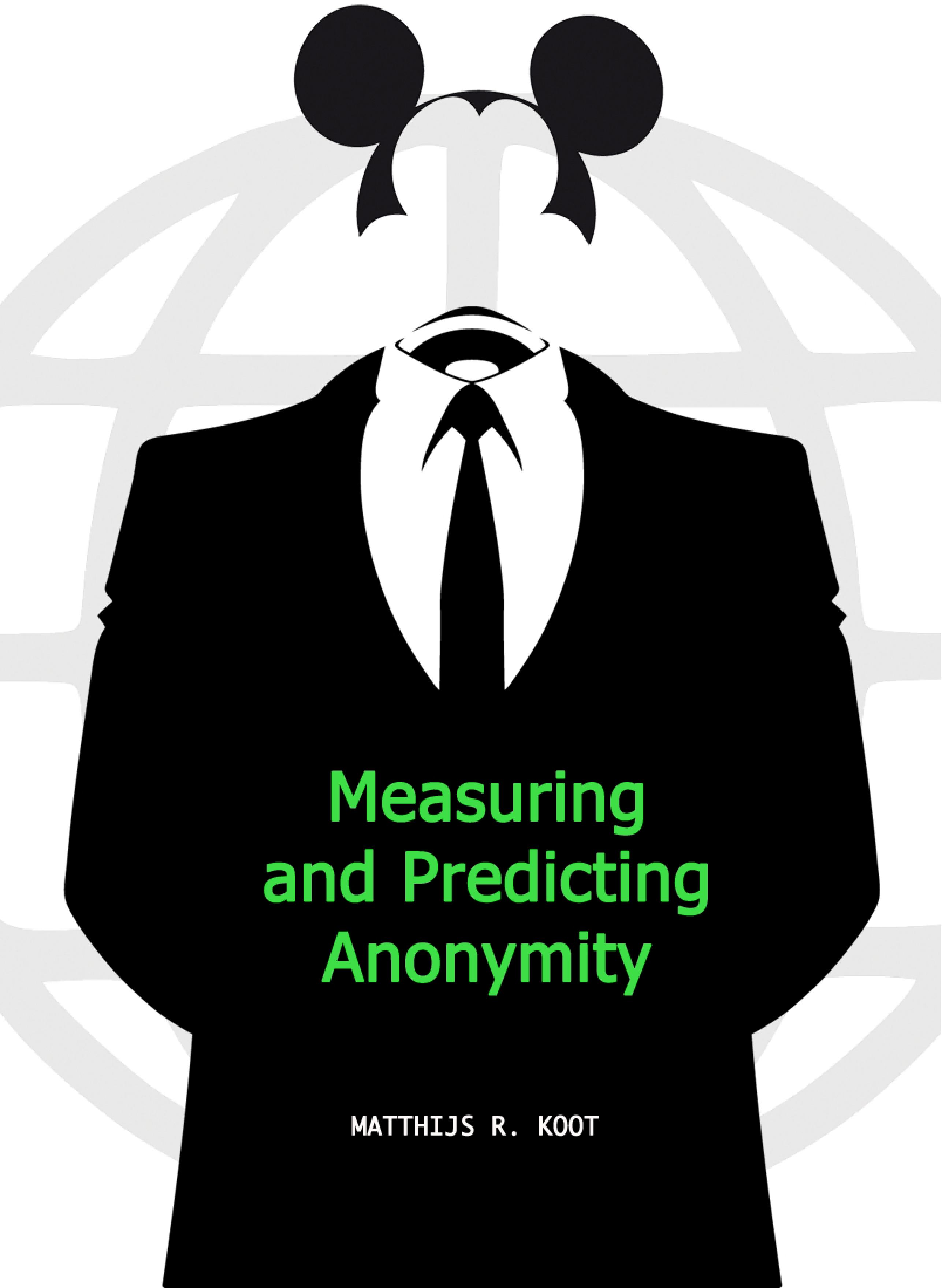


Measuring and Predicting Anonymity

MATTHIJS R. KOOT

Measuring and Predicting Anonymity

Measuring and Predicting Anonymity

Matthijs R. Koot

UNIVERSITEIT VAN AMSTERDAM

University of Amsterdam
Informatics Institute
Science Park 904
1098 XH Amsterdam

http://www.science.uva.nl/ii/

Measuring and Predicting Anonymity

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op woensdag 27 juni 2012, te 10.00 uur

door

Matthijs Richard Koot

geboren te Leeuwarden.

Promotor:	Prof.dr.ir. C.T.A.M. de Laat
Promotor:	Prof.dr. M.R.H. Mandjes
Overige leden:	Prof.dr. J.A. Bergstra
	Prof.dr.ir. C. Diaz
	Prof.dr.ir. B.J.A. Kröse
	Prof.dr. R.D. van der Mei
	Prof.dr. R.R. Meijer
	Prof.dr. L.A. Sweeney

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C012854

Cover design by Robert Koot.
Printed by Gildeprint Drukkerijen - The Netherlands.

The need for private life is neither new nor temporary,
and worthy of defense.

to my parents Giel and Ina Koot

and my dear brother Robert.

Contents

Acknowledgments **ix**

1 Introduction **1**

1.1 Terminology 4

1.2 Problem 7

1.3 Contribution 8

2 Background **11**

2.1 Early concepts 13

2.2 Information theory 16

2.3 Process calculi 18

2.4 Epistemic logic 21

2.5 k-Anonymity 23

2.6 Discussion 25

3 Empirical study **27**

3.1 Introduction 27

3.2 Background 29

3.3 Results 32

3.4 Discussion 39

4 Estimating uniqueness **41**

4.1 Introduction 41

4.2 Problem 43

4.3 Methodology: birthday problems 43
4.4 Experiments with demographic data 47
4.5 Discussion and future work 52

5 Analysis of singletons 55
5.1 Introduction 55
5.2 Mean number of identifiable objects 56
5.3 Variance of the number of identifiable objects 62
5.4 Probability of at least one singleton 65
5.5 Numerical experiments 68
5.6 Concluding remarks 68

6 Correlation and aggregation 71
6.1 Introduction 71
6.2 Analysis of singletons 73
6.3 Impact of interval-width 78
6.4 Multivariate distributions 83
6.5 Discussion 85

7 Practical applications 91
7.1 Preliminary model 92
7.2 Issues 96
7.3 What steps to take next 97
7.4 Other aspects 98
7.5 Limitations and future work 98
7.6 Conclusion 99

8 Conclusions and future work 101

A $\zeta(k,N)$ for $k = 1\ldots,50$ and $N = 1,\ldots,20$ 103

B Example analysis: questionnaire 105

C Publications 113

Bibliography 115

List of Figures 123

List of Tables 125

Abstract (English) 127

Abstract (Dutch) 131

Acknowledgments

I owe a large debt to prof.dr.ir. Cees de Laat, prof.dr. Michel Mandjes and Guido van ’t Noordende, whose help proved invaluable in what turned out to be a highly interesting and challenging endeavor. Luctor et emergo: I struggle and arise.

Amsterdam
June 2012.

Matthijs R. Koot

1 Introduction

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.

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 Coordinationpoints

[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[2].

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

[2] According 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 work[3]. 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-

[3] Chapter 2 will establish the background *within our own discipline.*

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.

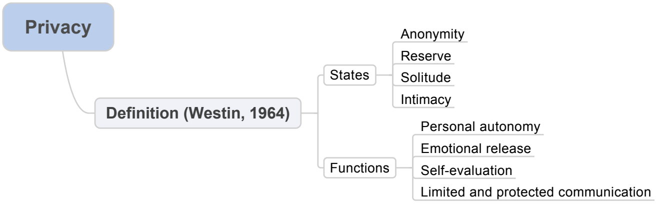


Figure 1.1: Privacy in 'functions' and 'states', according to Westin [87].

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[4]; *identification* comprises linking information to specific persons; *insecurity* comprises

[4] Note that throughout this thesis, we use the word 'aggregation' differently: we use it to mean generalization or grouping of data about different people.

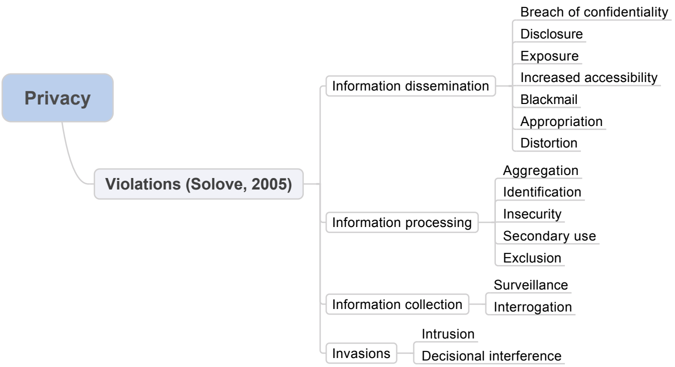


Figure 1.2: Taxonomy of privacy violations according to Solove [74].

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

— 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 who's 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 de-identified 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 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.

2 Background

This Chapter presents a study of existing literature on the analysis of anonymity. Section 2.5 will introduce k-anonymity, a concept that will be referred to repeatedly throughout this thesis. Busy readers may skip to that Section without risking unintelligibility of the remainder of this thesis.

Information systems for applications such as electronic voting, clinical healthcare and medical research should provide reliable security and privacy. Formal methods are useful to verify or falsify system behavior against specific properties, including aspects of security and privacy. The mathematics that underlie formal methods provide a more solid foundation for IT engineering than informal methods do; an important reason for this is the disambiguating and computer-verifiable nature of mathematical notation. Systems that are built on (or using) formal methods are thus expected to be more reliable[1].

We apply the vocabulary proposed by Pfitzmann and Hansen [64]. On December 2nd, 2011 the Internet Architecture Board announced[2] adoption of this document with the“[aim] to establish a basic lexicon around privacy so

[1] However, one must take into account that formal modeling remains a human activity and is, therefore, prone to human error, that mathematical specification of aspects about vague concepts like security and privacy is a difficult task and that in practice, typically only parts of systems can be proven correct due to the subtleties and complexity of real-life environments.

[2] `http://www.iab.org/2011/12/02/draft-on-privacy-terminology-adopted/` and `http://tools.ietf.org/html/draft-iab-privacy-terminology-00`.

that IETF contributors who wish to discuss privacy considerations within their work can do so using terminology consistent across the area". Note that this vocabulary did not exist before 2000 and has been scarcely referred to. It is sometimes difficult to compare existing literature without re-explaining the use of language. Key definitions:

Definition 2.1 *Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set.*

Citing from [64]: "[being] 'not identifiable within the anonymity set' means that only using the information the attacker has at his discretion, the subject is 'not uniquely characterized within the anonymity set'. In more precise language, only using the information the attacker has at his discretion, the subject is 'not distinguishable from the other subjects within the anonymity set'."

Definition 2.2 *Anonymity of a subject* from an attacker's perspective *means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set.*

Definition 2.3 *Unlinkability of two or more Items of Interest (IOIs, e.g., subjects, messages, actions, ...) from an attacker's perspective means that within the system (comprising these and possibly other items), the attacker cannot sufficiently distinguish whether these IOIs are related or not.*

The size of the anonymity set in Definitions 2.1 and 2.2 is the unit of measurement used throughout our work.

Privacy research related to electronic systems can roughly be divided in two topics:

- Data anonymity: unlinkability of an individual and (anonymized) data about him/her in databases;
- Communication anonymity: unlinkability of an individual and his/her online activity.

From Definition 2.2 it follows that anonymity is relative to a specific point of view: it depends on what the attacker knows *a priori* or can learn *a posteriori* about the system, its environment and its users.

The remainder of this Chapter is organized as follows: Section 2.1 describes early concepts; Section 2.2 refers to applications of information theory to research on anonymity; Section 2.3 refers to applications of process calculus; Section 2.4 refers to applications of epistemic logic; and Section 2.5 introduces to k-anonymity, a concept that will be used intensively throughout this thesis.

2.1 Early concepts

For the last two decades, research on identity hiding has largely been orbiting around the concept of a *mix* introduced by Chaum [18]. A mix is a system that accepts incoming messages, shuffles, delays and permutes them, and sends them to either the intended recipient or the next mix. The purpose of the intermediate processing is to provide anonymity. *What* anonymity is provided, to *whom*, to which *degree* and under what *assumptions* depends on the parameters of the mix design and the context of its usage.

Many mix systems have been proposed with subtle variations on the parameters of shuffling, delaying and permutation — 'permutation' being the use of cryptography to change message content so that to an observer, the input messages are, in terms of content, unlinkable to output message. Those parameters are dictated by either the purpose of the system (e.g. anonymous e-mail, anonymous file sharing, anonymous voting) or by assumptions about the conditions under which the system will be used (e.g. a specific threat model, need for interoperability with other systems, latency/throughput conditions).

Message-based mixes are designed to anonymize the communication of one-off, independent, potentially large-sized messages; such systems are typically designed to have high-latency and low-bandwidth properties. *Connection-based mixes* are designed to anonymize the communication of streams of small messages (e.g. packets); such systems are typically designed to have low-latency and high-bandwidth properties. It is sometimes mentioned that there is a trade-off between latency and anonymity, where high latency is associated with stronger anonymity, and low latency with weaker anonymity.

Two anonymity protocols that are often used to demonstrate formalizations of communication anonymity related to mixes are the *Dining Cryptographers* protocol by Chaum in 1988, and the *FOO92* voting scheme by Fujioka, Okamoto and Ohta in 1992 [17, 30]. A description of those protocols is beyond the scope of this thesis.

2.1.1 Degrees of anonymity

Anonymity is not a binary property; it is not either present or absent. Rather, a subject is more easily or less easily identifiable at any given time, and anonymity is a point on a scale. In 1998, Reiter and Rubin proposed a scales for degrees of anonymity, as depicted in Figure 2.1 [67]. This scale is an informal notion, but has aided discussion about anonymity systems.

Both in their original paper and most work that refers to that paper, a focus is given to three intermediate points (citation from [67]):

- *Beyond suspicion: A sender's anonymity is beyond suspicion if though the attacker can see evidence of a sent message, the sender appears no*

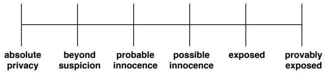


Figure 2.1: Degrees of anonymity according to Reiter and Rubin [67]

more likely to be the originiator of that message than any other potential sender in the system.

- *Probable innocence: A sender is probably innocent if, from the attacker's point of view, the sender appears no more likely to be the originator than not be the originator. This is weaker than beyond suspicion in that the attacker may have reason to expect that the sender is more likely to be responsible than any other potential sender, but it still appears at least as likely that the sender is not responsible.* Or: to the attacker, the subject has less than 50% chance of being the culprit.
- *Possible innocence: A sender is possibly innocent if, from the attacker's point of view, there is a nontrivial probability that the real sender is someone else.* Or: to the attacker, the subject has less than 100% chance of being the culprit.

