... 24

3 Towards Disclosure Analysis of High-Dimensional Data ......................... 27
   3.1 Modern Context for Disclosure Analysis: High-Dimensional Data .......... 27
      3.1.1 Sparsity ............................................................................. 28
      3.1.2 Fundamental Implications: Sensible Attributes as Quasi Identifiers and Inapplicability of Disclosure Guarantees ...................... 29
   3.2 From K-Anonymity to Unicity: Addressing High-Dimensionality .......... 32
      3.2.1 Sparsity and Inapplicability of General K-Anonymity Guarantees 32
      3.2.2 K-Anonymity to Unicity: Modeling Partial AK ........................... 37

4 An Information-Theoretic Framework for Disclosure Analysis of High-
   Dimensional Data ......................................................... 41
   4.1 From Unicity to Information Distance .............................................. 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1.1</td>
<td>Unicity to K-Disclosure: Deterministic to Probabilistic Re-Identification</td>
<td>42</td>
</tr>
<tr>
<td>4.1.2</td>
<td>K-Disclosure to EM-Disclosure: Identity Disclosure to Attribute Disclosure</td>
<td>44</td>
</tr>
<tr>
<td>4.2</td>
<td>An Information-Theoretic Framework for Disclosure Analysis of HD Data</td>
<td>47</td>
</tr>
<tr>
<td>4.3</td>
<td>Adversary Knowledge Assumptions</td>
<td>52</td>
</tr>
<tr>
<td>4.4</td>
<td>Empirical Results on Large-Scale Mobile Phone Metadata</td>
<td>55</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Ordinal Equivalence of Unicity and Information Distance Measures of Disclosure</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>Balancing Data Utility and Privacy: Informing Policy Formulation</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>Study Design</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>Results</td>
<td>64</td>
</tr>
<tr>
<td>6</td>
<td>Concluding Remarks</td>
<td>69</td>
</tr>
<tr>
<td>6.1</td>
<td>Disclosure Analysis</td>
<td>69</td>
</tr>
<tr>
<td>6.2</td>
<td>The Utility and Privacy Tradeoff</td>
<td>72</td>
</tr>
</tbody>
</table>
List of Figures

1-1 Example of a CDR record ................................................................. 14

2-1 (a) Published patients table. (b) Generalized patients table, where three equivalent classes with three records each are formed, based on generalized quasi-identifier value combinations. ........................................ 21

3-1 Major Latin American metropolis and its 56 districts for mobility analysis using mobile phone records (CDRs). ................................................................. 29

3-2 (a) Probability of k-anonymity of dataset $D$, $\@k = 2$ (in green); probability of any single record $i$ to be unique in $D$ (in blue); and their evolution as functions of dimensionality $d$. (b) Zoom on phase transition, range $d \in [15,20]$. ................................................................. 35

3-3 (a) Latin American metropolis at district spatial granularity. (b) Generalized down to 4 quadrants (North, South, East and West) to achieve k-anonymity $@k=2$. ................................................................. 37

4-1 Illustration of prior distribution $P(X_i^j)$, posterior distribution $P(X_i^j|A_i)$, and difference between the two as quantitative measure of disclosure. 48
Comparison of unicity and information distance-based measures of disclosure over a grid of spatiotemporal generalization levels. *EM-disclosure* stands for earth mover disclosure, and *KL-disclosure* stands for Kullback Leibler disclosure. On the x-axis are the combinations of spatiotemporal generalization, where $D$ denotes per-district and $M$ per-municipality spatial granularities (see table 4.2). Spatiotemporal generalization combinations are arranged on the x-axis in descending order with respect to unicity $\beta = 4$ values (which coincides with the order of values on each of the other 4 series).

Spatial generalization levels of the metropolitan area: (a) Municipality, (b) Districts, (c) ZIP codes.

Results of utility and disclosure risk assessments per generalization profile. The x-axis denotes data utility and the y-axis denotes disclosure risk. Each point represents the privacy and utility scores of a generalization profile. Note that utility axis $x$ is reversed for presentation purposes, so that minimization is preferred on both objectives; thus a preferred generalization profile is one that is closer to the southwest corner of the plots. Error intervals for unicity estimates are of order $O(.0001)$. (a) Transportation systems development application domain; (b) Infectious disease prevention and control application domain.
List of Tables

4.1 On the left, simple mobility data example, where cell $ij$ encodes presence of mobile user $i$ in spatiotemporal bin $j$. Number of users is $n = 4$, and dimensionality (number of spatiotemporal bins) is $d = 5$. On the right, basic concepts computed for each user, where $A_i$ is adversary knowledge for user $i$, $E_i$ is the equivalent class of user $i$, and $k$ is the size of $E_i$. ............................................. 43

4.2 Grid of spatiotemporal granularity levels over which $EM$-disclosure, $KL$-disclosure and unicity were computed. ................................. 55

5.1 Grid of spatiotemporal granularity levels over which $D$ was generalized, and data utility and disclosure risk assessed. ................................. 64

5.2 Unicity scores per generalization profile. ................................. 65
Chapter 1

Introduction

Large-scale datasets of human behavior have the potential to fundamentally transform the way we develop cities, fight disease and crime, and respond to natural disasters. However, understanding the privacy of these data sets is key to their broad use and potential impact, for these consist of sensitive information such as citizens’ geo-location. Moreover, recent research has shown adversarial methods that can successfully associate sensitive information in the datasets to individuals, even under pseudonymization of all personal identifiers.

In this work, we consider mobile phone metadata as a paradigmatic example of what is colloquially referred to as "big data". Due to its high granularity, high dimensionality, passive data generation process\textsuperscript{1}, and high potential value; mobile phone metadata represents most characteristic features of data types at the core of the so called "data revolution" – such as GPS, web browsing, and financial behavior data – in terms of both potential utility and privacy dimensions.

\textsuperscript{1}As opposed to data generated by surveys, for example, where subjects are explicitly asked to provide information.
1.1 Usefulness of Mobile Phone Metadata

Metadata is data about data. In the context of mobile phone usage this represents a record that a call was made (including a time stamp and a location with precision determined by the location of cell towers) but no information on the content of the call itself. Mobile Phone Metadata is most commonly referred to as CDRs (Call Detail Records)\(^2\). Figure 1-1 shows an example of the call detail record of a phone call.

Relevant characteristics of CDRs are: 1) the caller and receiver identities are pseudonymized, i.e., names and phone numbers are replaced by anonymous codes (called "hashes"); and 2) the geographic location of towers used for each communication provide an approximation of the user’s location.

CDRs are a particularly relevant data source for development purposes. CDRs are standard and generated by telecommunication companies on an ongoing basis. Moreover, handsets and airtime are becoming cheaper, leading to increased penetration and representativity, which by 2013 approached 89% in developing countries and 96% globally\(^{16}\).

There are several ways in which the location information in mobile phone records can be analyzed and used. Two of such uses are: population density and population mobility mapping. This information in turn has valuable applications in domains relevant and diverse as disaster response on earthquakes\(^5\) and floodings\(^{23}\), epidemics analysis of malaria\(^{15}\) and influenza\(^{10}\) outbreaks, socio-economic and poverty mapping in both the developed\(^9\) and developing worlds\(^{14}\), and transportation systems

\(^2\)Note that for a precise terminology, one must distinguish, for example, CDRs from DDRs (Data Detail Records), where the latter refers to records of mobile data transactions. However, most often in use the former generalizes the latter.
1.2 Privacy Risk in Pseudonymized Mobile Phone Metadata

Psuedonoymized metadata, and CDRs in particular, is metadata in which all personal information such as names, addresses and telephone numbers are replaced with unique and anonymous identifiers, most commonly generated by deterministic and standardized hashing functions which may not be reverse-engineered. This means that the sequence of records (calls) created by each individual continue to be associated with one another, but no explicit connection can be made between these records and the actual identity of the individual generating them.

In the past, privacy provided by pseudonymization coupled with institutional non-disclosure agreements (NDAs) has served as the basis for allowing controlled sharing of large CDR datasets. However, recent research has shown adversarial methods that successfully associate sensitive information in the datasets to individuals, even under pseudonymization of all personal identifiers.

A seminal study on reidentification of CDRs used mobility data from 1.5 million mobile phone subscribers in a small western country, where the location of an individual was specified hourly with a spatial resolution equal to that given by the carrier’s antennae, to demonstrate that just four random spatio-temporal points were enough to uniquely identify 95% of individuals in the CDR database [1]. Furthermore, the study showed that data can be coarsened, or made less granular, in order to reduce the chance of personal identification. This coarsening, more properly named spatial or temporal generalization, is one of several privacy-preserving techniques. As will be presented in this thesis, privacy-preserving techniques (PPTs) constitute the fundamental control variable that allows policy to regulate and tradeoff between data usefulness and privacy.
An active debate in academia and policy domains followed these first published results. The debate focused on methodologies for de-identification of datasets, such as k-anonymity [20], the extent to which these offer protection against potential attackers [4, 8], as well as on the likelihood of potential attackers procuring the additional knowledge (adversary knowledge) required to reidentify individuals in the pseudonymized database [13].

1.3 Research Questions and Structure of this Thesis

The driving motivation of this thesis is to better understand the tradeoff between data usefulness and data privacy in its modern ubiquitous, high-dimensional context; and to improve methodological frameworks and acumen for decision-making in public policy, corporate policy, and even the personal sphere. Methodologies for assessing disclosure risk are central towards this end. Chapters 2, 3, and 4 focus on disclosure analysis methodologies. Chapter 5 presents research results on the tradeoff between usefulness and privacy of mobile phone metadata.

Chapter 2 condenses and lays out the basic conceptual framework, vocabulary, and results that have been developed around disclosure analysis of datasets in the last two decades. These include basic concepts such as quasi-identifiers, equivalent classes, identity disclosure, and attribute disclosure; developments in disclosure measures such as k-anonymity, l-diversity and t-closeness; and privacy-preserving techniques (PPTs) such as spatiotemporal generalization, sampling and random noise addition.

Data at the core of the so called "big data" revolution is fundamentally high-dimensional. Chapter 3 analyzes fundamental new challenges that high-dimensional data presents for disclosure analysis, how previous methodologies outlined in section 2.2 are inadequate or incomplete in the modern context, and the virtues and limitations of methodologies that in the past few years have begun addressing such challenges.
Chapter 4 proposes and discusses a framework for disclosure analysis that integrates virtues of the high-dimensional adversary knowledge model introduced by unicity [1] [2], and its possible extensions, with the robust information-theoretic framework for disclosure analysis introduced by prior privacy literature – such as $t$-closeness and $\delta$-disclosure – in the context of low-dimensionality. Section 4.1 sequentially builds disclosure measures of $k$-disclosure and EM-distance as natural generalizations and developments of unicity, aimed at addressing its core limitations. Section 4.2 presents a robust information-theoretic framework for disclosure analysis, and shows EM-disclosure as a simple particular case of it. Section 4.3 analyzes and generalizes the partial adversary knowledge model introduced by work on unicity. Finally, section 4.4 tests methodologies proposed in this chapter, and compares results with those of prior methodologies such as unicity. Relevant implications and work paths forward are presented in section 6.1.

