Measuring and predicting anonymity

Koot, M.R.
With the emergence of computers and the internet, the collection, storage and processing of information about private lives is becoming ubiquitous. Large amounts of data about citizens are stored in various data sets, spread over databases managed by different organizations all around the world [3, 27, 70]. Data about individuals drives policy research on all sorts of topics: finance, health, and public administration, to name a few. Increasingly, data about individuals is also collected for purposes other than policy research: targeting advertising, personalized medicine, individual risk-profile based insurance, welfare fraud detection, and so on.

Suppose one is asked to anonymously fill out a questionnaire containing questions about privacy-sensitive subjects such as health and politics. At the end, one is asked to reveal age, gender and (partial) postal code. What is the privacy risk associated with revealing that additional information? Can one be sufficiently sure that revealing that information does not allow the pollster, or anyone else with access to the questionnaire form or the database which one’s answers probably end up in, to identify one afterwards by matching that information to public profiles on social media, or by asking a friend at the registry office or tax authority to match it to the database of named citizens? After all, that might enable the pollster to ‘hold answers against’ the respondent and to include in her analysis information about the respondent that the respondent were not asked for during the questionnaire, or decided not to disclose.
CHAPTER 1. INTRODUCTION

Motivated by the desire to establish a better understanding of privacy, and thereby take away some of the fear, uncertainty and doubt surrounding privacy problems, the objective of this thesis is to study techniques for measuring and predicting privacy. Ideally, we want to develop mathematical tools useful for privacy risk assessment at both the personal level and the population level.

Unfortunately, the word privacy suffers from semantic overload. Privacy can be approached from various perspectives, such as ethics, law, sociology, economics and technology (the latter being our perspective). Before focusing on how to measure, we first want to know what to measure and why. To that end, this introductory Chapter has a broad scope and first considers multidisciplinary aspects of privacy. A property shared between various perspectives is that privacy entails some desire to hide one’s characteristics, choices, behavior and communication from scrutiny by others. Such ‘retreat from wider society’ may be temporary, such as when visiting the bathroom, or more permanent, such as when opting for hermit life or choosing to publish using a pseudonym. Another prevalent property is that privacy entails some desire to exercise control over the use of such information, for example to prevent misuse or secondary use. Phrases commonly associated with privacy include “the right to be let alone”, meaning freedom of interference by others [85]; “the selective control of access to the self or to one’s group”, meaning the ability to seek or avoid interaction in accordance with the privacy level desired at a particular time [2]; and “informational self-determination”, meaning the ability to exercise control over disclosure of information about oneself. The latter phrase was first used in a ruling by the German Constitutional Court related to the 1983 German census.

It is unlikely that any reasonable person would accept that all their thoughts, feelings, social relations, travels, communication, physical appearance including the naked body, sexual preferences, life choices and other behavior are knowable by anyone, at any time, without restriction — not least because that exposes them beyond their control to yet unknown people and institutions in yet unknown situations, i.e., pose a risk to their personal security and/or feeling of security.

At the same time, transparency of the individual can reduce risk, including collective risk. In the Netherlands, for example, the welfare-issuing Dutch municipalities have commissioned an organization named Stichting Inlichtingenbureau\(^1\) to counter welfare fraud via linkage and analysis of data about welfare recipients. Stichting Inlichtingenbureau can link welfare registry data to judicial registry data for purposes of stopping fugitive convicts from receiving welfare and informing the Dutch Ministry of Justice of the whereabouts of fugitives. Nowadays, Stichting Inlichtingenbureau also provides services to the Dutch water control boards (‘waterschappen’), Regional Coordinatiopoints

\(^1\)Website: http://www.inlichtingenbureau.nl
Fraud Control (‘RCF - Kenniscentrum Handhaving’), Regional Reporting and Coordination function school dropouts (RMC), Central Fine Collection Agency (CJIB), Social Insurances Bank (SVB), and bailiffs.

Risk reduction can, at least theoretically, pervert into seeking a risk-free society [38] and suppress behavior that is permissible but deviates from social norms. Not unlike the ‘chilling effect’, i.e. the stifling effect that overly broad laws [29], profiling and surveillance [39] are claimed to have on legitimate behavior such as exercising the constitutional right to free speech. Although we are unaware of scientific evidence for such causality (it is beyond our expertise), one only needs to consider practices in certain parts of the world to be convinced that (being aware of) the possibility of being scrutinized can cause a person to change her behavior. Think of a person not expressing a dissenting opinion near a microphone-equipped surveillance camera at work or in a public space where that person would otherwise have done so; or a person not traveling to the red light district, even if one needs to be there for some other than the obvious reason, due to fear of abuse of data collected by real-time vehicular registration systems and public transport smart card systems. Perhaps both find alternative ways to achieve their goal; but it seems unwise to assume that that is always the case, and then disregard the effects that technology and human-technology interaction can have on the human experience to which privacy is essential. The need for risk reduction and accountability at the collective level can be at odds with the need for privacy at the personal level; what constitutes the ‘right’ balance will depend on context.