Halpern and O'Neill proposed a formal interpretation of such a scale using epistemic logic [33]. The authors use notations such as $K_i\varphi$ to model that agent i knows φ, and $P_i\varphi$ to model that agent i thinks that φ is possible. The formula $\theta(i, a)$ is used to represent "agent i has performed action a, or will perform a in the future". For example:

> Action a, performed by agent i, is *minimally anonymous* with respect to agent j in the interpreted system $\mathcal{I}$, if $\mathcal{I} \models \neg K_j[\theta(i, a)]$.

In this example, the agent i is minimally anonymous with respect to agent j if agent j does not know that agent i has performed action a. Another example:

> Action a, performed by agent i, is *totally anonymous* with respect to agent j in the interpreted system $\mathcal{I}$, if $\mathcal{I} \models \theta(i, a) \Rightarrow \bigwedge_{i' \neq j} P_j[\theta(i', a)]$.

In this example, the agent i is 'totally anonymous' with respect to agent j if agent j thinks it is possible that the action could have been performed by *any* of the agents. Note that this assumes that i and j are not the only two agents: otherwise, agent j knows that agent i must have performed the action.

Chatzikokolakis and Palamidessi proposed a revised formalization of probable innocence, building on the formalism of probabilistic automata [16]. Citing from [16]: "A probabilistic automaton consists in a set of states, and labeled transitions between them. For each node, the outgoing transitions are partitioned in groups called steps. Each step represents a probabilistic choice, while the choice between the steps is nondeterministic". The authors model anonymity by considering the execution paths of the automata across probabilistic action sets. The main contribution is that the authors' notion conveys both limits on an attacker's confidence in knowing which subject belongs to an observed event, and on the probability of detection.

2.1.2 Possibility, probability, and determinism

In anonymity theory, the notions of *determinism*, *non-determinism*, *possibility* and *probability* refer to choice types that are present in a system.

Deterministic models represent systems of which behavior only depends on internal states and is, therefore, predictable: at any given state, for some given (deterministic) input, there is only one possible transition. The system behaves the same for each execution.

Non-deterministic models represent systems of which behavior depends on some unpredictable external state and is, therefore, unpredictable itself; or at least very difficult to predict. Examples of external states are user input, schedulers, hardware timers/timing-sensitive programs, random variables and stored disk data. For anonymity, users and random number generators are two typical examples of non-deterministic aspects. *Angelic non-determinism* models choices as if the inputs are not arbitrary, but are always biased to guarantee success ('good' behavior). *Demonic non-determinism* models choices as if they are arbitrary, and never made with guarantee for success ('malicious' or 'ignorant' behavior).

Possibilistic models represent systems in which at any given state, there are N states to which transition is possible (N might be 1). No notion is made regarding the probability of each transition. In contrast to deterministic models, possibilistic models allow uncertainty; the models just do not explicitly describe it.

Probabilistic models are possibilistic models *with* probabilities. A probabilistic choice represents a set of alternative transitions where each transition is assigned a probability of being chosen; in contrast, a non-deterministic model has no notion of probability.

2.1.3 Anonymity set size

The most basic way to quantify anonymity is to use the anonymity set size. Suppose a message M was sent by subject s_1 from anonymity set S of size N, and suppose an attacker that detected M at the recipient but has no other knowledge. In the anonymity set size metric, anonymity is then quantified as

$$\text{anonymity set size} = \frac{1}{N} \tag{2.1}$$

For a set of size $N = 10$, the attacker can link M to s_1 only to a certainty of $\frac{1}{10}$. This metric assumes a uniform distribution of probabilities, and cannot be applied to situations where this equidistribution is not present. As most real-life systems deal with heterogeneous sets of subjects, this assumption almost never holds, and thus more refined metrics are needed.

2.2 Information theory

This Section refers to existing literature on the application of Shannon-entropy and Rényi-entropy to research on anonymity.

2.2.1 Shannon-entropy

In 2002, Serjantov and Diaz independently proposed the use of Shannon-entropy to establish anonymity metrics that lift the equiprobability requirement [72]. Shannon-entropy quantifies the level of uncertainty inherent in a set of data. In its (proposed) application to anonymity, the 'set of data' is the probability distribution over the possible links between a message M and its possible senders[3] S. It assumes that an attacker is able to estimate probabilities *a posteriori* after observing the system[4]. The Shannon entropy equation provides a way to estimate the average minimum number of bits needed to encode a string of symbols, based on the frequency of the symbols. Anything can be a symbol: letters like $\{A, B, C, ...\}$, persons like $\{subject_1, ...subject_n\}$, colors like $\{red, green, blue, ...\}$, et cetera. The (finite) set of possible symbols are referred to as the *source alphabet*. According Shannon, on average, the number of bits needed to represent the result of an uncertain event (e.g. production of a symbol) is given by its entropy. The Shannon-entropy formula:

$$H(S) = -\sum_{i=1}^{N} p(s_i) \log_2 p(s_i) \tag{2.2}$$

For anonymity, $H(S)$ (the H-symbol is borrowed by Shannon from Boltzmann H-Theorem in thermodynamics) denotes the number of additional bits the attacker needs to *perfectly* link a message M to its sender subject s_i from set S with size N (note that in the Pfitzmann-Hansen definition of 'anonymity from

[3] The proposed work only regards sender-anonymity; however, it may be suitable to measure receiver-anonymity or relationship-anonymity as well.

[4] 'Observing' might include passive attacks like statistical analysis, and/or active attacks like repetitive querying or other experiments to deduce knowledge.

an attacker's perspective', a subject is already non-anonymous if an attacker is able to 'sufficiently' identify the subject, and the attacker might very well be satisfied by a less-than-perfect link). To apply this probabilistic metric, the attacker has to assign a probability $p(s_i)$ to each subject s_i, where $p(s_i)$ is a value between 0 and 1 and $\sum_{i=1}^{N} p(s_i) = 1$. Suppose a particular $p(s_i) = 1$, then all the other $p(s_i)$ are 0 and $H(S) = 0$; this means the attacker has a perfect link. If all $p(s_i)$ are equal, the metric 'reduces' to the basic anonymity set size metric $H(S) = \log_2 |S|$.

The *degree of anonymity* is a quantification of the amount of information the system leaks about the probability distribution. The higher the degree, the less information is leaked. The maximum entropy of the system is expressed as H_M:

$$H_M = \log_2(N) \tag{2.3}$$

The degree d is a value between 0 and 1 and is determined by $H_M - H(S)$, then normalized by dividing by H_M:

$$d = 1 - \frac{H_M - H(S)}{H_M} = \frac{H(S)}{H_M} \tag{2.4}$$

Here, $d = 0$ if an attacker can link message M to its originating subject with probability 1, and $d = 1$ if it is equally likely to originate from any subject from S.

For example: suppose a system with an anonymity set of size $N = 10$, then maximum entropy $H_M = \log_2(10) \approx 3.32$ bits. Suppose that based on the outcome of passive or active observation of the system, the attacker estimates/deduces that s_4 is 10 times more likely to be the sender than the other nine subjects. The attacker will assign $p(s_4) = 0.5$ while keeping the rest uniform at $p(s_i) = \frac{1-0.5}{9} \approx 0.055$: then $H(S) \approx 2.58$ bits and the degree of anonymity $d = \frac{2.58}{3.32} \approx 0.77$. So, despite the single peak in probability assigned by s_4, the attacker is still lacking 2.58 bits of information needed to be fully confident and the system still provides a degree of anonymity 0.77 (with 1 being maximum). Indeed, this metric could also be applied as a measure of attack efficiency by using it to determine differences in unobservability. ('Unobservability' meaning "undetectability of an [Item of Interest (IOI, e.g., subjects, messages, actions, ...)] against all subjects uninvolved in it, and anonymity of the subject(s) involved in the IOI even against the other subject(s) involved in that IOI" [64].)

2.2.2 Rényi-entropy

Tóth, Hornák and Vajda argued that for some purposes of anonymity quantification a worst-case metric is preferable over the average case metric that Shannon-entropy provides [80]. In 2006, based on this notion, Clauß and

Schiffner proposed the use of Rényi-entropy as a generalization of Shannon-, Min- and Max-Entropy (and the authors provide the mathematical proof for this generalization) [20]. The Rényi-entropy formula:

$$H_\alpha(P) = \frac{1}{1-\alpha} \log_2 \sum_X p_i^\alpha \tag{2.5}$$

Here, the more α grows, the more $H_\alpha(P)$ approaches Min-Entropy (Min-Entropy is the situation where the attacker is certain that one subject is the originator and hence that the other subjects cannot possibly be the originator). The more α approaches zero, the more $H_\alpha(P)$ approaches Max-Entropy (Max-Entropy is the situation where from the attacker standpoint, all subjects are equally likely to be the originator). The more α approaches one, the more $H_\alpha(P)$ approaches Shannon-Entropy.

To overcome the strong influence of outliers, the authors propose the use of quantiles. Quantiles allow that lower bound outliers are cut off. With regard to this anonymity metric, it allows statements like: “10 bits of information are needed to address 90% of the source elements”. Whereas with Shannon-entropy, one can only make a statement regarding *all* of the source elements, and has to accept that the statement can be strongly influenced by outliers.

2.3 Process calculi

Process calculi are algebraic notations that can be used to (formally) model concurrent systems. They are typically associated with the area of theoretical computer science. The three major branches of process calculi are the *Calculus of Communicating Systems*, or *CCS* [55], *Communicating Sequential Processes*, or *CSP* [37] and *Algebra of Communicating Processes*, or *ACP* [6].

The word *process* refers to the *behavior of a system.* To cite formal methods researcher Jos Baeten, behavior is “the total of events or actions that a system can perform, the order in which they can be executed and maybe other aspects such as timing or probabilities” [4]. Process calculi try to capture different ways in which concurrent systems can be designed in terms of process creation (fork/wait, cobegin/end, etc), information exchange between processes (message passing, shared variables) and management of shared resources (semaphores, monitors, transactions, etc.) [65].

Considering that security and privacy are typically about concurring parties, concurrent processes are an intuïtive way to model security and privacy protocols, and process calculi have indeed been used extensively to formally define security properties and verify cryptographic protocols [65]. The following subsections describe examples of this.

2.3.1 Communicating Sequential Processes

In 1996, Schneider and Sidiropoulos proposed a definition of anonymity in CSP [71]. In CSP, systems are modeled in terms of *processes* that operate independently and interact with each other to perform *events* solely by passing *messages*. Events represent atomic communications or interactions. Processes are described in terms of the events that they may engage in. CSP is purely non-deterministic and has no notion of probability.

In the Schneider and Sidiropoulos model, anonymity is concerned with protecting the identity of users with respect to particular events or messages. They consider CSP trace semantics and use features of CSP to model anonymous message sending: parallel concurrent processes represent the anonymity set, and hidden events represent anonymous message sending (in theory, hiding an event makes it unobservable). If the sequences of events that are observable to an attacker are identical for any run (since the anonymous event was hidden), the result of the anonymous event is considered unlinkable to a specific process.

$$A = \{i.x | i \in \mathit{USERS}\}$$

A is the set of events that are supposed to be anonymous, and, therefore, will be *hidden*. An event $i.x$ is composed of its content x and the identity i of the agent that communicates it. *USERS* represents the users who want to communicate anonymously. Some process P provides anonymity if an arbitrary permutation P_A of the events in A, applied to the observables of P, does not change the observables:

$$P_A(\mathit{Obs}(P)) = \mathit{Obs}(P)$$

The authors demonstrate their model in automatic verification of the anonymity provided by the Dining Cryptographers protocol, using the *Failure Divergence Refinement* model-checking tool for CSP state machines.

2.3.2 π-calculus

π-calculus is a process calculus originally developed by Milner, Parrow and Walker as a continuation of CCS [56]. Its purpose is to describe concurrent systems whose configuration may change during execution. The main difference between π-calculus and earlier process calculi is that the former allows the passing of channels as data through other channels. This feature, called *mobility*, allows the network to change with interaction; i.e., it allows that topology changes after some input.

π-calculus can be used to represent processes, parallel composition of processes, synchronous communication between processes through channels, creation of new channels, replication of processes and non-determinism. Probabilistic π-calculus also allows representation of probabilistic aspects. In π-calculus there are two basic actions:

"c!x" : send value x on channel c (output action).
"c?x" : receive value x on channel c and bind it to the name x (input action).

2.3.3 μCRL / mCRL2

Chothia, Orzan, Pang and Dashti proposed a framework for automatically checking anonymity based on the process-algebraic specification language μCRL, which is based on Bergstra's ACP [19]. The authors introduce the notions of *player anonymity* and *choice anonymity*. *Player anonymity* refers to the situation where an attacker observed a certain event (e.g. a choice), and wants to link that event back to the originating subject(s). *Choice anonymity* refers to the situation where an attacker observed a subject, and wants to know which event(s) belong(s) to that subject.

The authors take the view that when participants in a (group) protocol wish to remain anonymous the authors wish to hide parts of their behavior and data; and state that a group protocol can be written as a parallel composition of participants and an environment process. Here, P and Q are process models written in μCRL, with P representing the player behavior and Q the environment (made up of entities that 'oversee' the protocol):

$$\text{Protocol(x)} = P_1(x_1) \parallel P_2(x_2) \parallel ... \parallel P_n(x_n) \parallel Q(n)$$

Here $x = (x_1, x_2, ..., x_n)$ is the *choice vector* of possible choices from a known domain; anonymity refers to the link between this value and the identity of the participant using it. The authors provide the following definitions of anonymity:

Choice indistinguishability: *Let* Protocol *be the specification of a protocol,* v_1 *and* v_2 *two choice vectors, and* Obs *an observer set. The set of all possible choice vectors is denoted by* CVS*. Then the relation* $\approx_{Obs}$: CVS × CVS *is defined as:*

$$v_1 \approx_{Obs} v_2 \text{ iff } \text{Protocol}_{Obs}(v_1) \approx \text{Protocol}_{Obs}(v_2).$$

Choice anonymity degree: *The choice anonymity degree (*cad*) of participant i w.r.t. an observer set* Obs *under the choice vector x is:*

$$\text{cad}_x(i) = |\{c \in \text{Choices}, \exists v \in \text{CVS} \textit{ such that } v_i = c \textit{ and } v \approx_{Obs} x \textit{ and } \forall_j \in Obs.v_j = x_j\}|$$

where $|\cdot|$ *denotes the cardinality of a set,* Choices *is the set of all possible choices,* CVS *is the choice vector set,* $v = \langle v_1, ..., v_n \rangle$ *and* $x = \langle x_1, ...x_n \rangle$*. We define the choice anonymity degree of participant i w.r.t.* Obs *as*

$$\mathrm{cad}(i) = \min_{x \in \mathrm{CVS}} \mathrm{cad}_x(i)$$

Player anonymity degree *The player anonymity degree (*pad*) of secret choice c, in a protocol with n players, w.r.t. an observer set* Obs *and the choice vector x is:*

$$pad_x(c) = |\{i \in \{1, ..., n\} \setminus Obs, \exists v \in \mathrm{CVS} \textit{ such that } v_i = c \textit{ and } v \approx_{Obs} x \textit{ and } (\forall_j \in Obs.v_j = x_j)\}|.$$

The player anonymity degree of secret choice c w.r.t. an observer set Obs *is*

$$pad(c) = \{0, \min_{x \in \mathrm{CVS}_{pad_x(c)} > 0} pad_x(c), otherwise$$

These definitions allow a precise way of describing the different ways that anonymity can break down, e.g. due to colluding insiders.