Chapter 5 moves on to investigate the tradeoff between data privacy and data usefulness associated to real-world applications that involve mobile phone metadata (referred to as CDRs). Section 5.1 describes the study design, which was conducted in collaboration between the Human Dynamics Laboratory (at the MIT Media Lab), and United Nations’ Global Pulse Initiative. We used 1 month of pseudonymized mobile phone metadata from 1.6 million users on a metropolis in the developing world. On the one hand, we developed maps of regional density and mobility patterns of the population, at various spatial and temporal granularities. We then assessed the information usefulness through in-depth interviews and a standardized questionnaire applied to a pool of 10 field experts in two domains: epidemics prevention and control, and transportation development. On the other hand, we incorporated disclosure results computed in section 4.4 over the same grid of spatiotemporal granularities, to elicit the corresponding utility-privacy tradeoff.

Section 5.2 presents main results. Consistent with previous results in the literature

---

3Global Pulse is a flagship innovation initiative of the United Nations Secretary-General on big data. For more information please visit www.unglobalpulse.org
[1], we show that four spatio-temporal points are enough to uniquely identify +95% of individuals, at a [ZIP code, 1 hour] spatiotemporal granularity. We also show that the ratio decreases as data is coarsened, and reaches values below 0.2% when data is specified at [District, 1 week] granularity, or lower. On the data privacy and usefulness tradeoff, we confirmed the existence of a utility-privacy tradeoff for experts in the study. However, Pareto optimality analysis reveals that the tradeoff is not strict. Section 6.2 presents policy implications of these results. Lastly, analysis showed that, due to ordinal equivalence and lemma 4.1 in section 4.4, chapter 5’s results on the utility-privacy tradeoff are invariant to assessing disclosure through information distance measures proposed in chapter 4.2.

Discussion and research paths forward are presented in Chapter 6.

Focus within the Data Privacy Landscape

Several complementary approaches exist to control data privacy risk while enabling its use. Some of these are: legal data use and non-disclosure agreements (DUAs and NDAs respectively); pseudonimized data publishing/sharing; precomputed queries or statistics, such as Telefonica’s Dynamic Insights4; differential privacy and restricted querying [3]; synthetic data and others. This thesis focuses on the publishing or sharing of pseudonymized microdata.

---

4 Telefonica Dynamic Insights provides information products like Smart Steps, which processes "Big Data" generated by Telefonica’s mobile network, and computes trends and patterns within the data to help understand consumer behaviour, answering questions of who, when, where, why and how, regarding consumer’s mobility patterns. Visit http://dynamicinsights.telefonica.com/blog/488/smart-steps-2 Smart Steps webpage for more information: http://dynamicinsights.telefonica.com/blog/488/smart-steps-2
Chapter 2

Disclosure Analysis of Pseudonymized Datasets

This chapter lays out a summary of the basic conceptual framework, vocabulary, and results that have been developed around disclosure analysis of pseudonymized datasets in the last two decades. Section 2.1 introduces basic concepts such as quasi-identifiers, equivalent classes, identity disclosure and attribute disclosure. Developments in disclosure measures and guarantees, such as k-anonymity, l-diversity and t-closeness, are presented in Section 2.2. Lastly, Section 2.3 describes privacy-preserving techniques (PPTs) such as spatiotemporal generalization, sampling and noise addition, and their role as key control variables that allow policy to regulate and tradeoff between data usefulness and privacy.

2.1 Basic Concepts and Vocabulary

Attribute Types

We start this section introducing basic definitions. Terminology and vocabulary follows that of the main literature published in the 21st Century on the topic – such as [19, 12, 11, 7].
When an agency publishes pseudonymized microdata, i.e., anonymous data at the individual level, it is most often done in the form of a table. Every record in the table, or dataset, has a number of attributes, such as name, telephone, address, income, health status, item purchased, time and date, etc. These attributes are classified in the following types:

- **Explicit Identifiers.** These are attributes that clearly identify individuals in the dataset, such as name, telephone, social security number, corporate ID, etc.

- **Quasi-Identifiers.** These are attributes that, when taken together, can potentially identify individuals in the dataset. Birth date, ZIP code and gender, are attributes traditionally considered quasi-identifiers.

- **Sensitive Attributes.** This are attributes that are considered sensitive, and thus potential attackers should not be able to link them to personal identities of individuals in the dataset. Health conditions, salary and amount of monetary transactions are common sensitive attributes.

**Disclosure Types**

Figure 2-1a shows a simple example of published data, where there is a simple pseudonymized ID (serial 1 to 9), Disease attribute is a sensitive attribute, and ZIP code and Age are quasi-identifiers. From it, we can appreciate how an attacker, if knowing the ZIP code and age of individual 1, would immediately identify his record, as the record has a unique combination of quasi-identifiers in the published table. This type of disclosure is called **identity disclosure**, also characterized as **reidentification**, and occurs when an attacker is able to identify uniquely the record of an individual in the dataset, and thus can associate or reidentify a pseudonymized ID. The set of information available to an attacker, in addition to the published data, is called **adversary knowledge** (ZIP code and age of individual 1 in this case).
A common technique for preventing reidentification (identity disclosure) is to generalize or coarsen information of the quasi-identifiers, so that more records coincide with the same set of quasi-identifiers. Figure 2-1b shows an example application of **quasi-identifier generalization** on the table of Figure 2-1a, where only the first 3 digits of ZIP code and the first digit of Age are published. In this case, an attacker with the ZIP code and Age information of individual 1 (adversary knowledge) is unable to reidentify uniquely, as there are three records that coincide with the quasi-identifier values.

An **equivalent class** is defined as the group of records that coincide with a set of quasi-identifier values. Figure 2-1b shows the existence of three equivalence classes, with three records each, that are formed once ZIP code and Age have been generalized.

It is relevant to note that even after quasi-identifier generalization, the sensitive attribute of individual 1 is disclosed. This happens because all records within the equivalent class of individual 1 have **Heart Disease**, therefore knowing the equivalent class is enough to know the sensitive attribute.

The above example illustrates the distinction between identity disclosure and attribute disclosure. **Identity disclosure** occurs when the pseudonymized ID is suc-
cessfully associated to a personal identity, i.e., the true identity of a record is revealed. **Attribute disclosure** occurs when a sensitive attribute is disclosed. The latter captures the essence of privacy concern on pseudonymized data publishing. The former is a sufficient, but not necessary condition for the latter.

2.2 Disclosure Guarantees and Measures

A sequence of measures of disclosure have been developed in the XXI century. The aim of such developments has been to 1) provide guidance on data treatments (most often quasi-identifier generalization) to achieve desired levels of privacy risk, and 2) provide means with which to assess disclosure or privacy risk associated to publication of pseudonymized datasets.

In early 2000’s, Samarati and Sweeney introduced the concept of **k-anonymity**. A k-anonymous dataset is one where each equivalent class has at least $k$ records. The table in figure 2-1b is 3-anonymous, in k-anonymity terminology. K-anonymity gained a lot of popularity, as it provided a simple assessment of disclosure risk, and algorithms were developed that achieve k-anonymity, given a $k$, by applying generalization of quasi-identifiers [19, 20]. However, as illustrated by figure 2-1b, the extent to which k-anonymity is able to account for and control disclosure risk of published datasets is **limited to identity disclosure**.

Several methodologies have been developed in recognition of this limitation. **1-Diversity** was proposed in 2007 [12]. The idea behind it is that, in order to avoid attribute disclosure, a dataset has to have diversity in the distribution of each sensible attribute of each equivalent class. Intuitively, if an adversary knows the equivalent class of an individual (through adversary knowledge of quasi-identifiers), then attribute disclosure will not occur if the distribution of the sensible attribute is diverse enough. Several definitions of diversity were then proposed, such as distinct-value $l$-diversity, entropy $l$-diversity, and recursive $(c,l)$-diversity [12].
Li et al. 2007 proposed the **information-distance framework** for defining and measuring disclosure. In this framework, disclosure is defined as the amount of information about sensitive attributes that an adversary gains due to publication of pseudonymized data [11, 7]. This amount is defined as the difference between the prior and posterior probability distributions of sensitive attributes, from the point of view of an adversary, and is quantified through information-theoretic measures of distance such as the **Earth Mover’s Distance** and the **Kullback-Leibler Distance** [18]. Prior adversarial knowledge is assumed to be the *common background knowledge*, i.e., the distribution of sensitive attributes over the entire population or table. This notion of information-distance was given the name of **t-closeness**.

Branches of this framework followed Li 2007. For example, Brickell and Shmatikov 2008 argued for the use of the Kullback-Leibler distance (KLD) as information distance measure, and showed the possibility of deriving bounds for attribute disclosure [7], if certain properties are met. In this work, a pseudonymized dataset is said to be **δ-disclosure-private** if the Kullback-Leibler distance between prior and posterior information of an adversary, denoted $D_{KL}(P, Q)$, is upper-bounded by a threshold $\delta$, for all equivalent classes (all combinations of quasi-identifiers). As pointed out by Li 2007, the disadvantage of this approach is that KLD cannot account for semantics higher than that of categorical variables [11], e.g., cannot account for the connectedness or contiguity in space and time of spatiotemporal segments in mobility data (e.g., CDRs), or the ordinal relationship among wage segments in economic data.

Overall, approaches developed during the first decade of the XXI century – such as k-anonymity, l-diversity, t-closeness, δ-divergence – where framed in a context where: sensitive attributes were assumed low-dimensional, and thus not themselves potential quasi-identifiers; and quasi-identifiers were assumed low-dimensional, and thus fully known by potential adversaries. In this context, the above frameworks sought to provide strong privacy guarantees, if specified data properties were met. As will be shown in Chapter 3, the modern context for disclosure analysis of pseudonymized mi-
crodata poses significant new challenges driven by high-dimensionality, characteristic of modern "big data".

2.3 Privacy-Preserving Techniques

Several data transformations have been proposed and used for abating disclosure risk in pseudonymized datasets. The most commonly used is generalization of quasi-identifiers, introduced in Section 2.1; others include sampling, noise addition, and random permutations. These are referred to as privacy-preserving techniques (PPTs).

PPTs are designed to decrease disclosure risk. However, as data is generalized, coarsened, sampled or infused with noise, its usefulness is most often eroded. Thus the form and extent to which PPTs are applied before data publication constitute the basic control variable with which policy can tradeoff between data utility and privacy. In what follows of this section some of the most relevant PPTs are described.

Generalization

The most common technique for preventing identity and attribute disclosure is to generalize or coarsen information of the quasi-identifiers, so that more records coincide with the same set of quasi-identifiers. Figure 2-1b shows a toy example of quasi-identifier generalization on the table of Figure 2-1a, where only the first 3 digits of ZIP code and the first digit of Age are published. In this case, an attacker with the ZIP code and Age information of individual 1 (adversary knowledge) is unable to reidentify uniquely, as there are three records that coincide with the quasi-identifier values. It is important to note that, as exemplified in Section 2.1, even after quasi-identifier generalization, sensitive attributes of individuals can be disclosed.

Relevant to mobile phone metadata is spatiotemporal generalization, where the geographic polygons and time intervals used to specify space and time of mobile phone communications are coarsened to hinder disclosure. Spatiotemporal general-
ization is discussed and used for analysis in chapters 3 and 5 of this thesis.