Certain personal information is considered ‘sensitive’ because it can, and has often shown to, catalyze stigmatization, social exclusion and oppression: ethnicity, religion, gender, sexuality, social disease, political and religious preference, consumptive behavior, whether one has been victim or culprit of crime, and so on. The need for private life, also in terms of being able to keep certain information to oneself, is therefore neither new nor temporary, and worthy of defense. Reducing misunderstanding and mistreatment through means of public education, especially the promotion of reason, critical thinking and empathy, is one step forward; forbidding discrimination through legislation is another; enabling privacy impact assessment and control over the disclosure of information about oneself, especially sensitive information, the topic of our thesis, is yet another.

The rise of social media and ubiquitous computing does not imply the ‘end’ or ‘death’ of privacy. Rather, as Evgeny Morozov paraphrased from Helen Nissenbaum’s book [61] in The Times Literary Supplement of March 12th, 2010: “the information revolution has been so disruptive and happened so fast (...) that the minuscule and mostly imperceptible changes that digital technology

2According to a trend report issued by the Dutch governmental Research and Documentation Centre (WODC), 368 bailiffs and 414 junior bailiffs were active during 2005: https://www.wodc.nl/images/ob247-summary_tcm44-59825.pdf
has brought to our lives may not have properly registered on the social radar”. In her 2.5-year ethnographic study of American youngsters’ engagement with social network sites, Boyd observed that youngster’s “developed potent strategies for managing the complexities of and social awkwardness incurred by these sites” [8]. So, rather than privacy being irrelevant to them, the youngsters found a way to work around the lack of built-in privacy. In conclusion: privacy is not dead. At worst, it is in intensive care, beaten up by overzealous and sometimes careless use of technology. It will return to good health, even if merely for economical reasons [5].

It remains unclear when the desire to retreat first emerged, and even whether it is only found in humans. From an evolutionary or biological perspective, privacy might be explained by the claim that hiding oneself and one’s resources from predators and competitors in the struggle for existence is beneficial for survival. The desire to retreat, then, is perhaps as old as the struggle for existence itself. This notion, however, seems very distant from common ideas about privacy. With more certainty, sociological study has traced the emergence of withdrawal from classical antiquity — distinguishing between ‘religiously motivated quest for solitude’ and the ‘lay quest for private living space’ [86]. Alternatively, privacy can be conceived as a means to ‘personal security’.

What is clear, is that privacy has been thoroughly studied. The next Section will address notable concepts and terminology proposed in disciplines other than our own (technology, that is), establishing a broad background for our work3. We then proceed by mapping our work to specific parts of that theory. Finally, wrapping up this introduction, we state the scientific contributions of this thesis. Throughout this thesis, we will develop methods and techniques for the quantification and prediction of identifiability in support of the analysis of privacy problems regarding the disclosure, collection and sharing of personal information. The questionnaire mentioned above is an example scenario to which our work is relevant. More importantly, our work is relevant to computer databases, which tend to be linked to other databases via computer networks and can be exposed to those seeking authorized and unauthorized access to the data.

1.1 Terminology

From a legal perspective, one of the early and most well-known comprehensive works on privacy dates from 1890, when US Supreme Court Justices Warren and Brandeis published “The Right to Privacy” in the Harvard Law Review [85]. In the 20th century, Castle Doctrine emerged in legislation of self-defense of one’s private space [54] — its name referring to the proverb “a man’s house is his castle”. During the 1960s, Westin, a legal scholar who fo-

---

3Chapter 2 will establish the background within our own discipline.
1.1. TERMINOLOGY

cused on consumer data privacy and data protection, described four ‘states’ and four ‘functions’ of privacy [87, 38]. Figure 1.1 shows our mind-map of his conceptualization. The four functions, or ‘ends’, or ‘reasons’ for privacy that Westin distinguishes are *personal autonomy*, e.g. regarding decisions concerning personal lifestyle; *emotional release*, e.g. of tensions related to social norms; *self-evaluation*, e.g. extracting meaning from personal experiences; and *limited and protected communication*, e.g. disclosing information only to trusted others. The four states, or ‘means’ to privacy that Westin distinguishes are *anonymity*, e.g. ‘hiding’ within a group or crowd; *reserve*, e.g. holding back certain communication and behavior; *solitude*, e.g. seeking separation from others; and *intimacy*, e.g. seeking proximity to a small group.