2.3.4 Other developments

Bhargava and Palamidessi proposed a notion of anonymity based on conditional probability, called *probabilistic anonymity.* The authors take into account both probability and non-determinism [7] and provide a mathematically precise definition by applying probabilistic π-calculus.

Deng, Pang and Wu proposed a probabilistic process calculus for describing protocols ensuring anonymity, and a notion of relative entropy to measure the degree of anonymity that can be guaranteed [25]. The authors quantify the amount of probabilistic information an anonymity protocol reveals and take both a priori and a posteriori knowledge into account, i.e. both knowledge that the attacker has about a system and its users beforehand, and the knowledge that the attacker learns from observing the protocol execution.

Deng, Palamidessi and Pang demonstrated the use of PRISM/PCTL for automatic verification of the notion of *weak anonymity* [24]. *Weak* refers to the notion that some amount of probabilistic information may be revealed by a protocol, e.g. through presence of attackers who interfere with the normal execution of the protocol or through some imperfection of the internal mechanisms. The authors study the degree of anonymity that a protocol can still ensure, despite the leakage of information.

Hasuo and Kawabe proposed *anonymity automata* as a means to provide simulation based proof of the notion of probabilistic anonymity introduced by Bhargava and Palamidessi [36].

2.4 Epistemic logic

Logic investigates and classifies the structure of arguments. Modal logic allows arguments with modalities such as necessity and possibility. Epistemic logic is a form of modal logic that is concerned with propositions of knowledge, uncertainty and ignorance. To anonymity, epistemic logic for multi-agent systems is most relevant. Epistemic logic extends propositional logic by adding an operator K to express the knowledge held by an agent (we use the terms agent and subject interchangeably). It is thereby possible to make statements such as:

$K_s p$: “subject s knows proposition p (and that it is true).”
$K_s \neg p$: “subject s knows that proposition p is false.”
$\neg K_s p$: “subject s does not know proposition p.”
$\neg K_s \neg p$: “subject s does not know that proposition p is false.”

Anonymity of an agent is defined as the uncertainty of the observer regarding a particular proposition which models sensitive information belonging to that agent. Epistemic analysis of multi-agent communication consists of [82]:

1. representing the initial knowledge or beliefs of the agents in a semantic model (e.g. in a so-called *Kripke structure* [46] using labels for individual agents and valuations for states);

2. representing the operations on the knowledge or beliefs of the agents as operations on semantic models;

3. model checking, to see if given formulas are true in the models that result from given updates.

Syverson and Stubblebine proposed the use of *group principals* as an approach to model anonymity in epistemic logic of multi-agent systems [78]. This means that knowledge can be modeled as a property of a group, rather than of an individual agents. Four types are proposed: a *collective group principal* that is expressed as $\star G$ (what this group knows is what is known by combining the knowledge of all the group members), an *and-group principal* that is expressed as $\& G$ (what this group knows is what is commonly known by all of its members, e.g. the common denominator), the *or-group principal* that is expressed as $\oplus G$ (what this group knows is what at least one member of the group knows) and the *threshold group principal* that is expressed as $n - G$ (what this group knows is anything known by any collective subgroup contained in G of cardinality at least n). They apply a small formal language to define anonymity properties *($(\geq n)$-anonymizable*, *Possible Anonymity*, *$(\leq n)$-suspected*, *$(\geq n)$-anonymous* and *Exposed*) using the group principals concept, specify an anonymity protocol similar to the Anonymizer.com anonymous web proxy service and assess the protocol against the anonymity properties. This work considers only possibilistic aspects.

Halpern and O'Neill proposed an alternative definition of anonymity using epistemic logic of multi-agent systems [33]. The authors build on earlier work in which a *runs and systems* framework was proposed for the analysis of security systems [34]. Anonymity is defined as the absence of specific knowledge at the observing agent about the anonymous agent and the actions the agent performs. This work considers probabilistic aspects. The authors include the following definitions, where Pr_j is a probability assigned by the attacker based on observations (i.e., assigned *a posteriori*), to the possibility θ that agent i executed action a):

> α-***anonymous:*** *Action a, performed by agent i, is α-anonymous with respect to agent j if $\mathcal{I} \vDash Pr_j[\theta(i,a)] < \alpha$.*
>
> ***Strongly probabilistically anonymous:*** *Action a, performed by agent i, is strongly probabilistically anonymous up to $\mathcal{I}_A$ with respect to agent j if for each $i' \in I_A, \mathcal{I} \vDash Pr_j[\theta(i,a)] = Pr_j[\theta(i,a)]$.*

Van Eijck and Orzan proposed the use of Dynamic Epistemic Logic (DEL) to model anonymity [82]. DEL distinguishes itself from other epistemic logics by the introduction of action models, which are Kripke structures describing information updates corresponding to various forms of communications [46]. These action models allow more intuitive specification, or even visualization, of the flow in a knowledge program, thus making it easier to express complex concepts like security and anonymity [82]. The authors propose a DEL verification method, provide automata-based tooling based on the μCRL toolset and the Construction and Analysis of Distributed Processes (CADP) model checker, and apply them to verify anonymity within the Dining Cryptographers and FOO92 protocols.

2.5 k-Anonymity

Over a decade ago, Sweeney proposed k-anonymity, a non-probabilistic metric for anonymity concerning entries in statistical databases such as released by data holders for research purposes [76, 77]. Sweeney's interest is in re-identifiability of persons based on their entries in such databases, e.g. through inferences over multiple queries to the database or linking between different databases (as depicted in Figure 2.2). A statistical database provides k-anonymity protection if the information for each person contained within cannot be distinguished from at least $k-1$ 'other individuals who appear in the database.

Sweeney applies set-theory to formalize the notions of a table, rows (or 'tuples') and columns (or 'attributes'), and the quasi-identifier concept introduced by Dalenius [21]. A quasi-identifier is a set of attributes that are individually anonymous, but in combination can uniquely identify individuals. Sweeney defines 'quasi-identifier' as follows [77] (note: throughout this thesis, we use 'quasi-identifier' in the less formal definition provided in Section 1.2):

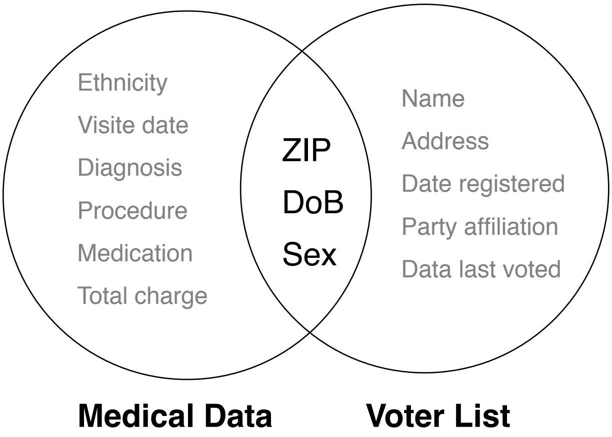


Figure 2.2: Linking to re-identify data [76]

Attributes. Let $B(A_1, ..., A_n)$ be a *table* with a finite number of tuples. The finite set of *attributes* of B is $\{A_1, ...A_n\}$.

Quasi-identifier. Given a population of entities U, an entity-specific table $T(A_1, ..., A_n)$, $f_c : U \rightarrow T$ and $f_g : T \rightarrow U'$, where $U \subseteq U'$. A quasi-identifier of T, written as Q_t, is a set of attributes $\{A_i, ..., A_j\} \subseteq \{A_1, ..., A_n\}$ where: $\exists p_i \in U$ such that $f_g(f_c(p_i)[Q_t]) = p_i$.

k**-Anonymity**. Let $RT(A_1, ..., A_n)$ be a table and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI_{RT}]$ appears with at least k occurrences in $RT[QI_{RT}]$.

The k-anonymity model assumes a global agent to calculate the metric. It also depends on the data holder's competence and willingness to correctly identify and work around quasi-identifiers. k-Anonymity protects against the 'oblivious‘ adversary targeting *anyone* (re-identifying anything he can, hoping to get lucky) as well as the adversary targeting a *specific individual.* One of the limitations of the original k-anonymity model is that it does not take into account the situation where the sensitive attribute has the same value for all k rows and is revealed anyway. l-Diversity was introduced to address this by requiring that, for each group of k-anonymous records in the data set, at least l different values occur for the sensitive column [50]. Further developments included t-closeness, m-invariance, δ-presence and p-sensitivity [10, 48, 59, 90]. Applications of k-anonymity to communication anonymity in mobile ad-hoc networks and overlay networks have been explored in [84, 89].

[49] provides a probabilistic notion of k-anonymity: a dataset is said to be probabilistically $(1-\beta, k)$-anonymous along a quasi-identifier set Q, if each row matches with at least k rows in the universal table U along Q with probability greater than $(1 - \beta)$. The authors also found a relation between whether a set of columns forms a quasi-identifier and the number of distinct values assumed by the combination of the columns. $(1 - \beta, k)$-anonymity is obtained by solving 1-dimensional k-anonymity problems, avoiding the so-called 'curse of dimensionality' that refers to problems arising from sparsity when data is in high dimensional space, e.g. "the exponential number of combinations of dimensions [that] can be used to make precise inference attacks" [1]. $(1 - \beta, k)$-Anonymity protects against the oblivious adversary, but claims to be insufficient against the adversary targeting a specific individual.

[35] reflects on k-anonymity by introducing the M-score measure, or 'misuseability weight', representing the sensitivity level of the data of each table an individual is exposed to — and, by extension, the harm that misuse of that data can cause to an organization if leaked by employees, subcontractors and partners.

Malin and Sweeney proposed a formal model of a re-identification problem that pertains to genomic data [51]. This model builds on the ideas from k-anonymity. The authors provide algorithms of re-identification that can be applied to systems handling genomic data, as tests of privacy protection capabilities.

Narayanan and Shmatikov demonstrated new statistical de-anonymization attacks against the publicly released Netflix Prize data set containing de-identified movie ratings of about 500,000 subscribers of Netflix [58]. The authors showed that, given a little prior knowledge of a certain subscriber, it is possible to identify, with high certainty, records related to that subscriber in the anonymized data set. The authors show that their findings apply in general to multi-dimensional microdata.

2.6 Discussion

This Chapter presented a study of literature on the analysis of anonymity. Four directions of research were distinguished: information theory, process calculus, epistemic logic and k-anonymity. The analysis of anonymity may involve deterministic, non-deterministic and probabilistic aspects, depending on the context in which it is discussed and the purpose it is supposed to serve. For any system that involves human input, modeling anonymity would involve notions of angelic and demonic non-determinism.

Which of the directions we should choose, considering our problem at hand depends on whether anonymity only needs to be quantified or also specified/proven. The information-theoretic metrics provide a practical and relatively lightweight approach to measure the level of anonymity that anonymizing systems provide in different environments and under different constraints, but cannot be used to specify an anonymizing system or proof (predict) that it provides any anonymity property. Process algebra and logic can be used for the latter, but, to our knowledge, do not provide means to *quantify* anonymity. In the literature that was reviewed on process algebra and epistemic logic, aspects that either cannot be expressed, or are very difficult to express are typically left out in the abstraction that are then examined — even though some of those aspects might be relevant for accurately understanding anonymity.

Because our primary interest is data anonymity, and we seek quantification rather than formal proofs, we decide that k-anonymity is the most relevant model for us. In Chapter 3, we will describe a large-scale experiment to see how k behaves in two real policy research databases in the Netherlands, and proceed to propose new methods and techniques to make predictions about data anonymity. By doing that, we establish the case for doing quantitative research on identifiability, as set out in Chapter 1 — keeping the questionnaire example in mind, but seeking relevance to the processing of personal data in general.

3 An empirical study of quasi-identifiers

Throughout this thesis we will develop techniques to measure and predict anonymity. In this Chapter[1] we first perform an empirical analysis to examine how identifiability may work out in practice for a range of example quasi-identifiers selected either by observed presence in real systems, by expectancy of the likeliness of presence, or simply by our curiosity for quantifying how a certain combination of information would (not) be re-identifying.

3.1 Introduction

To examine how problems of re-identifiability may work out in practice, we decide to experimentally probe the re-identifiability of Dutch citizens for quasi-identifiers found in real-world data sets. We analyzed real registry office data of Dutch citizens, gathered from municipalities.

A seminal work on re-identification was done by Sweeney [76, 77]. Using 1990 U.S. Census summary data, she established that 87% of the US population was uniquely identifiable by a quasi-identifier (QID) composed of three demographic variables [75, 76]:

Definition 3.1 $QID_{example} = \{$ *Date-of-Birth + gender + 5-digit ZIP* $\}$

[1] This Chapter is based on M. Koot, G. van 't Noordende and C. de Laat, *A Study on the Re-Identifiability of Dutch citizens*, Electronic Proceedings of HotPETS 2010, July 2010 [45].

In Massachusetts (U.S.) the Group Insurance Commission administers health insurances to state employees. Sweeney legitimately obtained a de-identified data set containing medical information about Massachusetts' employees from them, including details about ethnicity, medical diagnoses and medication [76]. The data set contained the variables described in $QID_{example}$. Sweeney also legitimately obtained the identified 1997 voter registration list from the city of Cambridge, Massachusetts, which contained the same variables. By linking both data sets, it turned out to be possible to re-identify medical records, including records about the governor of Massachusetts at that time.

Recalling Section 2.5, Sweeney proposed k-anonymity, a test asserting that for each value of a quasi-identifier in a data set, at least k records must exist with that same value and be indistinguishable from each other. This introduces a minimal level of uncertainty in re-identification: assuming no additional information is available, each record may belong to any of at least k individuals.

In a paper revisiting Sweeney's work [32], Golle observes a difference between his results and Sweeney's results. Golle states he was unable to explain that difference due to a lack of available details about the data collection and analysis involved in Sweeney's work. In particular, in Golle's study of the 2000 U.S. Census data, only ∼63% of U.S. citizens turned out to be uniquely identifiable, as opposed to ∼87% that Sweeney determined by studying the 1990 U.S. Census data. It remains unclear whether the difference should be attributed to inaccuracies in the source data, intermediate changes in the ZIP code system, or something else.