**Sampling**

Sampling is commonly used in addition to other PTTs such as generalization. The technique consists on publishing only a subset of randomly selected records from the full dataset. The rational behind sampling is to reduce disclosure risk by 1) exposing only a fraction of individuals to possible privacy attacks, and 2) to increase uncertainty to potential attackers by introducing membership uncertainty, i.e., uncertainty of weather or not the individual of interest is in the published sampled dataset. The sample fraction can be reduced as long as the desired statistical properties of interest of the data are conserved.

**Randomization**

Randomization techniques introduce random noise to a pseudonymized dataset. The general rational is to decrease accuracy of the data while preserving its desired statistical properties. Two relevant types of randomization are:

- **Random Noise Addition**: adds independent random noise to attribute values. For instance, the values of attribute Age in Figure 2-1a can be added noise so that they remain accurate only within a ±10% interval.

- **Random Permutations**: consist of random interchanges of attribute values among records in the dataset. This technique preserves the marginal distributions of attributes – as the resulting marginal is a permutation of the the original – but breaks correlations among them [21].
Chapter 3

Towards Disclosure Analysis of
High-Dimensional Data

3.1 Modern Context for Disclosure Analysis: High-Dimensional Data

The field of data privacy along the second half of the XXth century and beginning of the XXIst, was motivated by concern over disclosure of sensitive information contained in low-dimensional datasets such as census data and medical records (see for example [4, 20]). Today the advent of technologies for pervasive sensing, data storage and analysis, has allowed for unprecedented instrumentation of our societal and technical systems. The Promethean promise of the data revolution is to assist in building societies that are better in avoiding market crashes, ethnic and religious violence, corruption, reducing human death from disease outbreaks and natural disasters, control power outages, traffic congestion, etc. [22]. Chiefly, data at the core of this "big data" revolution is high-dimensional. Examples are human mobility data, financial and card transaction data, consumer behavior data, online behavior data such as social networking and web surfing, genetic data, etc.
This section analyzes fundamental new challenges that high-dimensional data presents for disclosure analysis, how previous methodologies outlined in section 2.2 are inadequate or incomplete in the modern context, and the virtues and limitations of methodologies that in the past few years have begun addressing such challenges.

3.1.1 Sparsity

Section 2.1 introduced basic concepts of attribute and disclosure types, adversary knowledge and equivalent classes. This concepts were illustrated by a simple example of medical data (Figure 2-1b), with two quasi-identifiers (ZIP Code and Age) and one sensitive attribute (Disease).

Consider now the case of human mobility patterns that can be extracted from mobile phone records (see Section 1.1). Mobility data specifies the presence of a person at a particular time and space (geographic polygon). Figure 3-1 shows a major Latin American metropolis and its partition in 56 districts. If spatial and temporal data granularity is specified at district and hourly levels, then one month of mobility data in this metropolis contains \((temporal\_bins)(spatial\_bins) = (24)(30)(56) = 40,320\) spatiotemporal bins where individuals can be present or not. Dimensionality of this dataset is 40,320.

Datasets of this high-dimensionality are **sparse**, even when containing dozens of millions of records. In the aforementioned human mobility case, the space of all possible mobility patterns \(\Omega\) is of size \(|\Omega| = (2)^{40,320}\). Let \(n\) denote the number of records in the dataset. Based on data provided by major mobile phone operator in the area, we know that \(n\) is in the order of millions of users in this metropolis (1.6M). We say the dataset is sparse because \(n \ll |\Omega|\), in this case \(O(1,000,000) \ll (2)^{40,320}\).
3.1.2 Fundamental Implications: Sensible Attributes as Quasi Identifiers and Inapplicability of Disclosure Guarantees

Records in datasets of high-dimensionality are only a sparse sample of the space of possible records Ω. Sparsity often entails that records are rather unique among themselves, i.e., the probability of two records coinciding in attribute values for several attributes is close to zero. Recent studies have demonstrated uniqueness property on large mobile phone datasets [1], card transaction datasets [2], and online movie review datasets [24]. Section 3.2 formalizes and illustrates the notion of high probability of uniqueness.

Furthermore, each of these studies show how sparsity can be exploited towards identity disclosure (reidentification) of pseudonymized records, as long as a potential attacker has access to a small set of a record’s data. The process consists of: 1) a potential attacker acquires knowledge of attribute values of a small subset of attributes
of a record, 2) the attacker verifies which of all records coincide with such subset of values, 3) if only one record coincides, then the attacker has successfully reidentified the record.

To illustrate record uniqueness induced by sparsity and the reidentification process, consider a very large database of song lyrics. The space of possible song lyrics – i.e., permutations of a bounded number of words – is extremely large. Thus, given a sequence of only 3 or 4 words from a certain song, we are likely to be able to identify uniquely which song among thousands the sequence belongs to.

**Sensitive Attributes as Quasi-Identifiers**

Sensitive attributes can function as quasi-identifiers in high-dimensional contexts. As in the case where a database contains distinct quasi-identifiers (e.g., age and ZIP code), and sensitive attributes (e.g., medical condition); all disclosure concepts introduced in Section 2.1 are meaningful and can be applied in high-dimensional contexts, where sensitive attributes function as quasi-identifiers:

- Given a subset of sensitive attribute values (quasi-identifiers), an equivalent class (EC) of records in a database can be defined;
- If the EC contains one record, identity disclosure occurs;
- Attribute disclosure occurs to the extent that the distributions of attributes within the EC differ from the distributions of attributes in the full table.

Dealing with issues associated with potential uses of sensitive attributes as quasi identifiers is core to the HD disclosure analysis techniques discussed and proposed in what follows of this thesis.

**Inapplicability of General Disclosure Guarantees**

A *k-anonymous* database is one where each equivalent class is guaranteed to have at least *k* records. Similar privacy guarantees are implicit in analogous definitions
of concepts like $l$-diversity and $t$-closeness (refer to section 2.2). Although we’ve discussed how the latter ones present fundamental advances from the former, they share in common a paradigm where: 1) sensitive attributes are low-dimensional; 2) quasi-identifiers are distinct from sensible attributes; 3) quasi-identifiers are assumed fully known by potential adversaries; and 4) based on (1), (2), and (3), strong privacy guarantees are possible, defined and pursued.

Sparsity in high dimensional databases makes the pursuit of general privacy guarantees inapplicable for two main reasons. First, models where potential adversaries know the full set of quasi-identifiers of an individual, applied to high-dimensional data, represent at best a rather trivial scenario. In HD contexts such assumption entails adversary knowledge (AK) of the full set of sensitive attributes, which is unlikely, and in which case there is no further information to be gained by a privacy attack. What follows of this thesis deals with the more general scenario characteristic of modern disclosure analysis: partial AK of sensitive attributes (quasi-identifiers). Refer to section 4.3 for analysis and discussion of AK assumptions of the different methodologies.

Second, meeting $k$-anonymity under full AK knowledge assumption, i.e., requiring that no record $i$ is unique ($|EC_i| > 1 \forall i$), requires generalization of sensitive attributes to extents that erode most data usefulness. This happens due to sparsity properties\(^1\) and is the topic of section 3.2.1. Alternatively, more modern disclosure analysis methodologies, such as unicity, shift focus from aiming at guaranties to measuring and controlling disclosure likelihood, under the context of partial AK. This likelihood paradigm for disclosure control in pseudonimized datasets can be compared with car security systems, where the design goal is not to develop guaranteed systems for car theft prevention under all circumstances, but ones that significantly hinder theft in scenarios where attackers devote reasonable amounts of resources.

In short, general disclosure guarantees are unfeasible due to HD sparsity proper-

\(^1\)A formal analysis of sparsity as rationale for inapplicability of k-anonymity guaranties is developed in section 3.2.
ties. Partial disclosure guaranties are contingent upon partial adversary knowledge conditions, for which appropriate analytical methodologies are needed.

3.2 From K-Anonymity to Unicity: Addressing High-Dimensionality

Recent literature has recognized the fundamental high-dimensional (HD) nature of modern data, such as CDRs, consumer card transaction data, and online behaviour data. It has also recognized the novel challenges for privacy that it entails, and developed techniques for modern disclosure analysis. One such measure that has gained significant attention in the past couple of years is unicity [1] [2]. This section first formalizes and illustrates rationales for inapplicability of k-anonymity on HD data. Then it introduces unicity, and shows it as an evolution of k-anonymity concepts for disclosure analysis in high-dimensional contexts. Finally, shortcomings of unicity are exemplified, as preamble and rationale for methodologies proposed and analyzed in section 4.

3.2.1 Sparsity and Inapplicability of General K-Anonymity Guarantees

Meeting k-anonymity on high-dimensional (HD) pseudonimized datasets – i.e., guaranteeing no record $i$ is unique: $|EC_i| > 1 \forall i \in [1, n]$ – requires generalization of quasi-identifiers to extents that erode most data usefulness. Loss in data value is particularly acute as quasi-identifiers generalized are the sensitive attributes themselves (e.g., spatiotemporal mobility data, rather than complementary socio-demographics; see section 3.1.2). This section addresses the question: how much information loss does meeting the general k-anonymity requirement entail in HD contexts? Working on a generic binary data matrix, we analyze how much dimensionality reduction must
be imposed through information generalization to achieve general \textit{k-anonymity}, even at the laxest privacy threshold \(k = 2\), and datasets where records \((n)\) are counted in the order of millions.

Let \(D^{n \times d}\) be a generic binary data matrix, with \(n\) rows and \(d\) columns. Denote each row vector as \(X_i \in D\) for \(i \in [1, n]\), and each column vector as \(X^j \in D\) for \(j \in [1, d]\). Note that any categorical data can be reduced to such a binary expression. In order for this matrix to be \textit{k-anonymous}, it is required to ensure there is no record with a unique combination of quasi-identifiers in the dataset (see section 2.2). Moreover, in contexts where \(d\) is high (say, \(d > 20\)), sensible attributes can be used as quasi-identifiers (see section 3.1.2). Thus:

Matrix \(D\) is \textit{k-anonymous} at \(k = 2\), if every record \(X_i \in D\) exactly matches at least one other record, i.e.: \(\forall i \exists i' \text{ s.t. } X_i = X_{i'}\), where \(i \neq i'\), and \(i, i' \in [1, n]\).

Let us assume a general case where data records in \(D\) are drawn from a population such that binary data cells \(X^j_i\) are i.i.d. draws from distribution \(X^j_i \sim \text{Bernoulli}(p)\). We are interested in the probability of matrix \(D^{n \times d}\) being 2-anonymous (laxest privacy guarantee), in \textit{k-anonymity} sense. We have that, for any \(i \neq i'\) and \(i, i' \in [1, n]\):

\[
P[X_i = X_{i'}] = \prod_{j=1}^{d} P[X^j_i = X^j_{i'}] = \prod_{j=1}^{d} p^2 + (1-p)^2
\]

\[
P[X_i = X_{i'}] = (p^2 + (1-p)^2)^d
\]

Let \(|EC_i| \in [1, n]\) denote the size of the equivalent class (EC) of record \(X_i\). Then \(|EC_i| - 1\), the number of records that exactly match record \(i\), follows a binomial distribution \(|EC_i| - 1 \sim \text{Binomial}(n, P[X_i = X_{i'}])\). Thus, the probability for any given record \(X_i\) to be unique in \(D\), denoted by \(P[|EC_i| = 1]\), is given by:

\[
P[|EC_i| = 1] = P[|EC_i| - 1 = 0] = \binom{n-1}{0} P[X_i = X_{i'}]^0 \left(1 - P[X_i = X_{i'}]\right)^{n-1-0}
\]
\[ P[|EC_i| = 1] = (1 - P[X_i = X_i'])^{n-1} \]  

(3.1)

Note that \( P[|EC_i| = 1] \) can also be understood as the expected number of unique records in \( D \).