![Privacy](image.png)

Figure 1.1: Privacy in ‘functions’ and ‘states’, according to Westin [87].

In the same era, Prosser, a legal scholar focusing on tort law, wrote that what had emerged from state and federal court decisions involving tort law were four different interests in privacy, or ‘privacy torts’ [66, 22]:

- intrusion upon the plaintiff’s seclusion or solitude, or into his private affairs;
- public disclosure of embarrassing private facts about the plaintiff;
- publicity which places the plaintiff in a false light in the public eye;
- appropriation, for the defendant’s advantage, of the plaintiff’s name or likeness.

More recently, in 2005, Solove, a legal scholar focusing on privacy, proposed a taxonomy of privacy violations that, unlike Prosser’s, does not only focus on tort law [74]. Figure 1.2 shows a map of that taxonomy. Solove describes the violations as follows. Categorized under *information processing* activity: *aggregation* comprises the combination of information about a person; *identification* comprises linking information to specific persons; *insecurity* comprises

---

4Note that throughout this thesis, we use the word ‘aggregation’ differently: we use it to mean generalization or grouping of data about different people.
lack of due diligence protecting (stored) personal information from leaks and improper access; secondary use comprises the re-use of information, without subject’s consent, for purposes different from the purpose for which it was originally collected; exclusion comprises not allowing the subject to know or influence how their information is being used. Categorized under information collection activity: surveillance comprises “watching, listening to, or recording of an individual’s activities”; interrogation comprises various forms of questioning or probing for information. Categorized under information dissemination activity: breach of confidentiality comprises “breaking a promise to keep a person’s information confidential”; disclosure comprises revealing (truthful) information that “impacts the way others judge [the] character [of the person involved]”; exposure comprises revealing “another’s nudity, grief, or bodily functions”; increased accessibility comprises “amplifying the accessibility of information”; blackmail comprises the threat to disclose personal information; appropriation comprises the use of the subject’s identity “to serve the aims and interests of another”; distortion comprises the dissemination of “false or misleading information about individuals”. Categorized under invasions: intrusion comprises acts that “disturb one’s tranquility or solitude”; decisional interference comprises “[governmental] incursion into the subject’s decisions regarding private affairs”. Section 1.2 will mention the violations that our work is primarily relevant to.

The last work we deem relevant as background to our research stems from 2010: Nissenbaum, a scholar in media, culture, and communication & computer science, conceptualized privacy as contextual integrity built from context-relative informational norms [61]. By that she means that whether some information flow constitutes a privacy violation, depends on its source context.
1.2. PROBLEM

— defined in terms of roles, activities, norms and values. We will reference Nissenbaum’s work again in Chapter 7.

1.2 Problem

We will now describe the specific research objectives that we address in this monograph. In an attempt to provide privacy, personal data that maps to single persons, i.e., microdata, is sometimes de-identified by removing ‘direct identifiers’ such as Social Security Numbers, names, addresses and phone numbers. De-identified data can still contain variables that, when combined, can be used to re-identify the de-identified data. Potentially-identifying combinations of variables are referred to as quasi-identifiers (QIDs) [21, 77]. The notion that quasi-identifiers can be used to re-identify people based on microdata poses questions on the usefulness of common de-identification procedures. Indeed, the question whether de-identification suffices to protect privacy in health research was recently posed in the American Journal of Bioethics [68].

Sweeney introduced the concept of $k$-anonymity, addressing this privacy risk by requiring that each quasi-identifier value (i.e., a combination of values of multiple variables) present in a data set must occur at least $k$ times in that data set, asserting that each record maps to at least $k$ individuals and hence obfuscating the link between records and individuals [77]. In common terminology, the group of $k$ individuals within which one is indistinguishable from $k - 1$ others is referred to as anonymity set (of size $k$) [64]. Motivated by the importance of privacy, as we argued, and considering the privacy risk posed from disclosure, collection and sharing of data about individual persons, we ask:

- To what extent is it possible to predict what (combined) information will turn out to be a perfect quasi-identifier, i.e., be unambiguously identifying for all persons in a group/population?
  
  — Example: “what is the probability that the combination of age, gender and (partial) postal code is uniquely identifying for all persons living in the postal code areas where my questionnaire is run?”

- For non-perfect quasi-identifiers, to what extent is it possible to predict the size of the anonymity sets?
  
  — Example: “what fraction of the citizens within this postal code area is uniquely identifiable by the combination of age and gender?”