In this Chapter, we analyze the identifiability of Dutch citizens by looking at demographic characteristics such as postal code and (partial) date of birth. By 'citizen' we refer to a person who is registered as an inhabitant of the Netherlands. We examine the re-identifiability only in the context of linking the data sets that are described, and not using any additional outside information. We limit ourselves to quasi-identifiers that we believe are most likely to be found in (identified) data sets elsewhere, based on commonly collected demographics. For two real-life data sets, the *National Medical Registration* (Dutch: "Landelijke Medische Registratie", or "LMR") and *Welfare Fraud Statistics* (Dutch: "Bijstands Fraude Statistiek", or "BFS"), we provide an assessment of two specific quasi-identifiers; many more quasi-identifiers exist in those data sets, involving e.g. ethnicity and marital status, but these are not discussed in this thesis. By using Dutch registry office data, we are confident that our results are likely to be very accurate, as we will argue in Section 3.2.3. That data is not collected via a census, but exists as a result of Dutch governmental administrative processes that citizens cannot opt out from. The registry offices are periodically subjected to audits that require very high data accuracy, which is tested via samples.

This Chapter is structured as follows: Section 3.2 describes our approach; Section 3.3 lists the results; and Section 3.4 discusses the results.

3.2 Background

In 2009, the Netherlands consisted of 12 provinces and 441 municipalities of varying size [14]. A municipality is an administrative region that typically spans several villages and cities. Municipal registry offices are the official record-keepers of persons residing in the Netherlands, and maintain identified data about them. De-identified data about individual citizens is available in a number of research databases. To illustrate our analysis we picked two, which we describe below. In Section 3.3 we assess, amongst others, re-identifiability of entries in these data sets.

3.2.1 Example data sets

The Dutch National Medical Registration (LMR) is a data collection program established in 1963, in which hospitals in the Netherlands participate by periodically sending in copies of medical and administrative information about hospital admissions and day care treatment. Example purposes of the LMR are the analysis of the effects of treatment, performance comparison between hospitals, and epidemiological studies. The LMR is currently managed by the Dutch Hospital Data foundation[2]. Statistics Netherlands, the Dutch organization for conducting statistical studies on behalf of the Dutch government, also receives annual copies of the full LMR data set for research purposes [15]. External researchers can currently request access to the records collected during 2005 and 2007 [11, 13]. These data sets contain only records about Dutch citizens; records about other patients are omitted. Each record in the LMR describes the hospital admission or day care treatment of a single individual, and multiple records may be present per individual. The 2005 and 2007 data sets each contain approximately 2.5 million records.

The Dutch Welfare Fraud Statistics (BFS) data set located at Statistics Netherlands contains records about investigations on suspected welfare fraud of Dutch citizens [12]. Each record in the data set maps to a single, completed investigation, and multiple records may be present per person. The information in the data set is provided by municipalities. Between 2002 and 2007, the average number of records (cases) per year was 38,161[3]. The BFS data set contains information at a different level of granularity than the LMR data set, which is the reason we selected it as a second example. For example, the LMR data set contains information about postal code, whereas the BFS data set does not.

Re-identified records from the BFS data set could be abused to embarrass or discriminate citizens that have been subject of fraud investigation. Similarly, re-identified records from the LMR data set could be abused to embarrass or

[2] `http://www.dutchhospitaldata.nl`

[3] Source: http://statline.cbs.nl

discriminate people based on medical history or medical conditions, potentially negatively impacting job or insurance prospects. Such consequences are at the disposal of the person possessing the (re-)identified records.

3.2.2 Approach and terminology

Recalling Section 1.2: a data set containing information about persons is said to be *de-identified* if 'direct' identifiers such as Social Security Numbers are omitted. A *quasi-identifier* is a variable or combination of variables which, although perhaps not intended or expected to identify individuals, can in practice be used for that purpose. A quasi-identifier may unambiguously identify a single individual, or reduce the number of possibilities to some small set of k individuals, the *anonymity set* [64]. A de-identified data set containing one or more quasi-identifiers can be *re-identified* by linking records to an *identified* data set containing the same quasi-identifying variable(s).

We assessed the (re-)identifiability of Dutch citizens by using quasi-identifiers composed of information about postal code, date of birth and gender information. We used registry office data of approximately 2.7 million persons, ~16% of the total population, obtained from 15 of 441 Dutch municipalities. The 15 municipalities and number of citizens are shown in Table 3.1. The sample contains small, mid-size and large municipalities. Although this selection is not random (we selected by number of citizens) or necessarily representative for the whole population, we considered the selection appropriate for our analysis, since it enables us to assess whether differences in re-identifiability are observable for small municipalities compared to large municipalities that contain a city, for example. The municipalities we selected are located in various parts of the country in such a way that there is no obvious bias due to geographical location of the municipalities in the countries — although the largest Dutch cities, Amsterdam, Rotterdam, and Den Haag, are located in the west of the Netherlands which is the most densely populated area of the Netherlands, known as "Randstad".

We requested a (nameless) listing of gender, full postal code and full date of birth of all citizens of 30 municipalities, and eventually obtained records of 15 municipalities, totalling approximately 2.7 million citizens. The remainder of this Chapter is based on analysis of this data. We distinctly discuss data only at municipal level; i.e. 'Amsterdam' refers to the municipality of Amsterdam rather than the city of Amsterdam.

We primarily focus on quasi-identifiers that match the LMR and BFS examples. The results, however, apply to *any* data set that contains these quasi-identifiers. We did not attempt to obtain access to data from the actual data sets, because for our purposes it suffices to know which possible quasi-identifying variables they contain, and the latter can be learned from public documents [11, 12, 13].

Table 3.1: Municipalities included in our study (ordered by number of citizens)

Municipality	# of citizens
Amsterdam	766,656
Rotterdam	591,046
Den Haag	487,582
Utrecht	305,845
Nijmegen	161,882
Enschede	156,761
Arnhem	147,091
Overbetuwe	45,548
Geldermalsen	26,097
Diemen	24,679
Reimerswaal	21,457
Enkhuizen	18,158
Simpelveld	11,019
Millingen a/d Rijn	5,915
Terschelling	4,751
TOTAL:	2,774,476

3.2.3 Data quality

Transactions between the Dutch government and Dutch citizens rely upon municipal registry offices as source of data about citizens — including the transaction of passport issuance. Registry office data is not free of error: data may be inconsistent with reality due to e.g. failure of citizens to report changes timely and truthfully, typographical errors and software errors [60]. The registry offices are required to undergo a periodical audit, which includes an integrity check of a random sample of the electronic person records. Each record from that sample is matched against other official files associated with the person whom the record is about, such as birth certificates. Each variable containing an incorrect value is counted as a single error, and the maximum allowed rate for errors in 'essential' fields like DoB and postal code is 1% of the sample set size: to pass the test, a 100-record sample cannot contain more than 1 error in essential fields. The sample size depends on the municipality size. During the 2002-2005 audit cycle, 339 of the 370 (92%) audited municipalities passed this test [60]. This suggests that Dutch registry offices are generally a reliable source of data. During our own data sanity checks we removed 11 records containing a postal code from outside the sampled municipalities, as those records would have caused false outliers[4]; the remainder passed all sanity checks.

[4] These cases may be related to moving citizens, e.g. pending handover of data between municipalities.

3.2.4 Postal codes in the Netherlands

In the Netherlands, a postal code consists of a four-digit number and a two-character extension — e.g. “1098 XG”, the postal code of our institution. The four-digit number is referred to as ‘4-Position PostalCode’ (*PC4*), and is located in exactly one town (city, village). A town may be divided into multiple PC4-regions: for example, our data contains eighty different PC4-regions for the city of Amsterdam, “1098” being one of them.

The two-character extension indicates a street, but often also a specific odd or even range of house numbers *within* that street. The full postal code is referred to as ‘6-Position PostalCode’ (PC6). A combination of full (*PC6*) postal code and house or P.O. box number uniquely indicates a postal delivery address in the Netherlands.

3.3 Results

This Section describes the results of our analysis. Section 3.3.1 describes an overall analysis of our input data. From the result data it becomes clear what combinations of variables can be used to single out individuals or small groups of citizens, and which combinations pose less of a privacy risk in that sense. Section 3.3.2 describes the potential re-identifiability of citizens in the LMR data set. Section 3.3.3 analyses the potential re-identifiability of citizens in the BFS data set. We use the following notations: *QID=Quasi-IDentifier*, *DoB=Date of Birth*, *YoB=Year of Birth* and *MoB=Month of Birth*.

By ‘quasi-identifier’ we refer to abstract variables, by ‘quasi-identifier value’ to a valuation of those variables. We use rounded values for the sake of readability. For each quasi-identifier, we counted the number of different (distinct) values in the data — this is the number of anonymity sets; the number of people sharing a specific quasi-identifier value represents the anonymity set size.

In addition to mean values, we provide quartiles and min-max values to give an indication of how a quasi-identifier maps citizens in anonymity sets of rather diverse or rather similar size[5]. We chose quartiles as a means to indicate the value distribution while maintaining some brevity and readability of tables. Another choice could have been made (e.g., for deciles or percentiles), however, none has a definite advantage over the other. By using quartiles we can state

[5] The lower (1st) quartile is the value separating the lower 25% of the values; the median value (2nd quartile) separates the higher half of the values from the lower half; the upper (3rd) quartile separates the higher 25% of the values. To illustrate: given a population of 500 persons, both (k=100,k=100,k=100,k=100,k=100) and (k=1,k=1,k=1,k=1,k=496) are possible outcomes that have a mean value of $k = 100$, while both sets are obviously very different. For the former set, all three quartiles are 100, as are both the minimum and maximum: *all* anonymity sets have size $k = 100$. For the latter set of numbers, minimum value and all quartiles are 1, but the maximum value is 496: this shows that the distribution is skewed. In our context, the latter means that a quasi-identifier maps citizens into anonymity sets of different sizes.

Table 3.2: Anonymity set size k for various (potential) quasi-identifiers

Quasi-identifier:	# of sets	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PC4	388	2	3,278	7,090	7,188	10,300	22,330
PC6	66,883	1	24	35	41	50	1,322
PC4+DoB	2,267,700	1	1	1	1	1	42
PC6+DoB	2,759,422	1	1	1	1	1	5
PC4+gender	776	1	1,652	3,536	3,594	5,151	11,730
PC6+gender	133,012	1	11	18	21	25	954
gender+YoB	221	1	5,219	14,570	12,550	19,740	25,580
gender+YoB+MoB	2,699	1	397	1,177	1,028	1,594	2,326
gender+YoB+MoB+PC4[a]	635,679	1	2	3	4	6	40
gender+YoB+MoB+municip.[b]	34,790	1	6	18	80	96	733
gender+DoB	71,318	1	21	40	39	54	571
gender+DoB+PC4	2,488,828	1	1	1	1	1	22
gender+DoB+PC6	2,766,475	1	1	1	1	1	4
town+gender	134	1	222	1116	20,700	3259	347,100
town+YoB	5,642	1	6	29	492	101	14,270
town+YoB+MoB	49,207	1	2	5	56	20	1,262
town+DoB	463,134	1	1	2	6	7	419
town+YoB+gender	10,492	1	4	17	264	60	7,515
town+YoB+MoB+gender	83,172	1	1	3	33	14	695
town+DoB+gender	697,875	1	1	2	4	5	226

[a] QID_A, see Section 3.3.2.
[b] QID_B, see Section 3.3.3.

properties of the distribution of anonymity set sizes such as "at most 25% of the anonymity sets are smaller than <1st quartile>" and "at most 50% of the anonymity sets are smaller than <median>".

3.3.1 Analysis over aggregated data

This Section describes the results of an analysis of the combined data of the citizens of all municipalities listed in Table 3.1. By including both small and large municipalities, covering the smallest villages (the smallest having two inhabitants) and largest cities (the largest having 684,926 inhabitants) in the Netherlands, the minimum and maximum anonymity set sizes represent the worst and best cases we expect to be found *anywhere* in the Netherlands. Furthermore, the statistics over the combined data indicate how strongly identifiable a quasi-identifier is for the overall population.

Throughout the next subsections, k denotes the anonymity set size; $k = 1$ means that some quasi-identifier value unambiguously identifies some individual, $k = 2$ means that the value is shared by two individuals, and so on. Table 3.2 shows the statistical characteristics of anonymity set size k for various (potential) quasi-identifiers. The column *'# of sets'* contains the number of different values present in our data for a given quasi-identifier, i.e., the number of anonymity sets. Generally, the higher this number, the weaker the level of privacy, because the anonymity sets will tend to be smaller. The min/max values denote the size of the smallest and largest anonymity set.

As an example, the median anonymity set size of PC6 is 35, the minimum

Table 3.3: Number of Dutch citizens per anonymity set size, for various quasi-identifiers

Quasi-identifier:	$k = 1$	$k \leq 5$	$k \leq 10$	$k \leq 50$	$k \leq 100$
PC4	0	9	19	345	996
PC6	429	6,109	25,103	1,459,939	2,354,255
PC4+DoB	1,861,081	2,754,465	2,765,932	2,774,476	-
PC6+DoB	2,744,653	2,774,476	-	-	-
PC4+gender	4	27	103	889	2,555
PC6+gender	1,854	31,262	184,803	2,342,242	2,629,017
gender+YoB	5	14	53	250	516
gender+YoB+MoB	55	356	712	4,478	9,674
gender+YoB+MoB+PC4[a]	137,035	279,100	2,196,950	2,774,476	-
gender+YoB+MoB+municip.[b]	2,186	22,565	59,597	244,152	619,671
gender+DoB	2,014	14,506	40,322	1,392,622	2,725,472
gender+DoB+PC4	2,240,461	2,765,067	2,772,205	2,774,476	-
gender+DoB+PC6	2,758,578	2,774,476	-	-	-
town+gender	4	4	28	372	896
town+YoB	499	3,172	7,225	50,985	103,145
town+YoB+MoB	10,083	61,073	112,850	287,173	394,844
town+DoB	185,042	596,769	1,045,559	2,730,668	2,750,700
town+YoB+gender	1,153	7,195	16,333	102,018	150,135
town+YoB+MoB+gender	22,260	109,126	170,351	398,601	826,744
town+DoB+gender	288,409	1,029,601	1,813,559	2,750,669	2,764,050

[a] QID_A, see Section 3.3.2.

[b] QID_B, see Section 3.3.3.

size is 1 and the maximum size is 1,322. This means that at most half of the values for PC6 have anonymity sets of sizes between 1 and 35, and that the sizes of the anonymity sets in the upper half are between 35 and 1,322.

From the quartiles it becomes clear that some quasi-identifiers are particularly strong, by which we mean that a large portion of the anonymity sets established by that quasi-identifier are of small size (e.g. $k = 1$ or $k \leq 5$). For example, for $\{PC4 + DoB\}$, Table 3.2 shows an anonymity set size of $k = 1$ for up to the 3rd quartile, meaning that 75% of the quasi-identifier values unambiguously identify a citizen. Looking at the lower quartiles, it also becomes clear that some quasi-identifiers are weaker identifiers: for $\{PC4\}$, only at most 25% of the sets are of size $k \leq 3,278$; for $\{gender + YoB\}$, at most 25% of the sets are of size $k \leq 5,219$. Overall, it turns out that quasi-identifiers containing both PC4 or PC6, as well as date of birth, are most identifying.