Finally, we have that the probability of matrix \( D^{n \times d} \) being \( 2 \)-anonymous, denoted by \( P[D_{k=2}] \), as a function of any given record \( X_i \) being unique, is given by:

\[ P[D_{k=2}] = (1 - P[|EC_i| = 1])^n \]

(3.2)

Consider the case where \( p = .5 \). Substituting in the above equations we get \( P[|EC_i| = 1] \) and \( P[D_{k=2}] \) as a function of \( n \) and \( d \):

\[ P[X_i = X_{i'}] = (.5^2 + (1 - .5)^2)^d = .5^d \]

\[ P[|EC_i| = 1] = (1 - .5^d)^{n-1} \]

(3.3)

\[ P[D_{k=2}] = (1 - (1 - .5^d)^{n-1})^n \]

(3.4)

**Insight and Analysis**

Now, the question is: is it possible for large HD datasets, e.g., of dimensionality \( d \) and size \( n \) similar to those of CDR datasets, to meet \( k \)-anonymity? To what extent must information generalization decrease dimensionality to achieve \( k \)-anonymity?

Consider the mobile phone records (CDRs) dataset described in section 3.1.1. Dimensionality of this dataset is \( d = 40320 \) – i.e., the number of bins in space and time where a person can be present or not, given an hourly temporal granularity and per-district spatial granularity. There are \( n = 1.6 \) million different users in the dataset. Plugging values \((n, d) = (1.6 \times 10^6, 40320)\) into equations 3.3 and 3.4, we get that \( P[|EC_i| = 1] = (1 - .5^d)^{n-1} \approx 1 \), and \( P[D_{k=2}] = (1 - (1 - .5^d)^{n-1})^n \approx 0 \). In words, the probability of any single record to be unique in \( D \) is approximately 1, and the probability of \( D \) being \( 2 \)-anonymous is approximately 0 (both approximations
Figure 3-2: (a) Probability of \textit{k-anonymity} of dataset \( D \), \( @k = 2 \) (in green); probability of any single record \( i \) to be unique in \( D \) (in blue); and their evolution as functions of dimensionality \( d \). (b) Zoom on phase transition, range \( d \in [15, 20] \).
hold even at precisions smaller than $10^{-100}$).

Uniqueness and the possibility of \textit{k-anonymity} in $D$ are driven by sparsity, which depends on the relative magnitudes of $d$ and $n$. Figure 3-2 shows the degree of generalization that must be applied to $D$ in order to allow for \textit{2-anonymity} in \textit{k-anonymity} sense. Generalization decreases $d$ while maintaining $n$ constant. In the case of mobility patterns from CDR data, spatial and temporal granularity are decreased, e.g., time and place of an SMS are registered only at daily and per-municipality granularities.

Figure 3-2 provides relevant insight in that:

- Data must be generalized to dimensionality levels as low as 16 to allow for \textit{2-anonymity} of $D$.

- For any dimensionality levels above 30 the expected fraction of unique records in $D$, i.e., $P[|EC_i| = 1]$, is 100%.

- Both $P[|EC_i| = 1]$ and $P[D_{k=2}]$ present stark phase transitions from 0 to 1 values (and vise versa) as dimensionality $d$ is changed. This characteristic implies that there are narrow threshold intervals on $d$ values, across which \textit{k-anonymity} is possible or nearly impossible.

In the context of the CDR dataset described, generalization down to $d = 16$ means that, for example, temporal granularity would be lowered to a weekly level and spatial granularity limited to a four quadrant – North, South, East and West – specification ($\text{temporal} \times \text{spatial} = 16$). Such coarsening, illustrated by figure 3-3, implies information loss that erodes most data usefulness. Furthermore, if dimensionality is raised to anything higher than 30, for instance by increasing temporal granularity to bi-weekly specification (i.e., distinguishing weekends from weekdays, $\text{temporal} \times \text{spatial} = 8 \times 4 = 32$), then the expected fraction of unique records in $D$ is 100%.

Results of this section where tested robust to varying parameter $p (X_i^j \sim \text{Bernoulli}(p))$. Further empirical research can apply generic analysis here conducted on relevant data.
types such as mobile phone records, card transaction data, and online behavior data; by plugging-in corresponding domain-specific joint distributions $P[X_i]$.

### 3.2.2 K-Anonymity to Unicity: Modeling Partial AK

*K-Anonymity* guarantees that every record $i$ in dataset $D$ has an equivalent class $EC_i$ of at least $k$ records. Similarly, concepts of *t-closeness* [11] and *δ-disclosure* [7] characterize a database with thresholds $t$ or $δ$ over which no equivalent class in $D$ discloses information, assuming full adversary knowledge $A_i$ of quasi-identifiers (see section 2.2). As discussed in sections 3.2 and 3.1.1, this paradigm is inapplicable in the context of high-dimensional data, sparsity, and partial adversary knowledge (AK). Rather, methodologies proposed in sections 4.1.2 and 4.2 of this thesis seek to borrow information-theoretic concepts from *t-closeness* and *δ-disclosure* to build disclosure measures that can be coupled with the partial AK model introduced by unicity [1] [2], as well as its possible extensions.

High-dimensional (HD) databases are sparse, and due to sparsity, sensitive attributes themselves can function as quasi-identifiers (see section 3.1.1). Models where potential adversaries know the full set of quasi-identifiers of an individual are in HD
contexts inapplicable or a rather trivial scenario, for two main reasons. First, such assumption implies AK of the full set of sensitive attributes, which appears unlikely, and in which case there is no further information to be gained by a privacy attack. Second, as shown in section 3.1.2 for *k-anonymity*, attempting to meet disclosure guarantees under full AK knowledge assumptions would require generalization of sensitive attributes to extents that erode most data usefulness.

Therefore we must consider scenarios where only partial knowledge is available to a potential attacker. For instance, cases where a potential attacker knows the home and work location of an individual *i*, but not *i*'s full mobility behavior. Recent work on *unicity* \cite{1} \cite{2} pioneered modeling partial AK. Taking a closer look to *unicity*'s methodology we can conceptualize its modeling of re-identification as estimating:

\[
P(ID) = P(ID|A)P(A)
\]  \hspace{1cm} (3.5)

where *ID* denotes identity disclosure and *A* stands for adversary knowledge.

*Unicity*'s modeling of partial adversary knowledge, *P(A)* term in equation 3.5, is its main advancement from *k-anonymity*. Analysis and discussion of this characteristic is conducted in section 4.3. In regards to term *P(ID|A)* of equation 3.5, *unicity* follows *k-anonymity*'s methodology in that privacy assessment is based upon the size of equivalent classes \(|EC_i|\) of records in *D*. In particular, *unicity* = 1 − \(F_{k=2}\), where \(F_{k=2}\) denotes the fraction of records in *D* that are 2-anonymous in *k-anonymity* sense. However, it is relevant to note that *unicity* introduces a paradigm shift from *k-anonymity*'s: it seeks an assessment of disclosure risk rather than to provide privacy guarantees; thus, rather than requiring all records to be 2-anonymous, it computes the fraction of records which are.

*Unicity* however presents a number of limitations, which motivate extensions and development of methodologies discussed in chapter 4, relevantly:

- Its limitation to deterministic re-identification, addressed in section 4.1.1.
- Its limitation to identity disclosure (re-identification), as opposed to the more general attribute disclosure; addressed in section 4.1.2.

- Its non-extensibility to assess the effects on disclosure risk of other relevant privacy protecting techniques (PPTs), such as sampling and random noise addition.

- Its non-extensibility to conduct disclosure analysis of data with higher semantics than purely categorical data, e.g., ordinality of income segments, continuity of space-time variables, or distance along a social network.
Chapter 4

An Information-Theoretic Framework for Disclosure Analysis of High-Dimensional Data

4.1 From Unicity to Information Distance

In this section we develop a sequence of disclosure measures that elaborate on unicity’s methodology. Recall disclosure equation 3.5 from section 3.2.2: \( P(ID) = P(ID|A)P(A) \). Methodologies in this section adopt unicity’s partial adversary knowledge model \( P(A) \) (discussed in section 4.3), but elaborate on and generalize disclosure component \( P(ID) \), sequentially addressing each of unicity’s aforelisted limitations. Throughout this section concepts are illustrated in the domain of mobility data, such as that extracted from mobile phone records (see section 1.1).

This section assumes basic understanding of concepts in sections 2.1, 2.2 and 3.2.

Basic Concepts and Notation

- Let \( D^{n \times d} \) be a binary data matrix, with \( n \) rows and \( d \) columns. Denote each row vector as \( X_i \in D \) for \( i \in [1,n] \), and each column vector as \( X_j \in D \)
for \( j \in [1, d] \). Note that \( D \) can represent categorical data, as any categorical variable can be reduced to its binary expression.

- Let set \( A_i \) denote partial adversary knowledge of record \( i \), where elements \( \{ j \mid j \in A_i, j \in [1, d] \} \), represent the attributes for which an adversary knows \( i \)'s information.

- We denote equivalent class \( E_i \), or \( E_i^{A_i} \) more explicitly, as the equivalent class of \( i \) given adversary knowledge \( A_i \). Any two records in \( E_i \) are equal in terms of attributes \( A_i \). More formally: \( X_i^j = X_i^{j'} \ \forall j \in A_i \) and \( \forall i' \in E_i \).

- Finally, let's denote \( u_k \) as the number of records \( i \) whose equivalent class \( E_i \) is of size \( k \), i.e., \( u_k = \{ i \mid |E_i| = k \} \), where \( k \in [0, n] \).

A Simple Working Example on Mobility Data

Consider mobility data extracted from mobile phone records described in section 3.1.1. Let data attributes \( j \) represent a spatiotemporal intervals or bins, e.g., a combination \((\text{hour } x, \text{district } y)\) (refer to section 3.1.1). Let record \( X_i \) be the CDR record of mobile phone user \( i \); and data cells \( X_i^j \) represent the presence of user \( i \) in spatiotemporal bin \( j \), where \( X_i^j = 1 \) encodes presence of user \( i \) in spatiotemporal bin \( j \). Table 4.1 shows a simple data example of this mobility data, and illustrates the aforelisted basic concepts and notation.

### 4.1.1 Unicity to K-Disclosure: Deterministic to Probabilistic Re-Identification

As described in section 3.2.2 unicity estimates the percentage of records in a dataset whose equivalent class \( E_i \) contains only record \( i \) itself. As such, unicity is defined as:

\[
\text{unicity}(D|A) = \frac{u_1}{n} \tag{4.1}
\]
Table 4.1: On the left, simple mobility data example, where cell $ij$ encodes presence of mobile user $i$ in spatiotemporal bin $j$. Number of users is $n = 4$, and dimensionality (number of spatiotemporal bins) is $d = 5$. On the right, basic concepts computed for each user, where $A_i$ is adversary knowledge for user $i$, $E_i$ is the equivalent class of user $i$, and $k$ is the size of $E_i$.