These questions can be answered relatively easily if quasi-identifiers follow the uniform distribution: in that case, they can be directly translated to so-called birthday problems. In reality, however, data about persons tends to not follow a
uniform distribution; and for non-uniform distributions, the mathematics that one would use to answer these questions becomes considerably harder. To our knowledge, no method yet exists for efficient approximation of these privacy metrics for the case of non-uniform probability distributions.

One complicating factor in quasi-identifier analysis is the effect of correlation between various numerical personal data. What is the effect on anonymity of adding or removing a piece of information that correlates to an existing piece of information in a quasi-identifier, versus adding or removing information that is not correlated to other information?

Another complicating factor is the effect on anonymity of collecting and sharing less specific or more specific information. Being able to assess this beforehand supports informed decision-making about what data (not) to collect.

In terms of Solove’s taxonomy, these questions primarily map to violations of disclosure, aggregation and identification. The main stakeholders of these questions are the persons whose data is involved, the data holders, and the policy makers responsible for making privacy policy, potentially taking into account social norms that have not been made explicit in legislation. Chapter 7 will return to this.

1.3 Contribution

Now that we stated the problem, we proceed to state our contributions to addressing that problem. Many improvements have been proposed to $k$-anonymity, but only address the situation in which data has already been collected and must be de-identified afterwards. A question remains: “can we predict what information can be used for identification, so that we may decide not to collect it, beforehand?” Our contributions are as follows:

- Chapter 2 surveys existing literature on the analysis of anonymity. Several branches of research are identified. We specify to which branch our thesis relates, and justify our choice to do research within that branch;

- Chapter 3 builds our case by inquiring into the identifiability of deIdentified hospital intake data and welfare fraud data about Dutch citizens, using large amounts of data collected from municipal registry offices. We show that large differences can exist in (empirical) privacy, depending on where a person lives;

- Anonymity can be quantified as the probability that each member of a group can be uniquely identified using a QID. Estimating this uniqueness probability is straightforward when all possible values of a quasi-identifier are equally likely, i.e., when the underlying variable distribution is homogenous. In Chapter 4, we present an approach to estimate anonymity for the more realistic case where the variables composing a QID follow a
1.3. CONTRIBUTION

non-uniform distribution. Using birthday problem theory and large deviations theory, we propose an efficient and accurate approximation of the uniqueness probability using the group size and a measure of heterogeneity named Kullback-Leibler distance. The approach is thoroughly validated by comparing approximations with results from simulations based on the demographic data we collected for our empirical study;

- Where Chapter 4 addressed the problem of every member in a group being unambiguously identifiable, Chapter 5 proposes novel techniques for characterizing the number of singletons, i.e., the number of persons having 1-anonymity and are unambiguously identifiable, in the setting of the generalized birthday problem. That is, the birthday problem in which the birthdays are non-uniformly distributed over the year. Approximations for the mean and variance are presented that explicitly indicate the impact of the heterogeneity, expressed in terms of the Kullback-Leibler distance with respect to the homogeneous distribution, on anonymity. An iterative scheme is presented for determining the distribution of the number of singletons. Here, our formulas are experimentally validated using demographic data that is publicly available, allowing others to replicate our work;

- In Chapter 6, we study in detail three specific issues in singletons analysis. First, we assess the effect on identifiability of non-uniformity of value distributions in QIDs. Suppose one knows the exact age of every person in a group; what is the effect on identifiability that some ages occur more frequently than others? Again, it turns out that the non-uniformity can be captured well by a single number, the Kullback-Leibler distance, and that the formulas we propose for approximation produce accurate results. Second, we analyze the effect of the granularity chosen in a series of experiments. Clearly, revealing age in months rather than years will result in a higher identifiability. We present a technique to quantify this effect, explicitly in terms of interval width. Third, we study the effect of correlation between the quantities revealed by the individuals; the leading example is height and weight, which are positively correlated. For the approximation of the identifiability level we present an explicit formula, that incorporates the correlation coefficient. We experimentally validate our formulae using publicly available data and, in one case, using the non-public data we collected in the early phase of our study;

- As a starting point for discussion, Chapter 7 gives preliminary ideas on how our work might fit in real-life society, taking into account various practical considerations.

We conclude our thesis in Chapter 8.
Appendix A contains a key intermediate result from Chapter 5, and shows, for varying $k$ and $N$, the probability that no singletons exists in a group of $k$ members that are uniformly assigned one of $N$ possibilities; i.e., the chance that no person within a group can be uniquely identified by some uniformly distributed quasi-identifier.

Appendix B discusses, as toy example, a non-sensitive anonymous questionnaire that was observed in real life. It explains how respondent anonymity degrades for each demographic that the respondent discloses. This Appendix is intended to inspire the reader to think about scenarios where analysis of anonymity is relevant.