We were surprised to find that PC4 postal codes exist which are shared by only two citizens: we had expected that PC4 codes always map to relatively large numbers of citizens. Upon closer inspection, it appears that the data is accurate: it represents the inhabitants of a new construction area in the harbor of Rotterdam. These pioneering citizens turn out to be unambiguously identifiable nation-wide by only their $\{PC4 + gender\}$ or $\{town + gender\}$ — albeit only until other citizens officially move in.

Table 3.2 also clearly shows that adding the two-character extension to the PC4 postal code strongly increases identifiability: the median anonymity set size for $\{PC4\}$ is 7,090, for $\{PC6\}$ only 35.

Whereas Table 3.2 focusses on the size distribution of the anonymity sets, Table 3.3 shows the actual number of *citizens* found in those anonymity sets. The larger the value in columns '$k = 1$', '$k \leq 5$' and possibly '$k \leq 10$', the larger the portion of the population that is covered by anonymity sets of those (small) sizes and the stronger the quasi-identifier identifies citizens. The numbers confirm that $\{PC6 + DoB\}$ is a strong identifier, because here nearly all citizens have $k = 1$; $\{PC6\}$ alone is not a strong identifier, because only a very small portion of the citizens have $k \leq 10$ (compared to $k \leq 50$). We also included columns for a few larger set sizes ($k \leq 50$ and $k \leq 100$) for illustrative purposes. For example, only 896 out of 2.7 million citizens are identifiable to a group of ≤100 by $\{town + gender\}$, so by themselves, those variables do not pose a significant privacy risk for most citizens. For readability, we replaced numbers by '-' when the total population is reached at some k.

From the numbers for quasi-identifier $\{gender + DoB + PC6\}$ it follows that approximately 99.4% of the Dutch citizens in our data set (2,758,578 out of 2,774,476) can be unambiguously identified by $\{gender + DoB + PC6\}$; and lastly, it turns out that 67.0% (1,861,081 out of 2,774,476) can still be unambiguously identified by $\{PC4 + DoB\}$.

3.3.2 Case: National Medical Registration

The LMR contains a large amount of information about hospital admissions and day care treatment: amongst others, it contains fields describing the hospital, the patient's insurance type, diagnosis codes, the treatment that was provided and the medical specialisms and disciplines involved [11, 13]. This information could be privacy-sensitive and it is generally treated as such, even when de-identified: i.e., access to the LMR and BFS data set is only granted to qualified applicants, for specific purposes, under specific conditions of confidentiality — Statistics Netherlands is very aware of privacy risk [88]. The LMR data set also contains demographic data about the patient. In particular, the LMR contains the following quasi-identifier:

Definition 3.2 $QID_A = \{\ PC4 + gender + YoB + MoB\ \}$

Our data contains 635,679 different anonymity sets for QID_A. We use k_A to denote the anonymity set sizes for this quasi-identifier. 137,035 people, ∼4.8%, are unambiguously identifiable by QID_A, that is, they are the only person in the anonymity set, which thus has k_A=1. Furthermore, we found 212,536 citizens to have $k_A = 2$; $260,244$ to have $k_A = 3$ and 282,644 to have $k_A = 4$ (most common size). Table 3.4 lists the statistical properties of the size of the anonymity sets established by this quasi-identifier. The municipality size is included for quick reference.

The numbers show that there is no large difference in anonymity between citizens of different-sized municipalities: the range of the medians is 1–5. The highest median anonymity set size is found in Amsterdam, the lowest is found in Terschelling. The latter means that half of the QID_A values found in Terschelling unambiguously identify a citizen.

The municipality size (*'# of citizens'*) and median anonymity set size (column *'Median'*) have a Pearson correlation coefficient of .60. The single largest anonymity set is found in Amsterdam and is of size 40. Based on the numbers shown in Table 3.3, the total percentage of citizens identifiable to a group of 10 or less by this quasi-identifier is ∼79.1% (2,196,950 out of 2,774,476).

Figure 3.1 visualizes the numbers in Table 3.4. Some large anonymity sets exist as outliers, especially for larger municipalities, but overall anonymity is approximately the same for all municipalities.

Note that there is a difference in constraints between registry office data and the hospital admission data set: whereas the year of birth is allowed to be zero by the Dutch registry offices — e.g. for immigrants about whom the date of birth is not fully known —, the LMR requires it to be non-zero and be estimated if unknown [79]. This means that LMR-records about a person who is officially registered with zero year of birth (in our data set we only found 3) will *not* be re-identified by quasi-identifiers involving the year of birth. On the other hand, the quality of data from the LMR and BFS depends on their sources (hospitals and municipalities); it is not asserted whether each record accurately represents reality [11, 12, 13] – note that any mismatch (error) prevents linkability, and thus improves privacy for the involved individual.

Table 3.4: Statistical summary of k_A, divided by municipality (ordered by median)

Municipality:	# of citizens	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Amsterdam	766,656	1	2	5	6	8	40
Rotterdam	591,046	1	2	4	5	6	33
Enkhuizen	18,158	1	2	4	4	6	20
Diemen	24,679	1	2	4	4	6	19
Den Haag	487,582	1	2	3	4	6	30
Utrecht	305,845	1	2	3	4	6	36
Enschede	156,761	1	2	3	4	5	31
Nijmegen	161,882	1	2	3	4	5	35
Arnhem	147,091	1	1	3	3	4	25
Millingen a/d Rijn	5,915	1	2	3	3	4	12
Simpelveld	11,019	1	1	3	3	4	12
Geldermalsen	26,097	1	1	2	2	3	16
Overbetuwe	45,548	1	1	2	3	4	18
Reimerswaal	21,457	1	1	2	2	3	11
Terschelling	4,751	1	1	1	1	2	10
OVERALL	2,774,476	1	2	3	4	6	40

3.3.3 Case: Welfare Fraud Statistics

In the BFS data set, we recognised the following as a potential quasi-identifier:

Definition 3.3 $QID_B = \{$ *municipality + gender + YoB + MoB* $\}$

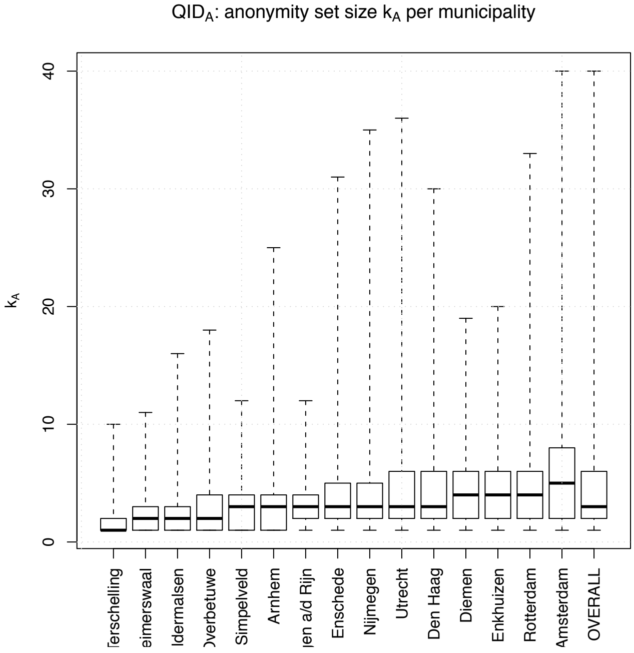


Figure 3.1: Box-and-whisker plot showing anonymity set sizes k_A, per municipality. Whiskers denote the minimum and maximum values; the boxes are defined by lower and upper quartiles and the median value is shown.

Our data contains 34,790 different anonymity sets for QID_B. 2,186 people, ~0.07%, are unambiguously identifiable by QID_B. Furthermore, we found 3,552 citizens to have $k_B = 2$; 5,064 to have $k_B = 3$ and 5,508 to have $k_B = 4$. The total percentage of citizens identifiable to a group of 10 or less is ~2.14% (59,597 out of 2,774,476). The single largest anonymity set is found in Amsterdam and is of size 733.

Table 3.5 lists the statistical properties of k_B per municipality. The numbers show that regarding the BFS, large differences in anonymity exist between citizens of different-sized municipalities: the range is 1–733. The highest median anonymity set size is 310, found in Amsterdam, the lowest is 2, found in Terschelling. Municipality size and median anonymity set size have a Pearson correlation coefficient of .99; the median anonymity set size is rather constant at ~0.04% (1/2,500) of the population size.

Figure 3.2 visually represents the numbers in Table 3.5. Note that the range on the vertical axis is much larger than in figure 3.1. It is clear that citizens from large municipalities tend to have much stronger anonymity than citizens from small municipalities, which is something to remember when dealing with de-identified data about citizens from small municipalities.

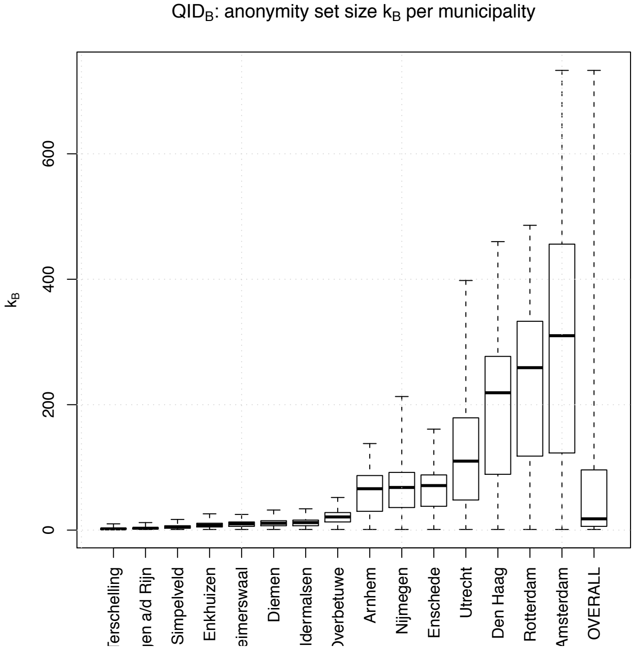


Figure 3.2: Box-and-whisker plot showing anonymity set sizes k_B, per municipality. Whiskers denote min-max values.

Table 3.5: Statistical summary of k_B, divided by municipality (ordered by median)

Municipality:	# of citizens	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Amsterdam	766,656	1	123	310	296	456	733
Rotterdam	591,046	1	118	259	228	333	486
Den Haag	487,582	1	89	219	188	277	460
Utrecht	305,845	1	48	110	121	179	398
Enschede	156,761	1	38	71	64	88	161
Nijmegen	161,882	1	36	68	66	92	213
Arnhem	147,091	1	30	66	60	87	138
Overbetuwe	45,548	1	13	21	20	28	52
Geldermalsen	26,097	1	7	12	12	16	34
Diemen	24,679	1	7	11	11	15	32
Reimerswaal	21,457	1	6	10	10	13	25
Enkhuizen	18,158	1	5	8	8	11	26
Simpelveld	11,019	1	3	5	5	7	17
Millingen a/d Rijn	5,915	1	2	3	3	4	12
Terschelling	4,751	1	1	2	3	3	10
OVERALL	2,774,476	1	6	18	80	96	733

3.4 Discussion

This Chapter established the identifiability of Dutch citizens using information about postal code, date of birth and gender. We studied real registry office data of approximately 2.7 million citizens, ∼16% of the total population, obtained from 15 of 441 Dutch municipalities of varying size. We assessed the re-identifiability of records about these individuals in known data sets about hospital admissions and welfare fraud.

It turns out that approximately 99.4% of the sampled population is unambiguously identifiable using PC6 postal code, gender and date of birth, and 67.0% by PC4 and date of birth alone. Regarding the quasi-identifier found in the LMR data set, approximately 4.8% of the sampled population is unambiguously identifiable and 79.1% is identifiable to a group of 10 or less. Regarding the quasi-identifier found in the BFS data set, approximately 0.07% of the sampled population is unambiguously identifiable and 2.14% is identifiable to a group of 10 or less; for small municipalities, however, the anonymity set sizes become much smaller and re-identifiability higher.

As far as we know, we are the first to study re-identifiability using authoritative registry office data. Comparing to Sweeney [75, 76] and Golle [32], who's studies relied on census data, our study relies on data from the data source that is authoritative during Dutch passport issuance, which is not prone to the intricacies of survey-based data collection. We only cover a portion of the Dutch citizens, ∼16%, but are confident that the results for that portion are accurate. For the quasi-identifiers we chose to analyze, we also provide the minimum and maximum anonymity set sizes that can be expected to be found *anywhere* in the Netherlands.

The results suggest that, considering the quasi-identifier in the National Medical Registration data set, someone who is able to access registry office data can re-identify a large portion of records with relatively high certainty. Considering the quasi-identifier in the Welfare Fraud Statistics data set, the re-identification risk is generally lower, but strongly depends on municipality size.

One could argue about the plausibility of the threat scenario underlying the two cases we picked: we assume an adversary who is able to access non-public records from both registry offices and Statistics Netherlands. Access to the data sets at Statistics Netherlands, including the LMR and BFS data sets, is only granted to qualified applicants, for specific purposes, under specific conditions of confidentiality [88]. Thus, obtaining data may require an investment that is disproportional to the expected gain of re-identifying records from these particular data sets to begin with. We note, however, that our results apply to *any* de-identified data set containing the assessed quasi-identifiers. For a data set that does not contain other quasi-identifiers than those discussed in this Chapter, our results provide an upper and lower bound of anonymity. Also, registry offices are not the only source for identified data, and *any* identified database containing these quasi-identifiers with sufficiently large coverage of the total population may be used; suitable data sets may also exist at, e.g., information brokers, marketing agencies and public transport companies. Besides, preventing registry office data itself from being used for re-identification may be difficult: the 441 municipalities are autonomous gatekeepers to their citizen's data and that citizen data is already necessarily exchanged on a regular basis for a variety of legitimate purposes [63]. It is difficult to protect data that has many legitimate users and uses.

These results are, by themselves, useful as input for privacy impact assessments involving data about Dutch citizens. It remains a matter of policy what value of k can be considered *sufficiently strong* anonymity for particular personal information. Conceivably this is be estimated via regular risk calculations, i.e., chances multiplied by impact, assuming that *impact* takes into consideration aspects such as 'misusability' of the information, emotional harm, social harm and other harm that may result from its disclosure.

4 Efficient probabilistic estimation of quasi-identifier uniqueness

4.1 Introduction

In Chapter 3 we analyzed quasi-identifiers in two data sets containing information about hospital intakes and welfare fraud. The quasi-identifier in the hospital intake data set consisted of 4-digit postal code, gender, month of birth and year of birth, and in the welfare fraud data set it contained the municipality rather than the 4-digit postal code. The objective of the study was to assess the level of anonymity enjoyed by persons present in the data sets. The results were roughly comparable to the results obtained by Sweeney in the U.S. For example, 67.0% of the sampled population turned out identifiable by date of birth and four-digit postal code alone, and 99.4% by date of birth, full postal code and gender.