Where $u_1 = \left| \left\{ i \mid |E_i| = 1 \right\} \right|$, $D$ is the data, and $A$ is the set of adversary knowledge.

One limitation of unicity as a measure of disclosure is that it only accounts for cases where re-identification happens with absolute certainty. That is, cases where an adversary seeking to reidentify $i$ uses $A_i$ to narrow down an equivalent class $E_i$ with $|E_i| = 1$, and thus can deterministically match a known identity with record $i$.

In large datasets such as mobile phone records or card transaction records, even when $|E_i| > 1$, an adversary might still use $A_i$ to narrow down its search for $i$ to a small number of records. When $|E_i| \ll n$, relevant adversarial information gains can be ignored by unicity. A natural extension of unicity that accounts for adversarial information gains towards non-deterministic re-identification is $k$-disclosure, defined as:

$$ k\text{-disc}(D|A) = \frac{1}{n} \sum_{k=1}^{n} \frac{u_k}{k} $$

Expanding equation 4.2:

$$ k\text{-disc}(D|A) = \frac{1}{n} \sum_{k=1}^{n} \frac{u_k}{k} = \frac{1}{n} \sum_{k=1}^{3} \frac{u_k}{k} = \frac{1}{n} u_1 + \frac{1}{2n} u_2 + \frac{1}{3n} u_3 + \ldots $$

Equation 4.3 shows $k$-disclosure as a simple generalization of unicity, where the first term of the sum corresponds to unicity's equation 4.1. As unicity, $k$-disc \in
We can see from expansion 4.3 how \textit{k-disclosure accounts for adversarial information gains useful towards non-deterministic re-identification}. For example, the second term in the sum accounts for those records which could be narrowed down to an equivalent class of only two records by a potential adversary \((|E_i| = 2)\); so does the \(k\)th term for records with \(|E_i| = k\).\footnote{The \(k\) number of sum terms can be limited, if appropriate}

Applied to the simple data example of table 4.1, \textit{unicity} = \(\frac{\text{un}}{n} = \frac{1}{4} = .25\), and \textit{k-disc} = \(\frac{1}{14} + \frac{1}{24} + \frac{12}{34} = \frac{13}{24} \approx .54\). Note that, for instance, in the case of user 2, a potential adversary uses \(A_2\) to increase the probability of correctly re-identifying user 2 from \(\frac{1}{n}\) to \(\frac{1}{|E_2|} = \frac{1}{2}\). \textit{K-disclosure} captures this difference in probabilities, unaccounted for by \textit{unicity}. Yet, as discussed in the following section 4.1.2, its focus remains limited to identity disclosure, thus ignoring the more general attribute disclosure.

\subsection*{4.1.2 K-Disclosure to EM-Disclosure: Identity Disclosure to Attribute Disclosure}

\textit{K-disclosure} accounts for adversarial information gains useful for non-deterministic re-identification. However, it preserves core limitations of \textit{unicity} as measure of disclosure: its focus on re-identification – i.e., identity disclosure – rather than attribute disclosure (section 2.1). This is the same core limitation of \textit{k-anonimity}, pointed out and analyzed in the context of low-dimensional data by [12], [11] and [7] in the first decade of our century (section 2.2).

Take the simple example on table 4.1. If we know that user 2 was present in spatiotemporal bin 1 (i.e., \(A_2 = \{1\}\)), then we know that user 2’s record must be in either row 1 or row 2 (i.e., \(E_2=\{1,2\}\)). \textit{K-disclosure} models this a case as a \(\frac{1}{2}\) or 50\% probability of re-identification. Yet, a per-attribute analysis reveals that user 2’s privacy has been more significantly compromised. Due to the fact that user 2 and user 1’s records are similar, an attacker would know with 100\% certainty that user 2 \{wasn’t, was, wasn’t\} present in spacetime bins \{2,3,5\} respectively, and with a 50\%
chance that user 2 was present in spacetime bin 4. This is attribute disclosure, and we want to construct disclosure measures that account for it.

We can define as disclosure measure the earth mover distance of user $i$ on attribute $j$, denoted by $EM_{ij}$, as:

$$EM_{ij}(D|A) = \left| \frac{|E_{ij}|}{|E_i|} - \frac{|D_{ij}|}{n} \right|$$  \hspace{1cm} (4.4)

Where $|E_{ij}|$ is the number of records $i' \in E_i$ where $X_i^{j} = 1$, i.e., $|E_{ij}| = \{i' \mid i' \in E_i \land X_i^{j} = 1\}$; and analogously, $|D_{ij}|$ is the number of records $i' \in [1,n]$ where $X_i^{j} = 1$, i.e., $|D_{ij}| = \{i' \mid i' \in [1,n] \land X_i^{j} = 1\}$.

On the simple example of table 4.1, for user 2 and attribute 3:

$$EM_{23} = \left| \frac{|E_{23}|}{|E_2|} - \frac{|D_{23}|}{4} \right| = \left| \frac{2}{2} - \frac{3}{4} \right| = .25$$

The earth mover distance $EM_{ij}$ as measure of disclosure, which we refer to as $EM$-disclosure, captures the probabilistic information gain on user $i$’s $j$th attribute, from using knowledge $A_i$ on data $D$. The first term is the probability with which an attacker knows that $X_i^{j} = 1$ once using knowledge $A_i$ on data $D$. The second term is the probability with which an attacker knows that $X_i^{j} = 1$ by looking at the full data $D$. Their subtraction in absolute value measures the difference between prior and posterior probabilities $P(X_i^{j} = 1)$ and $P(X_i^{j} = 1 \mid A_i)$.

Note that, whereas unicity and $k$-disclosure are defined for the entire dataset $D$, $EM$-disclosure $EM_{ij}$ is defined per user and attribute, providing much more detailed information on disclosure risk of pseudonymized datasets. Thus we can compute meaningful disclosure statistics on subsets of $D$. For example, we define earth mover distance of user $i$, denoted $EM_i$, as the average disclosure along $i$’s attributes:

$$EM_i(D|A) = \frac{1}{d} \sum_{ij} \left| \frac{|E_{ij}|}{|E_i|} - \frac{|D_{ij}|}{n} \right|$$  \hspace{1cm} (4.5)
A per-user measure of disclosure provides useful insight for analyzing the inequality of the distribution of disclosure risk among individuals whose information is contained in $D$, for example. We can analogously define earth mover distance of attribute $j$, denoted $EM_j$, as the average disclosure for all users on attribute $j$:

$$EM_j(D|A) = \frac{1}{n} \sum_{i} \left( \frac{|E_{ij}|}{|E_i|} - \frac{|D_{ij}|}{n} \right)$$

A per-attribute measure of disclosure, for example, by understanding which attributes present higher and lower disclosure risk, can provide highly valuable insight on guiding the implementation of privacy protecting techniques (PPTs), such as generalization (see section refsec:PPTs). Finally, we can compute an earth mover disclosure statistic for the full dataset $D$, simply denoted as $EM$:

$$EM(D|A) = \frac{1}{nd} \sum_{i,j} \left( \frac{|E_{ij}|}{|E_i|} - \frac{|D_{ij}|}{n} \right)$$

An overall disclosure statistic can be useful for aggregate comparisons of privacy risk among a database under different implementation of privacy protecting techniques (PPTs), or for comparison among EM's measure of disclosure and other measures such as unicity and $k$-disclosure (see section 4.4.1).

**EM looks beyond identity disclosure (re-identification) and accounts for attribute disclosure.** Consider again the case of user 2 in figure 4.1. **EM-disclosure** acknowledges that the attacker has full certainty on 2’s $\{1, 2, 3, 5\}$ attributes, as $P(X_2^j = 1 | A_2) = \left\{ \frac{2}{3}, \frac{2}{3}, \frac{1}{2}, \frac{1}{2} \right\}$; and combines it with priors $P(X_2^j = 1) = \left\{ \frac{2}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{0}{4} \right\}$ to compute:

$$EM_2 = \frac{1}{5} \left\{ |\frac{2}{2} - \frac{2}{4}| + |\frac{2}{2} - \frac{1}{4}| + |\frac{2}{2} - \frac{3}{4}| + |\frac{2}{2} - \frac{3}{4}| + |\frac{2}{2} - \frac{0}{4}| \right\} = \frac{.5 + .25 + .25 + .25 + 0}{5} = .25$$

**Interpretation in the Context of Mobility Data**

Earth mover disclosure has a meaningful interpretation. Consider again the
context of mobility data, such as the one that can be extracted from CDRs (see section 1.1). In this context, entries $X_i^j = 1$ in $D$ denote that user $i$ was present in spatiotemporal bin $j$. Hence, $EM_{ij}$ can be read: given that we know John (user $i$) made a call within time and place $j' \in A_i$, we are $EM_{ij}$ more certain that he was/was not also present at spacetime $j$. Fundamentally, $EM$-disclosure captures correlations between records' attributes – people’s mobility patterns in this context – and measures the probabilistic knowledge that can be gained by leveraging these correlations to connect adversary knowledge to unknown sensitive information.

This section showed earth mover disclosure as a natural generalization and development from unicity, designed to alleviate several of its limitations. The following section shows earth mover disclosure as a simple case of a robust information-theoretic framework for disclosure analysis. Section 4.4.1 analyzes differences in using unicity versus earth mover disclosure in the context of high dimensional data.

4.2 An Information-Theoretic Framework for Disclosure Analysis of HD Data

We can conceptually define disclosure that occurs due to certain action as the difference in adversarial (or simply undesired) knowledge prior and posterior to the action. This conceptualization of disclosure was first proposed by Li 2007 in the context of low-dimensional data [11].

Within the field of data privacy, this thesis focuses on the risk and value of publishing nominal \footnote{nominal data refers to data disaggregated at the individual level, e.g., census data specifying sociodemographic variables of each individual.} pseudonimized data, as opposed to simple aggregates or statistics. As [11] and [7], the action we want to measure disclosure for is the publishing of nominal pseudonimized data.

Thus, we can define prior knowledge as that which follows from general statistics
Figure 4-1: Illustration of prior distribution $P(X_i^j)$, posterior distribution $P(X_i^j|A_i)$, and difference between the two as quantitative measure of disclosure.

of the population. In sparse data such as the case of mobility patterns, where prior probability $P(X_i^j)$ is the probability that any person in particular was present in space and time $j$, prior $P(X_i^j) \rightarrow 0$. On the other hand, posterior probability $P(X_i^j|A_i)$ is the the probability of $i$ having been present in space and time $j$, that can be computed given pseudonimized data $D$ and complimentary adversarial knowledge $A_i$.

We can quantify this difference as the distance between prior and posterior distributions $P(X_i^j)$ and $P(X_i^j|A_i)$, in an information-theoretic sense. Two methodologies for quantifying this distance are the earth mover distance (EMD) and the Kullback-Leibler divergence (KLD).

Earth Mover Distance

The earth mover distance (EMD) is defined as the minimum work needed to convert one distribution into another. The visual analogy pictures one distribution as a mounts of earth dispersed on the floor, and the other as holes in the ground; the earth mover distance between the two is pictured as the minimum amount of earth that has to be transported to make the ground even.

More formally work is defined as the product of distance times mass. Let $R = (r_1, r_2, ..., r_m)$ and $S = (s_1, s_2, ..., s_m)$ be probability distributions. And let $F^{m \times m}$ be a flow matrix, where $f_{ij}$ denotes the amount of mass that flows from $r_i \in R$ to $s_j \in S$. 

48
Then work is defined as:

$$W = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

The aforementioned visual analogy assumes a common euclidean norm. Yet, an interesting characteristic of EMD is the ability to specify measures of distances $d_{ij}$ among values in the support of the distribution. By defining such values, EMD can incorporate desired semantics into the distance calculation. Designing measures of distances EMD can account for semantics in the data. For instance, if variables $j$ in data matrix $D$ would codify monthly income levels of individuals, then we would like to take into account that income segment [$1,000, \$1,500] is more related to segment [$800, \$1,000] than segment [$3,000, \$4,000]. More generally, any data semantics higher than binary or categorical, require measures and corresponding distances that account for it. In the case of spatiotemporal data, it would be desired to use a measure that accounts for distances across space and time, where moving probability mass from to contiguous polygons should cost less work than moving mass long distances; similarly across time.

Once distance is defined, then $EMD(R, S)$ is defined as the minimum work to transform $R$ into $S$:

$$\min_{F} W(R, S) = \min \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

such that $\forall i \in [1, m]$:

$$r_i = -\sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ij} = s_i$$

Finding $EMD(R, S)$ thus imply solving the optimization problem above.

Section 4.1.2 developed the earth mover distance measure of disclosure as a generalization and extension of unicity, which addressed at least two of its core limitations: limitation to deterministic disclosure and limitation to identity disclosure. Here we show it as a simple case of a robust framework: disclosure quantified as EM information distance between adversary’s prior and posterior distributions.
Consider the case where distributions $R$ and $S$ have binary support $\sigma_R = \sigma_S = \{0, 1\}$. Now let’s define distance $d_{ij}$ simply as $d_{ij} = 1 \ \forall i, j \in 0, 1$. Then work minimization problem 4.8 translates into

$$\min_F W(R, S) = \min_F \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij} = \min_F \sum_{i=0}^{1} \sum_{j=0}^{1} f_{ij} = \min_F (f_{00} + f_{01} + f_{10} + f_{11}) \quad (4.10)$$

$$r_i = \sum_{j=0}^{1} f_{ij} \quad s_i = \sum_{j=0}^{1} f_{ij} \quad (4.11)$$

Minimization problem 4.10 has a trivial solution, namely: if $r_1 < s_1$ then $\{f_{00}, f_{01}, f_{10}, f_{11}\} = \{0, s_1 - r_1, 0, 0\}$, and if $r_1 > s_1$ then $\{f_{00}, f_{01}, f_{10}, f_{11}\} = \{0, 0, r_1 - s_1, 0\}$. In any case, we have that

$$W^*(R, S) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}^* = |r_1 - s_1| \quad (4.12)$$

Now, in the context of disclosure analysis, let $R$ be prior probability $P(X_i^j)$ and $S$ be posterior probability $P(X_i^j|A_i)$ (see section 4.1.2). Note that all distributions have binary supports $\sigma = \{0, 1\}$. Then we have that

$$\text{EMD}(P(X_i^j), P(X_i^j|A_i)) = \text{EMD}(R, S) = W^*(R, S) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}^*$$

$$= |r_1 - s_1| = \left| P[X_i^j = 1] - P[X_i^j = 1|A_i] \right| = \left| \frac{D_{ij}}{n} - \frac{E_{ij}}{E_i} \right| \quad (4.13)$$

Expression 4.13 above is equal to the proposed definition of earth mover measure of disclosure, $\text{EM-disclosure}$ described in section 4.1.2. Thus, the methodology proposed in section 4.1.2 is only a particular case where $D$ is binary and distances in support $\sigma$ are simply defined as $d_{ij} = 1$. It is a particular case of the more robust framework for assessing disclosure, where it is quantified as EM information distance between adversary’s prior and posterior distributions. Interesting research avenues
are are extensions of this methodology to assessing disclosure on data with higher semantics: for example, by defining distances over hierarchical categories, space and time, or social networks.

**Kullback Leibler Divergence**

Other well known notions of distance between two probability distributions exist. A relevant one is the Kullback Leibler Divergence (KLD). Denoted by $KLD$, the Kullback Leibler Divergence between distributions $R$ and $S$, with discrete support $\sigma$ is defined as:

$$KL(R, S) = \sum_{t \in \sigma} r_t \log \frac{r_t}{s_t}$$  \hspace{1cm} (4.14)

In the context of disclosure analysis, we call $KL$-disclosure the KLD used as measure of disclosure, defined as:

$$KL_{ij}(P(X^j_i|A_i), P(X^j_i)) = \sum_{t \in \sigma} P(X^j_i = t|A_i) \log \frac{P(X^j_i = t|A_i)}{P(X^j_i = t)}$$  \hspace{1cm} (4.15)

Note that, similar to $EM$-disclosure ($EM_{ij}$) and different than unicity and $k$-disclosure, $KL$-disclosure ($KL_{ij}$) is defined per user and attribute, providing much more detailed information on disclosure risk of pseudonimized datasets. Similar to derivations in section 4.1.2 we can compute meaningful disclosure statistics on subsets of $D$, such as per-attribute, per-individual and overall disclosure measures (equations 4.6, 4.6 and 4.7 respectively).

An important characteristic of the use of $KL$-disclosure is the possibility of deriving bounds [7]. In the context of disclosure analysis, this implies the possibility of deriving bounds on the information that an adversary can gain from data $D$ using adversary knowledge $A$. A practical disadvantage of $KL$-disclosure is that its interpretation is less intuitive as disclosure measure; in particular to non-academic audiences, such as policy decision makers in public and private sectors (see section 4.1.2 for $EM$-disclosure’s interpretation, in contrast).
4.3 Adversary Knowledge Assumptions

Prior work on the information distance paradigm for disclosure, which section 4.2 builds on, rose in the context of low-dimensionality. Similar to \textit{k-anonymity}, focus was placed on being able to derive general privacy guarantees. Concepts of \textit{t-closeness} \cite{11} and \textit{\delta-disclosure} \cite{7} characterize a database \( D \) with thresholds \( t \) or \( \delta \) over which no equivalent class in \( D \) discloses information, assuming full adversary knowledge \( A_i \) of quasi-identifiers. As discussed in sections 3.2 and 3.1.1, this paradigm is inapplicable in the modern context of high-dimensional data, sparsity, and partial adversary knowledge. Rather, methodologies proposed in sections 4.1.2 and 4.2 of this thesis seek to borrow information-theoretic concepts from \textit{t-closeness} and \textit{\delta-disclosure} to build disclosure measures that can be coupled with the partial adversarial knowledge model introduced by unicity \cite{1} \cite{2}, as well as its possible extensions.

High-dimensional (HD) databases are most often sparse, and due to sparsity, sensitive attributes themselves can function as quasi-identifiers (see section 3.1.1). Models where potential adversaries know the full set of quasi-identifiers of an individual are in HD contexts inapplicable or a rather trivial scenario, for two main reasons. First, such assumption implies adversary knowledge (AK) of the full set of sensitive attributes, which is unlikely, and in which case there is no further information to be gained by a privacy attack. Second, as shown in section 3.1.2 for \textit{k-anonymity}, attempting to meet disclosure thresholds such as \textit{unicity}, \textit{k-disclosure}, \textit{t-closeness}, etc., under full AK knowledge assumption, requires generalization of sensitive attributes to extents that erode most data usefulness.

Therefore we must consider scenarios where only partial knowledge is available to a potential attacker. For instance, cases where a potential attacker knows the home and work location of an individual \( i \), but not \( i \)'s full mobility behavior. Work on \textit{unicity} pioneered modeling partial AK. Taking a closer look to this methodology – and recalling section 3.2.2– we can conceptualize \textit{unicity}'s modeling of re-identification as
estimating:

\[ P(ID) = P(ID|A)P(A) \]  \hspace{1cm}(4.16)

where \( ID \) denotes identity disclosure and \( A \) stands for adversary knowledge.

Similarly, we can define the above conceptualization for the more comprehensive concept of attribute disclosure (see section 2.1)

\[ P(disclosure) = P(disclosure|A)P(A) \]  \hspace{1cm}(4.17)

where \( disclosure \) denotes attribute disclosure, which can be quantified by measures presented in sections 4.1.2 and 4.2.

Methodology in [1] and [2] chooses at random \( p \) data points for sample \( S \) of individual records \( X_i \) in dataset \( D \). The \( p \) data points constitute adversary knowledge \( A_i \) for each individual in \( S \). For example, \( p \) data points can be 3 calls or SMS issued by mobile phone user \( i \) from some spatiotemporal bins \( \{j_1, j_2, j_3\} \). This methodology is in fact a **Monte Carlo stochastic simulation**\(^3\) [17] of probability of re-identification, or probability of disclosure. In Monte Carlo terminology:

1. The input variable is \( A \), for which a domain and probability distribution \( P(A) \) is assumed;

2. Deterministic computation of the quantity of interest is conducted in \( ID|A \);

3. Output distribution \( P(ID) \), or quantities of it — i.e., \( unicity = \frac{u_i}{n} \), see section 4.1.1 — are estimated.

Under this scope, the question is: what domain and distribution \( P(A) \) is unicity implementing? Is it a sensible assumption? how can we extend, adapt, or improve it?

---

\(^3\)Monte Carlo simulation is a class of computational algorithms that rely on repeated random sampling to derive numerical estimations of quantities related to complex probability distributions [17]
Domain of $P(A)$ in unicity’s methodology is determined by parameter $p = |A_i|$, $\forall i$, the number of randomly chosen data points $X^i_j$ to be included in $A_i$ for each record $X_i$ in sample $S$. For example, if $p = 2$, then domain of $P(A_i)$ are all different combinations of 2 attributes $(j_1, j_2) \in [1, d]^2$; if $p = 4$, then domain of $P(A_i)$ are all different combinations of 4 attributes $(j_1, j_2, j_3, j_4) \in [1, d]^4$. A probability distribution $P(A)$ is then defined over the domain, i.e., the likelihood of attributes $j$ to be included as one of $p$ elements of $A_i$. It is relevant to note that the choice of parameter $p$ is arbitrary. Work in [1], for instance, tests and reports results for different values of $p$ ($p = 3$ and $p = 4$). Therefore, quantities estimated with $P(A)$ as input – e.g., unicity – are not absolute measures of disclosure risk, but ordinal measures that allow for decisions among alternatives on, for example, different data treatments by use of PPTs, or different public or corporate data policies. This limitation to ordinal or relative disclosure risk assessment is likely to be understood in academia, but generates a degree of misunderstanding in policy domains.