One of the common challenges in k-anonymity and its developments is the recognition of quasi-identifiers (QIDs). The method we develop in this Chapter[1] provides a new way of efficiently estimating the likelihood that a given set of attributes will function as a *perfect quasi-identifier*, i.e., that each value of a quasi-identifiers unambiguously identifies an individual. That quantification may be useful as a worst-case metric in privacy impact assessments and policy

[1] This Chapter is based on M. Koot, M. Mandjes, G. van 't Noordende and C. de Laat, *Efficient probabilistic estimation of quasi-identifier uniqueness*, Proceedings of NWO ICT.Open 2011, November 2011 [43].

research.

Usually, QIDs are addressed *after* data has been collected, and each data collection deals with QIDs for itself. In our scenario, a data collector (perhaps Statistics Netherlands) collects data and publishes a single number representing the heterogeneity of the QID distribution over the records in his table. That number, the *Kullback-Leibler distance* that will be introduced shortly, represents the distribution skew in the prior data collections. Using that number, our method allows future data collectors to predict properties of QIDs *before* collecting data; and possibly use that information to decide on what (not) to collect and possibly to decide what the impact of combining earlier-collected data may have on privacy.

For QIDs consisting of personal attributes that do not change, such as date of birth, or that rarely change, such as postal code, the method introduced in this Chapter provides an efficient approximation of the probability that every (QID) value in a group of people unambiguously identifies an individual. An entity such as Statistics Netherlands, which has access to enormous amounts of data, might publish precomputed tables that data collectors can use to decide what data (not) to collect. Chapter 7 will elaborate on this.

As a follow up to Chapter 3, the primary question this Chapter addresses is: *'Can we develop a methodology to determine the probability that all persons in a group can be uniquely identified by quasi-identifier X?* This can then be used as a measure of anonymity. The main contribution of our work is that we provide a sound technique to accurately approximate this probability. We translate our question in terms of a birthday problem, and then rely on probabilistic techniques.

The main problem is that, unlike in the classical birthday problem [57], the probability distribution for many variables and thus for many QIDs is non-uniform, i.e., not all possible values occur with equal frequency. This heterogeneity is dealt with by adjusting the outcome of the homogeneous birthday problem (in which all outcomes *are* equally likely) by a measure of heterogeneity, the *Kullback-Leibler distance* [47]. As mentioned, the techniques used are of a probabilistic nature; more specifically, we borrow elements from *large-deviations theory* [23, 52].

It is emphasized that the stated question is of interest both to adversary ('which quasi-identifiers should I want?') and the anonymous subject ('which quasi-identifiers should I avoid?'). Our method will be demonstrated using demographic data from the Netherlands, but the approach can be applied to *any* population.

The remainder of this Chapter is organized as follows. In Section 4.2 we formally describe the problem in terms of a birthday problem with unequal probabilities. Section 4.3 presents an approximation for the uniqueness probability under heterogeneity, where the deviation from the uniform situation is captured by the Kullback-Leibler distance. In Section 4.4 we validate the approximation, and use the approximation to run a number of experiments. The Chapter is concluded in Section 4.5, by a discussion and outlook.

4.2 Problem

The problems we come across in this Chapter can be regarded as *generalized birthday problems.* In the 'classical' birthday problem [28, 83] there are k individuals, each of whom is assigned (uniformly, independently) a value from the set $\{1, \ldots, N\}$. It is a straightforward exercise in probability theory to check that the probability that all values ('birthdays') are unique is given by

$$\pi_{\mathrm{u}}(k, N) = \frac{N}{N}\frac{N-1}{N}\cdots\frac{N-k+1}{N} = \frac{N!}{(N-k)!N^k}.$$

However, things complicate in case the outcomes $\{1, \ldots, N\}$ are *not* equally likely. To study this situation, suppose that F_i outcomes have probability α_i/N, for $i = 1, \ldots, d$ (that is, there are d groups within which the probabilities are uniform again). Here it is assumed that $F_1 + \ldots F_d = N$ (each outcome is a member of one group) and $F_1\alpha_1 + \ldots F_d\alpha_d = N$ (the total probability is 1). For this generalized birthday problem, it is not possible to write down a clean expression for the uniqueness probability (although it can be evaluated numerically in quite an efficient way [41]). However, as we will show in this Chapter, we succeeded in developing an accurate approximation. This approximation is based on the Kullback-Leibler distance, which is a measure for heterogeneity within the population. It turns out that the more heterogeneous the population is, the lower the uniqueness probability. In addition, it is shown that assuming all outcomes are equally likely (so that the above explicit formula can be applied) leads to quite substantial estimation errors.

To simplify the exposition, we use a very simple quasi-identifier in our examples: age. We experimentally assessed the quality of our approximation using real data about the Dutch population: the distribution of age in all Dutch municipalities, which vary in size (1k–750k citizens). Different from our study in Chapter 3, the data we use here is publicly available from Statistics Netherlands, so as to remove a threshold for those desiring to reproduce our results[2].

4.3 Methodology: birthday problems

As mentioned above, the uniqueness probability can be calculated straightforward in case all outcomes are equally likely. In this Section we present an approximation for the situation where this is *not* the case, that is, the situation in which probabilities of the outcomes $1, \ldots, N$ differ from $1/N$.

[2] Statistics Netherlands, StatLine: http://statline.cbs.nl

4.3.1 Approximations for general birthday problems

In this subsection we describe a way to find an approximation for the uniqueness probability in the non-uniform scenario. The approximation relies heavily on the idea of 'Poissonization'.

Approximations for the uniform case. We briefly describe a classical approximation for the uniform case (i.e., $d = 1$), and show that this approximation is exact in a particular asymptotic regime. To this end, observe that

$$\begin{aligned}\pi_{\mathrm{u}}(k, N) &= \exp\left(\sum_{i=0}^{k-1} \log\left(1 - \frac{i}{N}\right)\right) \\ &\approx \exp\left(-\frac{1}{N}\sum_{i=0}^{k-1} i\right) \approx \exp\left(-\frac{k^2}{2N}\right). \qquad (4.1)\end{aligned}$$

This approximation can be formally justified if k scales like $\sqrt{N}$: applying 'Stirling',

$$\begin{aligned}\pi_{\mathrm{u}}(a\sqrt{N}, N) &= \frac{N!}{(N-k)!N^k} \\ &\sim e^{-a\sqrt{N}}\left(1 - \frac{a}{\sqrt{N}}\right)^{N-a\sqrt{N}} \to e^{-\frac{a^2}{2}}, \qquad (4.2)\end{aligned}$$

where the convergence is due to Lemma 4.1, included at the end of this Chapter. Plugging in $a := k/\sqrt{N}$ indeed yields approximation (4.1).

Poissonization for the uniform case. We show that assuming that k is not given but drawn from a Poisson distribution with mean k yields, remarkably enough, the same asymptotic (4.2). To this end, suppose that the sample size is Poisson distributed with mean k. An elementary conditioning argument yields that this gives the uniqueness probability

$$\pi_{Pois,\,u}(k, N) = \sum_{i=0}^{N} e^{-k}\frac{k^i}{i!}\frac{N!}{(N-i)!N^i} = e^{-k}\left(1 + \frac{k}{N}\right)^N.$$

As before, an approximation of the type $\exp(-k^2/(2N))$ can be justified, because

$$\pi_{Pois,\,u}(a\sqrt{N}, N) = e^{-a\sqrt{N}}\left(1 + \frac{a}{\sqrt{N}}\right)^N \to e^{-\frac{a^2}{2}},$$

applying Lemma 4.1.(ii). In other words, even though we randomize the number of samples, we obtain the same approximation.

The non-uniform case. We now consider the situation where F_i (for $i = 1, \ldots, d$) of the outcomes have probability α_i/N, with $F_1 + \ldots F_d = N$ and

$F_1\alpha_1 + \dots F_d\alpha_d = N$. As argued earlier, if the α_i are not uniform, then computing the uniqueness probability $\pi(k, N)$ is not straightforward. The idea of Poissonization does ease this task considerably, though, as we will show.

It is first observed that when sampling k times according to the mechanism described above, the number of these samples that are from group i (with $i = 1, \dots, d$) has a multinomial distribution with parameters k and (probability vector) $(\alpha_1 F_1/N, \dots, \alpha_d F_d/N)'$. Suppose instead the number of samples from group i is Poisson distributed with mean $(\alpha_i F_i/N)\cdot k$ (rather than the described multinomial distribution). Then the uniqueness probability essentially reduces to the product of the uniqueness probabilities *within each of the* d *groups* (use independence!). Therefore, in self-evident notation,

$$\begin{aligned}\pi_{\mathrm{P}ois}(k, N) &= \prod_{i=1}^{d} \pi_{\mathrm{P}ois,\,u}\left(\alpha_i F_i \cdot \frac{k}{N}, F_i\right) \\ &\approx \exp\left(-\frac{k^2}{2N^2}\sum_{i=1}^{d}\alpha_i^2 F_i\right), \end{aligned} \tag{4.3}$$

and then the idea is to approximate $\pi(k, N)$ by $\pi_{\mathrm{P}ois}(k, N)$, as we did in the uniform case. In [9, Thm. 4] this approximation was made precise, in the sense that, with $f_i := F_i/N$ being the fraction of all individuals that is of type i, as $N \to \infty$,

$$\pi(a\sqrt{N}, N) \to \exp\left(-\frac{a^2}{2}\sum_{i=1}^{d}\alpha_i^2 f_i\right).$$

4.3.2 Impact of non-uniformity

A perhaps naive idea could be to ignore the heterogeneity and to simply use the 'homogeneous formula' (4.1). In this subsection we show that such an approach could lead to highly inaccurate estimates — evidently, the more heterogeneous the population is, the less accurate such an approximation. To study this effect, we further asses the impact that non-uniformity has on the uniqueness probability.

Uniform distribution maximizes uniqueness probability. The approximation of the uniqueness probability for the non-uniform case is majorized by the approximation for the uniform case. This can be explained as follows. First observe that we need to prove that $\sum_{i=1}^{d}\alpha_i^2 f_i \geq 1$, given that $\sum_{i=1}^{d} f_i = \sum_{i=1}^{d}\alpha_i f_i = 1$ (where it is noted that the minimum value 1 is attained when all α_i coincide). Let the random variable A have the value α_i with probability

f_i. As variances are non-negative, we evidently have

$$\sum_{i=1}^{d} \alpha_i^2 f_i = \mathbb{E}A^2 \geq (\mathbb{E}A)^2 = 1,$$

which proves our claim. The fact that the uniform distribution actually *maximizes* the uniqueness probability has been observed before, cf. [40, 69]. More specifically, it means that all perturbations from the uniform distribution *reduce the uniqueness probability.*

Distances between distributions. Observing that

$$\frac{\exp(-\frac{a^2}{2})}{\exp(-\frac{a^2}{2}\sum_{i=1}^{d}\alpha_i^2 f_i)} = \exp\left(\frac{a^2}{2}\sum_{i=1}^{d}(\alpha_i^2 f_i - 1)\right),$$

we conclude that

$$\frac{1}{2}\sum_{i=1}^{d}(\alpha_i^2 f_i - 1)$$

is a measure for discrepancy between the uniform distribution and the non-uniform distribution under consideration. There are several distance measures between distributions, the most prominent perhaps being the Kullback-Leibler distance [47]. Below we argue that, at least for small perturbations, our discrepancy metric essentially reduces to the Kullback-Leibler distance.

Indeed, if α_i is not too different from 1, the Kullback-Leibler distance with respect to the uniform distribution, say κ, can be evaluated as follows. First observe that

$$\kappa = \sum_{i=1}^{d}\left(Nf_i\frac{\alpha_i}{N}\right)\log\left(\frac{Nf_i\dfrac{\alpha_i}{N}}{Nf_i\dfrac{1}{N}}\right) = \sum_{i=1}^{d}\alpha_i f_i \log\alpha_i.$$

Now let α_i equal $1+\beta_i\varepsilon$ for ε small; $\sum_{i=1}^{d}\alpha_i f_i = 1$ then entails that $\sum_{i=1}^{d}\beta_i f_i = 0$. Using the Taylor expansion $\log(1+x) = x - x^2/2 + O(x^3)$, it follows that

$$\begin{aligned}
\kappa &= \sum_{i=1}^{d}(1+\beta_i\varepsilon)f_i\log(1+\beta_i\varepsilon)\\
&= \sum_{i=1}^{d}(1+\beta_i\varepsilon)f_i\left(\beta_i\varepsilon - \frac{1}{2}\beta_i^2\varepsilon^2\right) + O(\varepsilon^3)\\
&= \frac{1}{2}\sum_{i=1}^{d} f_i\beta_i^2\varepsilon^2 + O(\varepsilon^3).
\end{aligned}$$

Now replacing $\beta_i\varepsilon$ by $\alpha_i - 1$, and using $\sum_{i=1}^d \alpha_i f_i = 1$, we arrive at the approximation, for ε small:

$$\kappa \approx \frac{1}{2}\sum_{i=1}^{d}(\alpha_i^2 f_i - 1).$$

In other words,

$$\frac{\pi_{\mathrm{u}}(k,N)}{\pi(k,N)} \approx \frac{\exp\left(-k^2/2N\right)}{\exp\left(-k^2/2N \cdot \sum_{i=1}^d \alpha_i^2 f_i\right)} \approx \exp\left(\frac{k^2}{N}\cdot\kappa\right).$$

As a consequence, we obtain the following elegant approximation for the uniqueness probability in the heterogeneous case:

$$\pi(k,N) \approx \pi_{\mathrm{u}}(k,N)\cdot e^{-k^2/N\cdot\kappa} \approx e^{-(\frac{1}{2}+\kappa)k^2/N}.$$

In other words, to approximate the uniqueness probability for the non-uniform case, we have to take the uniqueness probability for the uniform case, and raise it to the power κ. This κ, the Kullback-Leibler distance, measures the discrepancy of the distribution relative to the uniform distribution. More specifically, the larger κ, the more heterogeneous the distribution is, the smaller the uniqueness probability. It is noticed that the approximation formula is consistent with the one for the uniform case; then $\kappa = 0$.

4.4 Experiments with demographic data

In this Section we run two sets of experiments: (i) experiments in which we validate our approximation formula, as was deduced in the previous Section; (ii) experiments in which we assess the impact of heterogeneity, where all computations are based on our approximation formula.

4.4.1 Validation of the approximation formula

In our validation experiment we have considered the following setup, focusing on the level of anonymity one has after revealing her or his age. Supposing that a group of k individuals is considered, our objective is to determine the probability that each of them has a unique age.