Once domain of $P(A_i)$ is defined by specifying parameter $p$, unicity’s methodology in [1] implements a probability distribution $P(A_i)$. In it, it is relevant to note that different attributes $j \in [1, d]$ are not included in $A_i$ at random with equal probability, rather, each attribute $j$ is picked at random with a probability proportional to the amount of communications (i.e., calls, SMS, etc.) that user $i$ had on space and time $j$. For example, an employed adult will typically engage in more communications during working hours from his work location, and at nighttime from his home location; correspondingly, $P(A_i)$ models a higher probability of adversary knowledge on such times and locations for individual $i$. Conversely, for a call made at some space and time $j$ out of $i$’s typical mobility patterns, $P(A_i)$ implemented by unicity models adversary knowledge of $i$’s presence in such $j$ as less likely than that of $i$’s presence in home or work locations.

$P(A_i)$ implemented in [1] and [2] is thus a reasonable assumption, absent better domain-specific understanding of the likelihood of adversarial knowledge. How-
Spatial Granularity  

<table>
<thead>
<tr>
<th>Spatial Granularity</th>
<th>Temporal Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>District</td>
<td>1 h</td>
</tr>
<tr>
<td>(D, 1h)</td>
<td>(D, 6h)</td>
</tr>
<tr>
<td>Municipality</td>
<td>1 h</td>
</tr>
<tr>
<td>(M, 1h)</td>
<td>(M, 6h)</td>
</tr>
</tbody>
</table>

Table 4.2: Grid of spatiotemporal granularity levels over which EM-disclosure, KL-disclosure and unicity were computed.

ever, one key area of opportunity for advancing disclosure analysis of pseudonimized datasets is precisely the generation of better, domain-specific models of likelihood of AK obtained from sources external to the published dataset. Finally, it is relevant to note that if an improved model for $P(A)$ is available, it can seamlessly be plugged into the Monte Carlo simulation framework here described, synthesized by equation 4.17.

### 4.4 Empirical Results on Large-Scale Mobile Phone Metadata

In this section we compute EM-disclosure, KL-disclosure and unicity disclosure measures for the mobile phone records (CDRs) dataset described in section 3.1.1. We compute these measures over a grid of various levels of spatial and temporal generalization applied to the database, and compare results among methodologies. Generalization was applied to coarsen the CDR database according to the grid of spatiotemporal granularities shown in table 4.2.

**KL-disclosure**

As mentioned in section 4.2, we can define $KL$ as an overall measure of disclosure for database $D$. Analogous to construction of $EM$ from $EM_{ij}$, we define:

$$KL(D|A) = \frac{1}{nd} \sum_{(i,j)} KL_{ij}$$  \hspace{1cm} (4.18)
\[
KL(D|A) = \frac{1}{nd} \sum_{(i,j)} \sum_{t \in \sigma} P(X^j_i = t|A_i) \log \frac{P(X^j_i = t|A_i)}{P(X^j_i = t)}
\]

In the case of binary variables \(X^j_i\), and continuing with notation introduced in section 4.1, we have that:

\[
KL(D|A) = \frac{1}{nd} \sum_{(i,j)} \sum_{t=0}^{1} \left( \frac{|E_{ij}| \log |D_{ij}|}{|E_i|} \right)
\]

(4.19)

4.4.1 Ordinal Equivalence of Unicity and Information Distance Measures of Disclosure

Unicity’s range is \(unicity(D|A) \in [0, 1]\), as it is the fraction of records \(i\) that are unique in their equivalent class \(|E_i|\). Moreover, we note from equation 4.4 in section 4.1.2 that \(EM_{ij}\) is the difference of two probabilities, thus we have that \(EM_{ij} \in [0, 1]\) and \(EM = \frac{1}{nd} \sum_{(i,j)} EM_{ij} \in [0, 1]\). However, it’s relevant to note that \(KL\)-disclosure is not comparable in absolute terms to \(EM\)-disclosure and unicity, as it can take values out of range \([0, 1]\).

Figure 4-2 shows results from computing unicity using 3, 4, and 5 random points of adversary knowledge (see sections 3.2.2 and 4.3), as well as information distance-based measures \(EM\)-disclosure and \(KL\)-disclosure. On the x-axis of figure 4-2 are the spatiotemporal generalization levels applied to dataset \(D\), specified in table 4.2, which are laid out along the x-axis in descending order in respect to values of unicity \(\oplus p = 4\). We call a combination of spatial and temporal granularities a generalization profile. In this analysis, eight generalization profiles were applied to \(D\) and assessed by five disclosure measures.

Relevantly, all five measures of disclosure preserved ordinality of the generalization profiles assessed. Although profiles were laid out on the x-axis in descending order with respect to unicity \(\oplus p = 4\), all five series in figure 4-2 are monotonically decreasing. Let \(g_i\) with \(i \in [1, 8]\) denote the eight generalization profiles evaluated;
Figure 4-2: Comparison of unicity and information distance-based measures of disclosure over a grid of spatiotemporal generalization levels. EM-disclosure stands for earth mover disclosure, and KL-disclosure stands for Kullback Leibler disclosure. On the x-axis are the combinations of spatiotemporal generalization, where \( D \) denotes per-district and \( M \) per-municipality spatial granularities (see table 4.2). Spatiotemporal generalization combinations are arranged on the x-axis in descending order with respect to unicity @\( p = 4 \) values (which coincides with the order of values on each of the other 4 series).
and let $m_j(g_i)$ with $j \in [1, 5]$ denote disclosure assessment of profile $i$ under measure $j$ (e.g., $m_2(g_3)$ can denote unicity $@p = 4$’s assessment of profile $(D, 6h)$). Then, results here presented hold the **ordinal equivalence property**:

$$m_j(g_i) \geq m_j(g_{i'}) \iff m_{j'}(g_i) \geq m_{j'}(g_{i'}) , \forall i \neq i', \forall j \neq j'$$  \hspace{1cm} (4.20)

This property means that any generalization profile that under one methodology is more private than another profile, will also be more private under any other ordinally equivalent methodology. Ordinal equivalence has been here shown empirically for dataset $D$, yet a germane line of research is to elicit the mathematical properties of $D$ and $A$ (e.g., high-dimensionality, sparsity, uniqueness) under which ordinal equivalence holds.

Ordinal equivalence is of high relevance, as policy decision-making based on any measure from a set of ordinally equivalent measures will yield the same policy choices. In particular, ordinal equivalence preserves Pareto dominance and Pareto optimality properties of choices being assessed. For instance, chapter 5 conducts Pareto analysis of a grid of generalization profiles applied to a large CDR dataset. Each profile is evaluated in terms of its data disclosure risk and its data usefulness, and Pareto analysis is conducted over these two objectives. In this case, unicity $@p = 4$ is used as disclosure measure on the privacy objective. Results of this section 4.4.1 on dataset $D$ indicate that, due to **ordinal equivalence**, Pareto dominance and optimality results of chapter 5 would be unchanged if other measures of unicity (e.g., $@p = 3$), or more robust information distance measures of $EM$-disclosure or $KL$-disclosure, were used instead. The following **lemma** formalizes this notion:

**Lemma 4.1**: Let $x$ be a profile evaluated on objectives $O_1 \in \mathbb{R}$ and $O_2 \in \mathbb{R}$. Let also $T : \mathbb{R} \rightarrow \mathbb{R}$ be a transformation of objective $O_1$ that meets the **ordinal equivalence property**, such that $O_1(x) > O_1(y) \iff T(O_1(x)) > T(O_1(y))$. Then, $x$ is Pareto optimal in $\{O_1, O_2\} \iff x$ is Pareto optimal in $\{T(O_1), O_2\}$
Proof: If \( x \) is Pareto optimal in \( \{O_1,O_2\} \), assuming a minimization problem without loss of generality, then \( \nexists y \) such that \( O_1(y) < O_1(x) \) and \( O_2(y) < O_2(x) \). Now, due to ordinal equivalence property of \( T \), we know that \( \nexists y \) such that \( T(O_1(y)) < T(O_1(x)) \) and \( O_2(y) < O_2(x) \). Therefore \( x \) is Pareto optimal in \( \{O_1,O_2\} \). The reverse path is true as long as \( T \) can be inverted.

Lemma 4.1 formalizes the notion that ordinal equivalence property preserves Pareto optimality properties across diverse measures of the same objective. Extensions to multiple objectives and transformations follow from the above proof.

It is relevant to note that, although ordinal equivalence was shown among methodologies for these dataset and generalization profiles, information distance disclosure measures are likely to be preferred in more sophisticated cases where: other PPTs than generalization were to be implemented, such as sampling and noise addition; consideration of higher data semantics was desired, such as continuity over space and time, ordinality of income segments, or distance over a social network; analysis of information disclosure bounds was desired; finer disclosure analysis per-sensitive attribute or per-user was desired; or where attribute disclosure is likely to occur even when identity disclosure does not (see section 4.1.1). Finally, germane investigation should elicit mathematical properties of \( D \) and \( A \) under which ordinal equivalence holds.
Chapter 5

Balancing Data Utility and Privacy: Informing Policy Formulation

This chapter moves on to investigate the tradeoff between data privacy and data usefulness associated to real-world applications that involve mobile phone metadata (referred to as CDRs). Section 5.1 describes the study design, which was conducted in collaboration between United Nations’ Global Pulse Initiative\(^1\) and the Human Dynamics Laboratory (at the MIT Media Lab). Section 5.2 presents main results.

Analysis in this chapter considers disclosure risk and data usefulness in conjunction. We focus on the case of mobility patterns that can be extracted from mobile phone records (see section 1.1), which are a particularly relevant data source for development purposes. CDRs are standard and generated by telecommunication companies on an ongoing basis, see section 1.1 for description of CDRs. Moreover, handsets and airtime are becoming cheaper, leading to increased penetration and representativity, which by 2013 approached 89% in developing countries and 96% globally [16].

There are several ways in which location information in mobile phone records can be analyzed and used. Two of such uses are: population density and population

\(^1\)Global Pulse is a flagship innovation initiative of the United Nations Secretary-General on big data. For more information please visit www.unglobalpulse.org
mobility mapping. This information has valuable applications in domains relevant and
diverse as disaster response on earthquakes [5] and floodings [23], epidemics analysis
in both the developed [9] and developing worlds [14], and transportation systems
development [6].

Given the demonstrated potential of mobile phone data, it is incumbent upon the
data-driven development community to examine the utility-disclosure risk landscape
under various privacy preserving treatments in relevant domains of application. The
following sections presents first steps along these lines.

5.1 Study Design

We used 1 month of pseudonymized mobile phone metadata from 1.6 million users on
a metropolis in the developing world (dataset described in section 3.1.1). On the one
hand, we developed maps of regional density and mobility patterns of the popula-
tion, at various spatial and temporal granularities. We then assessed the information
usefulness through in-depth interviews and a standardized questionnaire applied to
a pool of 10 field experts in two domains: epidemics prevention and control, and
transportation development. On the other hand, we incorporated disclosure results
computed in section 4.4 over the same grid of spatiotemporal granularities, to elicit
the utility-privacy tradeoff.

Dataset $D$ was generalized according to a grid of spatial and temporal data gran-
ularities. We call spatiotemporal generalization levels generalization profiles. Ta-
ble 5.1 specifies the grid of generalization profiles used in this study, where levels of
spatial granularity are \{ZIP code, District, Municipality\}, illustrated in figure 5-1;
and levels of temporal granularity are \{1 hour, 6 hours, 12 hours, 1 week\}.

For each generalization profile disclosure and utility assessment were conducted.
Disclosure assessment was based on unicity $@p = 4$ as measure of disclosure (see
Figure 5-1: Spatial generalization levels of the metropolitan area: (a) Municipality, (b) Districts, (c) ZIP codes.
Temporal Granularity

<table>
<thead>
<tr>
<th>Spatial Granularity</th>
<th>1 h</th>
<th>6 h</th>
<th>12 h</th>
<th>1 week</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZIP code</td>
<td>(ZIP, 1h)</td>
<td>(ZIP, 6h)</td>
<td>(ZIP, 12h)</td>
<td>(ZIP, 1w)</td>
</tr>
<tr>
<td>District</td>
<td>(D, 1h)</td>
<td>(D, 6h)</td>
<td>(D, 12h)</td>
<td>(D, 1w)</td>
</tr>
<tr>
<td>Municipality</td>
<td>(M, 1h)</td>
<td>(M, 6h)</td>
<td>(M, 12h)</td>
<td>(M, 1w)</td>
</tr>
</tbody>
</table>

Table 5.1: Grid of spatiotemporal granularity levels over which $D$ was generalized, and data utility and disclosure risk assessed.

sections 3.2.2 and 4.3).

Assessment of data usefulness was conducted by a series of in-depth interviews and a standardized questionnaire applied to a pool of 10 field experts in two application domains: infectious disease prevention and control and urban transportation systems development. Through this instruments we gathered field experts’ perspectives on how mobile phone data is more or less useful at different levels of generalization.

Map visualizations of the generalized datasets, such as those in figure 5-1, were produced and used for both interviews and questionnaires. Each expert was asked to answer how valuable the generalized datasets would be if available to their professional activity and area of expertise. Usefulness assessments were given in a quantitative scale ranging from 0 (no utility) to 10 (high utility). Quantitative questions were complemented by qualitative questions regarding the potential use of CDR data. It is worth mentioning the need to provide to experts an explanation of the nature of CDR data prior to their assessment.

5.2 Results

Table 5.2 shows unicity ($\@p = 4$) scores for each generalization profile. Aligned with previous studies on privacy of pseudonymized CDR datasets [1], we show that four spatio-temporal points of adversary knowledge are enough to uniquely identify $+95\%$ of individuals in $D$, at a [ZIP code, 1 hour] spatiotemporal granularity. We also see that the ratio (unicity) decreases as data is coarsened, and reaches values below $0.2\%$. 64
Combining unicity results with usefulness assessments from the questionnaire, we can analyze the tradeoff between usefulness and disclosure risk of $D$ as data is generalized. Figure 5-2 shows results of this comparison. The x-axis denotes data utility and the y-axis denotes disclosure risk. Each point in figure 5-2 represents the privacy and utility scores of a generalization profile.

Figure 5-2 shows the Pareto front, i.e., the set of generalization profiles which are non-dominated in Pareto sense by other profiles\(^2\). The Pareto front is central to multiobjective decision analysis as it’s formed by the set of options that constitute the underlying tradeoff among objectives. Dominated options will not be chosen and do not imply a tradeoff, as, by definition, exists an alternative option superior in all objectives.

Utility results of this study are not meant to be representative, but a demonstration of core concepts and methodologies for decision analysis on data utility-privacy tradeoffs. Results are based on a sample of 10 experts, thus standard errors on utility are not meaningful. Towards statistical analysis of utility-privacy tradeoffs in relevant domains, future research should extend this methodology to larger samples on well-defined populations.

From figure 5-2 we first confirm existence of a tradeoff, for the interviewed experts: a negative relationship between utility and privacy of the pseudonimized dataset, represented by the Pareto front’s negative slope. However, note that the tradeoff is not strict, as not all profiles lie on the Pareto front. This implies that not every gain

\(2\)A profile $x$ is dominated by profile $y$ if $y$ is equal or better than $x$ in all objectives.
Figure 5-2: Results of utility and disclosure risk assessments per generalization profile. The x-axis denotes data utility and the y-axis denotes disclosure risk. Each point represents the privacy and utility scores of a generalization profile. Note that utility axis x is reversed for presentation purposes, so that minimization is preferred on both objectives; thus a preferred generalization profile is one that is closer to the southwest corner of the plots. Error intervals for unicity estimates are of order $O(.0001)$. (a) Transportation systems development application domain; (b) Infectious disease prevention and control application domain.
in privacy entails loss in utility, and conversely, not every gain in utility entails privacy loss. Most relevantly, it implies that policy decision-making on data treatments (e.g., generalization profiles) should be guided by an understanding of the underlying privacy-utility tradeoff, as suboptimal policies can otherwise be implemented, incurring in potentially significant inefficiencies and unnecessary privacy or utility losses (e.g., profile (M,1) in figure 5-2).
Chapter 6

Concluding Remarks

6.1 Disclosure Analysis

This thesis conceptualizes, relates, and generalizes salient methodologies for disclosure analysis of published pseudonymized data that have been developed in the last two decades, such as: \textit{k-anonymity}, \textit{t-closeness}, and \textit{unicity}. Data at the core of the so-called "big data" revolution is fundamentally high-dimensional. The transition from low to high data dimensionality marks a paradigmatic divide among disclosure methodologies, where the latter context presents fundamental new challenges (chapter 3), namely: sparsity, characteristic of high-dimensionality; sensitive attributes functioning as quasi-identifiers; and inapplicability of earlier paradigms of full adversary knowledge assumptions and pursuit of general privacy guarantees.

In particular, prior work on the information distance paradigm for disclosure, which section 4.2 builds on, rose in the context of low-dimensionality. Similar to \textit{k-anonymity}, focus was placed on deriving general privacy guarantees. Concepts of \textit{t-closeness} \cite{DBLP:journals/corr/abs-1507-08657} and \textit{δ-disclosure} \cite{DBLP:journals/corr/abs-1507-08657} characterize a database with thresholds $t$ or $δ$ over which no equivalent class in $D$ discloses information, assuming full adversary knowledge of quasi-identifiers. This paradigm is inapplicable in the modern context of high-dimensional data, sparsity, and partial adversary knowledge. Rather, methodologies
proposed in sections 4.1.2 and 4.2 of this thesis sought to borrow information-theoretic concepts from \textit{\textit{t-closeness}} and \textit{\textit{\delta-disclosure}} to build disclosure measures that can be coupled with the partial adversarial knowledge model introduced by \textit{unicity} \cite{1} \cite{2}, as well as its possible extensions (section 4.3).

Results on a large CDR dataset showed \textit{ordinal equivalence} among \textit{unicity} measures and information distance disclosure measures \textit{EM-disclosure} and \textit{KL-disclosure}. Investigation is needed to elicit the mathematical properties of \textit{D} and adversary knowledge \textit{A} under which ordinal equivalence holds. However, information distance disclosure measures are likely to be preferred for more sophisticated uses: where other PPTs than generalization were to be implemented, such as sampling or noise addition; where consideration of higher data semantics was desired, such as continuity over income segments or spatiotemporal variables, ordinality of preferences, or distance over a social network; where analysis of information disclosure bounds was desired; where finer disclosure analysis per-sensitive attribute or per-user was desired (section 4.1.2); or where attribute disclosure is likely to occur even when identity disclosure does not (section 4.1.1).

\textbf{Extensions and Future Work}

Work here presented on disclosure analysis points to several complementary research paths, that could:

- Elicit conditions on \textit{D} and adversary knowledge \textit{A} over which ordinality property among disclosure disclosure measures holds.

- Investigate improved domain-specific adversary knowledge models, \textit{P(A)}, that can be plugged into the Monte Carlo framework described in section 4.3.

- Investigate extensions of information distance measures to higher semantics, such as continuity over income segments or spatiotemporal variables, ordinality of preferences, or distance along social networks.
Investigate the use of disclosure bounds (e.g., Cramer Rao bounds), and their possible role on machine learning-based data engineering tasks towards optimal transformations of pseudonimized datasets.

The Big Picture

We are still distant from accurate probabilistic estimates of disclosure risk, in absolute terms. Such assessments would require not only to specify a complete model for the probability of disclosure conditioned to the availability of adversary knowledge \(P(disclosure|A)\), and a complete model of the probability of adversary knowledge \(P(A)\); but also a complete model of the value (or Loss) that any possible disclosure event implies. Recalling equation 4.17 from section 4.3, a complete disclosure risk model includes:

\[
disclosure\ risk = \text{Loss}(disclosure)P(disclosure|A)P(A)
\]  

(6.1)

Literature on disclosure risk is not close from coding accurate social, corporate, or even individual loss functions for disclosure. Moreover, a relevant and challenging research avenue forward is precisely the improvement of adversary knowledge models \(P(A)\) in eq. 6.1). In words of Frank R. Field III (MIT, 2014): "We are constantly trapped in the duality between what one can measure and what one wants to know... 'I cannot predict the future; nevertheless, I will try to predict the future'".

However, modern disclosure measures provide assessments valuable for disclosure analysis, management, and policy decision-making (e.g., analysis in chapter 5). In this sense, disclosure assessments and protection achieved by use of these measures can be simplistically compared to car security systems: where the design goal is not to develop guaranteed systems for car theft prevention under all possible circumstances, nor to provide absolute probabilistic estimates on the likelihood of each variant of theft, for each vehicle; but rather, systems that significantly hinder theft in scenarios where attackers devote reasonable amounts of resources, as well as insightful estimates
and analytic frameworks that enable design and decision-making over such systems.

6.2 The Utility and Privacy Tradeoff

Chapter 5 investigates the tradeoff between data privacy and data usefulness related to mobile phone metadata (CDRs) and its real-world applications. The study assessed information usefulness through in-depth interviews and a standardized questionnaire applied to a pool of 10 field experts in two domains: epidemics prevention and control and transportation development. Disclosure measurements for the same grid of spatiotemporal granularities were incorporated to the analysis, towards eliciting the utility-privacy tradeoff.

On the disclosure side, four spatio-temporal points were enough to identify uniquely +95% of individuals, at a [ZIP code, 1 hour] spatiotemporal granularity – consistent with main results in the literature [1][2]. As the dataset was coarsened in space and time, the ratio (unicity) decreased to values below 0.2% for data specified at [District, 1 week] granularity or lower. We confirmed the existence of a utility-privacy tradeoff for experts in this study, i.e., a positive relationship between reidentification risk and data utility. However, Pareto analysis reveals that several generalization profiles are Pareto-suboptimal, thus the tradeoff is not strict. Non-strictness implies that not all privacy gains entail utility loss, and conversely, not all utility gains entail privacy loss. Results thus suggest that data policy decisions should rest on an understanding of the underlying privacy-utility tradeoff, as suboptimal policies can otherwise be implemented, unnecessarily incurring in potentially significant privacy or utility losses. Lastly, analysis showed that, due to ordinal equivalence and lemma 4.1 in section 4.4, chapter 5’s results on the utility-privacy tradeoff are invariant to assessing disclosure through information distance measures proposed in chapter 4.2.

Overall, work in chapter 5 provides a decision-theoretic framework for policy analysis and decision-making on the data utility-privacy tradeoff, and demonstrates gen-
eral properties of such tradeoff in two relevant application domains. However, results here presented do not aim at representativity. Extended application of this methodology over a domain-representative sample is needed prior to its use for public or corporate policy formulation.
Bibliography