Now the key observation is that the distribution of age is in general *not* uniform: some ages have a higher frequency within the population than others. It means that we are in the heterogeneous setting of the previous Section.

Our experiments are based on the age distribution of all 428 Dutch municipalities that existed in 2010. For each of them we computed the Kullback-Leibler distance κ; let κ_j be the Kullback-Leibler distance of municipality j.

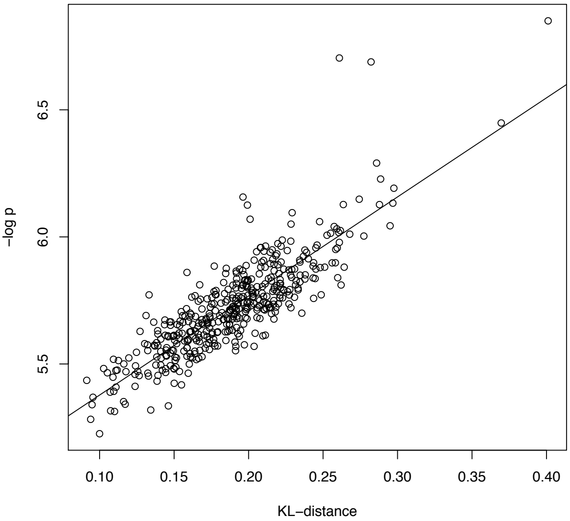


Figure 4.1: For all Dutch municipalities: the Kullback-Leibler distance and the estimated uniqueness probability, when revealing age.

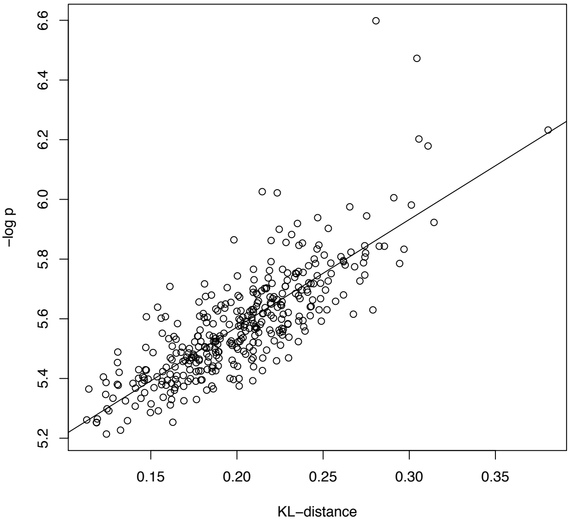


Figure 4.2: For all Dutch municipalities: the Kullback-Leibler distance and the estimated uniqueness probability, when revealing age and gender.

More specifically, with φ_{ij} the fraction of the population with age i (for i ranging between 0 and the maximum age, say M) in municipality j (where obviously $\sum_{i=0}^{M} \varphi_{ij} = 1$ for all j), we have

$$\kappa_j = \sum_{i=0}^{M} \varphi_{ij} \log \frac{\varphi_{ij}}{1/(M+1)};$$

the $1/(M+1)$ is the uniform density on $\{0, \ldots, M\}$. In our experiments we took $M = 94$ (thus neglecting a tiny fraction of the population).

In our experiments we took $k = 29$, such that under uniformity we would have a uniqueness probability $\pi_{\rm u}(29, 95) = 0.84\%$. The approximation of the uniqueness probability p_j for municipality j is therefore $0.84 \cdot 10^{-2} \cdot e^{-k^2/N \cdot \kappa_j}$. The accuracy of this approximation for municipality j can be validated by sampling (independently) n_+ groups of size k from age distribution $(\varphi_{0j}, \ldots, \varphi_{Mj})$, and to check for each of these samples whether all individuals included are unique (if yes, then increase counter n). Then the uniqueness probability of municipality j can be estimated by $\hat{p} := n/n_+$. To guarantee that this estimate is sufficiently reliable, we should have that the ratio of confidence interval's half-width and the estimate (known as the *relative efficiency*) is below some predefined number r, say, 10%, which means that

$$\frac{t_\alpha \sigma(\hat{p})}{\hat{p}} < r,$$

where $\sigma(\hat{p})$ is the standard error of the estimate, which roughly equals

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_+}} \approx \sqrt{\frac{\hat{p}}{n_+}},$$

and t_α is the t-value corresponding to confidence α (1.96 for $\alpha = 0.95$). An easy computation shows that the number n_+ of experiments needed to make sure that the relative efficiency is below r, is $t_\alpha^2/(r^2\hat{p})$. In the setting of this experiment, with $r = 0.1$ and a uniqueness probability of roughly one percent, and choosing $\alpha = 0.95$, it turns out that we have to sample until the number of 'unique samples' (that is, the n_+) is about 400. This procedure gives us reliable estimates for the uniqueness probabilities of all municipalities; we call these $\hat{p}_1$ up to $\hat{p}_{428}$.

The question is to what extent the approximation

$$p_j = 0.84 \cdot 10^{-2} \cdot e^{-k^2/N \cdot \kappa_j}$$

is valid, and to this end we can now compare the $0.84 \cdot 10^{-2} \cdot e^{-k^2/N \cdot \kappa_j}$ with the $\hat{p}_j$, for $j = 1$ up to 428. If these numbers would exactly match, then we would have that $\log(0.84 \cdot 10^{-2}) - k^2/N \cdot \kappa_j = \log p_j$, or, in other words, that the logarithm of the uniqueness probability depends linearly on the Kullback-Leibler distance. To study the validity of this relation, we plotted in Figure 4.1 the value of κ_j against $\log \hat{p}_j$; each dot represents one municipality j.

The main conclusion from Figure 4.1 is that there is a remarkably good fit, in that the cloud resembles a straight line quite well. The line drawn represents the *least squares fitting*. The percentage of variance that can be explained by the estimator, usually denoted by R^2, provides a measure of the quality of the fit; we obtained $R^2 \approx 0.72$ (popularly: the estimator explained 72% of the variance). We ran the same experiment but then for target probabilities in the order of 10^{-3} and 10^{-4} (rather than the 0.83% of the above experiment); these yield values of the R^2 of even 0.79 and 0.82, respectively.

Another general conclusion is that the use of $\pi_{\mathrm{u}}(k, N)$ without correction by $e^{-\kappa}$ would lead to substantially overestimating the uniqueness probability. Noting that $e^{-5.8} = 3.0 \cdot 10^{-3}$ (where -5.8 is a typical value for $\log p_j$, as seen in Figure 4.1) indicates that the naive estimate $\pi_{\mathrm{u}}(29, 95) = 8.4 \cdot 10^{-3}$ is usually off by a factor of about 3, due to the heterogeneity that was not taken into account.

We performed the same experiments for the combination age and gender (that is, $M = 95 \times 2 = 190$). We took $k = 41$, where it is noted that $\pi_{\mathrm{u}}(41, 190) = 0.95\%$. Figure 4.2 shows that the same effects apply as in the situation in which just age was considered.

4.4.2 Additional experiments

In this Section we report the outcomes of a number of additional experiments; in the numerics we rely on the approximation formula that was developed in Section 4.3.1, and validated in Section 4.4.1.

In a first experiment we study the effect of the group size k; we return to our example of Section 4.4.1, in which the individuals reveal their ages. For clarity of exposition, we chose two municipalities (Laren and Urk) that differ substantially in Kullback-Leibler distance κ (Laren has a κ of 0.0914, Urk has 0.4011). This difference is reflected clearly in the uniqueness probability, as displayed in Figure 4.3. We approximately have

$$\pi(k, N) \approx \exp\left(-\left(\frac{1}{2} + \kappa\right)\frac{k^2}{N}\right).$$

If we would assume uniformity, then $\kappa = 0$; the resulting graph has been displayed as well.

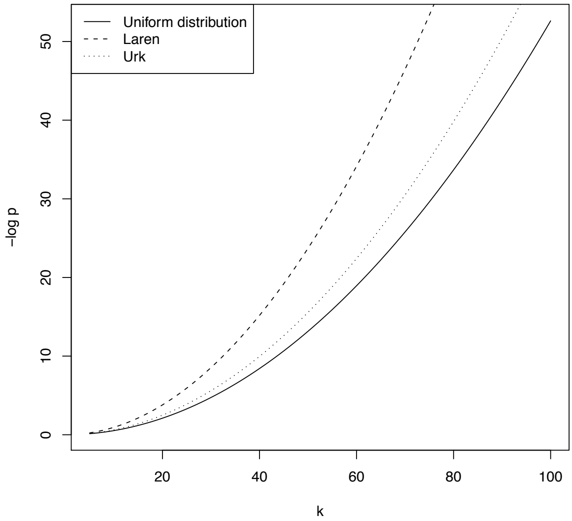


Figure 4.3: For two Dutch municipalities: the uniqueness probability as a function of the group size k; also the curve under uniformity has been added.

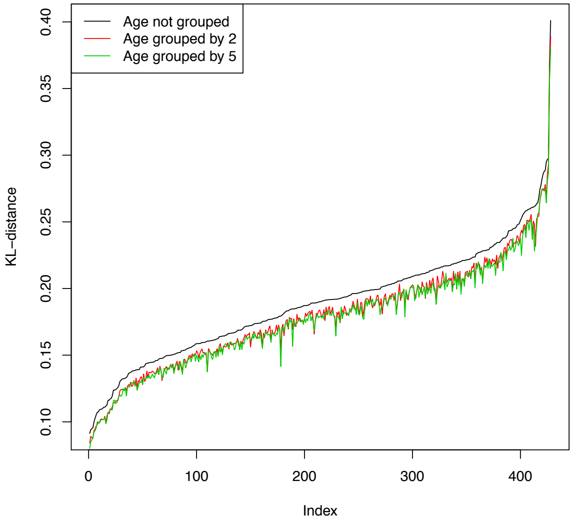


Figure 4.4: For all Dutch municipalities: the effect of aggregated (age) statistics on the KL-distance.

Our next experiment is inspired by the fact that quite often the data available is relatively coarse-grained and aggregated. For example, in the context of Figure 4.2 we had information on the number of individuals that were of any given (age, gender)-pair (there were $95 \times 2 = 190$ such pairs). Suppose, however, that we have less information: we only know the number of males and females, and per age the number of individuals (that is, just 97 numbers, where of course the sum over all ages should match with the sum of the male and female). For this situation the same questions can be posed; notice that the machinery developed in this Chapter does not immediately apply.

Figure 4.4 provides an indication of the effect that aggregated statistics of age have on the Kullback-Leibler distance for age. The figure shows the Kullback-Leibler at the level of individual ages (i.e., not grouped), at the level of age groups of 2 ('age 0-1', 'age 2-3', 'age 4-5', etc.) and age groups of 5 ('age 0-4', 'age 5-9', 'age 10-14', etc.). The horizontal axis is a meaningless index of the municipalities, which for clarity of exposition were ordered by Kullback-Leibler distances for the non-grouped scenario.

4.5 Discussion and future work

One of the common challenges in k-anonymity and its developments is the recognition of quasi-identifiers. The method we proposed in this Chapter provides a new way of efficiently estimating the likelihood that given set of attributes will function as a perfect quasi-identifier, i.e,. that each value of a quasi-identifier unambiguously identifies an individual.

We proposed an approximation for the uniqueness probability when sampling k objects from a population of N, for the situation where the N outcomes are not equally likely. The deviation with respect to the uniform distribution is captured by the Kullback-Leibler distance. The approximation clearly shows how the heterogeneity affects the anonymity: the more heterogeneous the population is, the lower the uniqueness probability. In terms of k-anonymity: the more heterogeneous the population is, the lower the probability that every record in a table will unambiguously identify an individual through the approximated QID.

We emphasize that the anonymity metric used in this Chapter (that is, the uniqueness probability) does not unambiguously reflect the effect for an individual. For instance, if the individual has an age that is relatively rare within the population (the person is relatively old, for instance), then of course he or she is more likely to be identifiable.

Our approximation has several restrictions. First, it can only be applied when the number of subjects k is smaller than the number of quasi-identifier values N. Second, we assumed that while the adversary does not know which identity belongs with each quasi-identifier value, he *does* know the set of identities of those whose data is present within the de-identified data set; this holds, for example, if the adversary attempts to link an identified data set containing all citizens in a municipality to a de-identified data set that also contains all citizens in that municipality. In Chapter 5 and Chapter 6 we will look into different settings.

While the approximation formula allows data holders and policy makers to make predictions about future data collection, and individuals to predict what information the population to which one belongs may better (not) disclose at the end of a survey, there are still a number of challenging open questions. For example, age and gender (as in Figure 4.2) are roughly independent of each other, which makes all computations easier, but quite often when considering multiple quasi-identifiers such a property does not hold. Consider age and marital status: in the Netherlands there will be near-to-zero married people younger than 18 (Dutch law provides for rare exceptions, but none below age 16), therefore, being a widow at a young age is highly unlikely. The question arises how these dependencies should be dealt with.

A useful lemma

Lemma 4.1 *In the scaled heterogeneous model, as $N \to \infty$,*

$$\frac{\mathbb{C}ov(S_i(N), S_j(N))}{N} \to -a^3\alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i+\alpha_j)a}.$$

Proof: From the expressions in Section 5.3, it is straightforward that

$$\begin{aligned} &\frac{\mathbb{C}ov(S_i(N), S_j(N))}{N(\alpha_i f_i\, \alpha_j f_j)} \\ =\quad & a(Na-1)\left(1-\frac{\alpha_i+\alpha_j}{N}\right)^{Na} - a^2 N\left(1-\frac{\alpha_i}{N}\right)^{Na}\left(1-\frac{\alpha_j}{N}\right)^{Na} \\ \sim\quad & a^2 N\left(\left(1-\frac{\alpha_i+\alpha_j}{N}\right)^{Na} - \left(1-\frac{\alpha_i}{N}\right)^{Na}\left(1-\frac{\alpha_j}{N}\right)^{Na}\right), \end{aligned}$$

where $f(n) \sim g(n)$ denotes that $f(n)/g(n) \to 1$ as $n \to \infty$. We have, due to L'Hôpital's rule, for $A, B \in \mathbb{R}$,

$$\lim_{N\to\infty} \frac{\left(1-\frac{A+B}{N}\right)^{Na} - \left(1-\frac{A}{N}\right)^{Na}\left(1-\frac{B}{N}\right)^{Na}}{\frac{1}{N}} = \psi'(0),$$

with $\psi(x) := (1-(A+B)x)^{a/x} - (1-Ax)^{a/x}(1-Bx)^{a/x}$. Using Taylor expansions, we find $\psi'(0) = -aABe^{-a(A+B)}$. Now plugging in $A = \alpha_i$ and $B = \alpha_j$ yields the stated result. □

5 Analysis of singletons in generalized birthday problems

5.1 Introduction

Consider again a population of k people, each of them independently assigned a certain 'feature' (for instance: gender, birthday, age, ...) which is element of $\{1, \ldots, N\}$; in case of gender $N = 2$ (we simplify reality for clarity of exposition), in case of birthday $N = 365$ (neglecting leap years), in case of age N can be taken, say, 95. We assume the distribution of the feature over $\{1, \ldots, N\}$ is given, which is not *a priori* assumed to be uniform (birthday and gender will be roughly uniform, whereas age will not). In the literature this setting is often referred to as that of the *generalized birthday problem* (see Chapter 4), or, alternatively, the birthday problem with unequal probabilities. There is a vast literature on this topic, e.g. [26, 31, 40, 41, 53, 62].

Some of the outcomes will be assigned to just one of the k people in the population; we call these *singletons*. The objective of this Chapter[1] is the analysis of the distribution of the number of singletons S. We subsequently address its mean and variance, as well as a computational scheme for evaluating the distribution of S. It is noted that existing literature, and also Chapter 4, primarily focus on the probability that all k samples are unique (where it was

[1] This Chapter is based on M. Koot and M. Mandjes, *The analysis of singletons in generalized birthday problems*, Probability in the Engineering and Information Sciences, April 2012) [42].

obviously assumed that $k \leq N$).

Similar to Chapter 3, this Chapter will assume that the adversary knows, beforehand, the set of identities of those whose data is present within the de-identified data set.

The contributions of this Chapter are as follows. Our results cover both the homogeneous setting (that is, all outcomes $1, \ldots, N$ being equally likely, that is, have probability $1/N$) and the heterogeneous setting. In the latter, we assume there are F_i 'bins' that have probability α_i/N; obviously we require that $F_1 + \cdots F_d = N$ and $\alpha_1 F_1 + \cdots \alpha_d F_d = N$.

- In Section 5.2 we first derive an explicit expression for the mean number of singletons $\mathbb{E}S$. We then scale the number of samples and number of outcomes per group by N, that is, $k \equiv aN$ and $F_i \equiv f_i N$. We then show that the mean number of singletons in the scaled model, that is, $\mathbb{E}S(N)$ can be accurately approximated by

  $$\mathbb{E}S(N) \approx aN\, e^{-a} \left(1 + \frac{a}{2}(a-2)\kappa\right),$$

  where κ is the Kullback-Leibler distance [23, 47] between our heterogeneous distribution and the homogeneous one. This approximation nicely reflects the impact of heterogeneity on the number of singletons. As we will argue, this effect is both quantitatively and qualitatively different for different values of a: for low values of a, $\mathbb{E}S(N)$ is increasing in κ, whereas for high values of a, $\mathbb{E}S(N)$ is decreasing in κ. We illustrate the theory by an example.
- In Section 5.3 we perform a similar analysis for the variance of S. Again we first derive an exact expression for $\mathbb{V}\mathrm{a}r S$, and then consider approximations in the scaled model.
- Section 5.4 first develops a recursive algorithm that identifies the full distribution of S for the homogeneous case. A crucial role is played by a technique to find the probability of *no* singletons, i.e., $\mathbb{P}(S = 0)$. Then it is demonstrated how to extend the analysis to the heterogeneous case, for which also a more explicit approximation is presented.
- Section 5.5, finally, is devoted to numerical experiments. Based on demographic data of all Dutch municipalities, we estimate the heterogeneity κ, and then assess the accuracy of the approximations for $\mathbb{E}S$ and $\mathbb{V}\mathrm{a}r S$.

5.2 Mean number of identifiable objects

In this Section we analyze the mean number of singletons. We find an exact expression, as well as approximations that show how the heterogeneity affects this quantity.

5.2.1 Explicit expressions

We first consider the homogeneous case: suppose one throws k balls into N bins, uniformly at random. Then, the probability that a given bin contains exactly one ball (a 'singleton') is

$$k \cdot \frac{1}{N}\left(1-\frac{1}{N}\right)^{k-1}; \tag{5.1}$$

here we make use of the fact that the number of balls in that bin obeys a binomial distribution with parameter k and $1/N$. As there are N bins, it follows that the mean number of singletons is

$$\mathbb{E}S = k\left(1-\frac{1}{N}\right)^{k-1}.$$

The result for the homogeneous case is standard, but interestingly it can be extended to the heterogeneous setting in a straightforward manner. Let there be F_i bins that have probability α_i/N; obviously $F_1 + \cdots F_d = N$ and $\alpha_1 F_1 + \cdots \alpha_d F_d = N$. Let N_i the number of balls that end up in bins of type i, and let S_i be the number of singletons among them; observe that N_i has a binomial distribution with parameters k and $\alpha_i F_i/N$. It is clear that

$$\mathbb{E}S_i = kF_i\left(1-\frac{\alpha_i}{N}\right)^{k-1}\frac{\alpha_i}{N}. \tag{5.2}$$

We arrive at the following statement.

Proposition 5.1 *In the heterogeneous model defined above, the mean number of singletons equals*

$$\mathbb{E}S = k\sum_{i=1}^{d} F_i\left(1-\frac{\alpha_i}{N}\right)^{k-1}\frac{\alpha_i}{N}.$$

We now consider the number of singletons $S(N)$ in the asymptotic regime in which there are aN balls, and F_i is scaled by N (that is, $F_i \equiv f_i N$). After straightforward calculus we find the following result.

Proposition 5.2 *In the scaled heterogeneous model defined above, the mean number of singletons satisfies, as $N \to \infty$,*

$$\frac{\mathbb{E}S(N)}{N} \to a\sum_{i=1}^{d}\alpha_i f_i e^{-\alpha_i a}.$$

This result essentially states that the number of singletons equals roughly Na (the number of balls), but thinned by a factor $\sum_{i=1}^{d}\alpha_i f_i \exp(-\alpha_i a)$; from the requirement that $\alpha_1 f_1 + \cdots \alpha_d f_d = 1$ it is immediately seen that this factor is smaller than 1.

5.2.2 Impact of heterogeneity: an approximation

Similar to Chapter 4 and [53], we can assess the impact of heterogeneity by parameterizing $\alpha_i = 1+\beta_i\varepsilon$, for ε typically small; evidently, $\beta_1 f_1 + \cdots \beta_d f_d = 0$. Relying on the Taylor series $e^x = 1 + x + x^2/2 + O(x^3)$, it is now immediate that

$$\begin{aligned} a\sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} &= a e^{-a} \sum_{i=1}^{d} f_i (1+\beta_i\varepsilon) e^{-\beta_i \varepsilon a} \\ &= a e^{-a} \sum_{i=1}^{d} f_i (1+\beta_i\varepsilon)\left(1 - \beta_i\varepsilon a + \frac{1}{2}(\beta_i\varepsilon a)^2\right) + O(\varepsilon^3) \\ &= a e^{-a}\left(1 + \frac{a}{2}(a-2)\sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2\right) + O(\varepsilon^3). \qquad (5.3) \end{aligned}$$

The Kullback-Leibler distance κ of the non-uniform probabilities $(1+\beta_i\varepsilon)/N$ with respect to the uniform probabilities $1/N$ reads, as described in Chapter 4,

$$\kappa := \sum_{i=1}^{d} f_i N \left(\frac{1+\beta_i\varepsilon}{N}\right) \log\left(\left(\frac{1+\beta_i\varepsilon}{N}\right) \Big/ \left(\frac{1}{N}\right)\right) = \frac{1}{2}\sum_{i=1}^{d} f_i (\beta_i\varepsilon)^2 + O(\varepsilon^3).$$

This suggests the approximation

$$\mathbb{E}S \approx k e^{-k/N}\left(1 + \frac{k}{N}\left(\frac{k}{N} - 2\right)\cdot\kappa\right). \qquad (5.4)$$

It can even be computed what the fraction γ_j of bins is that is covered by j balls, when sampling aN balls. In the homogeneous case this leads to the known result that

$$\gamma_j = \lim_{N\to\infty} \binom{aN}{j}\left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{aN-j} = e^{-a}\frac{a^j}{j!};$$

we recognize the Poisson distribution. In the heterogeneous model described above,

$$\gamma_j = \sum_{i=1}^{d} f_i e^{-\alpha_i a}\frac{(\alpha_i a)^j}{j!}.$$

Notice that the γ_j sum to 1, as desired (recall they represent *fractions*). Again parameterizing $\alpha_i = 1 + \beta_i\varepsilon$, for ε small, we obtain

$$\begin{aligned} \gamma_j &= \sum_{i=1}^{d} f_i e^{-a}\frac{(1 - \beta_i\varepsilon a + \frac{1}{2}(\beta_i\varepsilon a)^2)(1 + j\beta_i\varepsilon + \frac{1}{2}j(j-1)(\beta_i\varepsilon)^2)}{j!} + O(\varepsilon^3) \\ &= e^{-a}\frac{a^j}{j!}\left(1 + (a^2 + j(j-1) - 2ja)\cdot\frac{1}{2}\sum_{i=1}^{d} f_i\beta_i^2\varepsilon^2\right) + O(\varepsilon^3). \end{aligned}$$

Replacing a by k/N and $\frac{1}{2}\sum_{i=1}^{d} f_i\beta_i^2\varepsilon^2$ by κ, this gives an approximation for the fraction of bins covered by j balls, as a function of k, N, and the 'non-homogeneity' κ:

$$\gamma_j \approx e^{-k/N}\frac{(k/N)^j}{j!}\left(1+\left(\frac{k^2}{N^2}+j(j-1)-2j\frac{k}{N}\right)\kappa\right). \tag{5.5}$$

5.2.3 Remarks, example

Remark 5.3 The above findings also enable us to compute the fraction ϕ_j of people being in a group of size j; cf. the concept of k-anonymity in privacy [1, 77]. After an elementary computation, we obtain for $j = 1, 2, \ldots$

$$\phi_j = \frac{j\gamma_j}{\sum_{\ell=1}^{\infty}\ell\gamma_\ell} = \sum_{i=1}^{d} f_i\alpha_i e^{-\alpha_i a}\frac{(\alpha_i a)^{j-1}}{(j-1)!}.$$

As before, this can be approximated by an expression in terms of k, N, and κ only, under $\alpha_i = 1+\beta_i\varepsilon$:

$$\phi_j \approx e^{-k/N}\frac{(k/N)^{j-1}}{(j-1)!}\left(1+\left(\frac{k^2}{N^2}+j(j-1)-2j\frac{k}{N}\right)\kappa\right). \tag{5.6}$$

It is a matter of elementary calculus to verify that both the approximation of γ_j (as given in Eqn. (5.5)) and the approximation of ϕ_j (as given in Eqn. (5.6)) add up to 1 (summing over $j = 1, 2, \ldots$), as it should.

Remark 5.4 We now study for which value of k the above approximation (5.4) is maximized. We do so by looking at the scaled version:

$$\max_{a\geq 0} ae^{-a}\left(1+\frac{1}{2}a(a-2)\kappa\right).$$

This yields the first order condition

$$e^{-a}\left((1-a)-\frac{a\kappa}{2}(a-1)(a-4)\right)=0,$$

yielding the optimizer $a = 1$ (which is easily seen to be a maximizer for $\kappa < \frac{2}{3}$).

We observe that (5.4) first increases in k, reaches a maximum N/e at $k^\star = N$, and then decreases, with limiting value 0 as $k \to \infty$. This qualitative behavior can be understood easily. For small k there are few singletons, as there are few samples; for large k quite likely all possible outcomes have been sampled more than once, also resulting in a low number of singletons.

For instance in case of birthdays, assuming they are equally spread over the 365 days, then sampling 365 individuals maximizes the number of identifiable objects, which is (on average) 134.

Remark 5.5 Expression (5.3) confirms the claim that (for $a \leq 2$, at least) the mean number of singletons is maximized by the uniform distribution (that is, $\beta_i = 0$ for all $i = 1, \dots, d$) — this is due to the absence of a linear (in ε) term in the expression in (5.3).

It is observed the mean number of singletons decreases in κ for small a (that is, $a < 2$), but increases for large a (that is, $a > 2$). This can be intuitively understood.

- For small a, most bins will be empty or occupied by just one or two balls. Then heterogeneity leads to a smaller number of singletons, as it increases the probability that two balls end up in the same bin.
- For large a, most bins will be occupied by multiple balls. The more heterogeneity, the larger the probability that it is actually just one ball, thus leading to more singletons.

Example 5.6 Consider the following (somewhat atypical) example. Suppose one has data of a set of individuals, consisting of (a) postal code, and (b) age. Assume that ages range from 0 to 94, and (for the moment) that all these ages are equally likely — below we indicate how to deal with heterogeneity. Now suppose that k people share a postal code, and that k needs to be chosen so as to optimize the number of uniquely identifiable individuals.

If there is no penalty imposed on the number of postal codes introduced, it is evident that it is optimal to give any individual her or his own postal code. It is more realistic to assume that there are costs, say C, for every postal code introduced. If the set of people has size M, then about $(M/k) \cdot k \exp(-k/N)$ individuals can be uniquely identified. We are therefore faced with the optimization problem

$$\max_k M e^{-k/N} - C\frac{M}{k};$$

observe that the value of M is irrelevant when determining the optimum group size $k^\star$.

It is a matter of elementary computation to conclude that for $C = 1$ one should have 10 individuals per postal code; for $C = 10$ we obtain 38. Adaptation to the heterogeneous case is straightforward: then

$$M e^{-k/N}\left(1 + \frac{k}{N}\left(\frac{k}{N} - 2\right) \cdot \kappa\right) - C\frac{M}{k}$$

should be maximized.

5.2.4 Continuous model

The result of Proposition 5.2 can be further refined. We now present its continuous counterpart. Let $\varphi(\cdot)$ be a continuous density on $[0, 1]$, and define the probability that an arbitrary ball is put in bin i by

$$\Phi_{i,N} := \int_{(i-1)/N}^{i/N} \varphi(x)\mathrm{d}x.$$

Then, in the scaled model, due to Proposition 5.1,

$$\begin{aligned}\lim_{N\to\infty} \frac{\mathbb{E}S(N)}{N} &= a \lim_{N\to\infty} \sum_{i=1}^{N} (1-\Phi_{i,N})^{aN-1}\Phi_{i,N} \\ &= a \lim_{N\to\infty} \sum_{i=1}^{N} \left(1-\frac{1}{N}\varphi\left(\frac{i}{N}\right)\right)^{aN-1} \frac{1}{N}\varphi\left(\frac{i}{N}\right).\end{aligned}$$

Now it is a matter of straightforward analysis to derive the following result.

Proposition 5.7 *In the scaled heterogeneous model defined above, the mean number of singletons satisfies, as $N \to \infty$,*

$$\frac{\mathbb{E}S(N)}{N} \to a \int_0^1 \varphi(x) e^{-\varphi(x)a} dx.$$

Example 5.8 Consider the density $\varphi_\gamma(x) = \gamma(x-\frac{1}{2})+1$, with $\gamma \in [-2,2]$. The substitution $y := a(\gamma(x-\frac{1}{2})+1)$ substitution yields

$$a \int_0^\infty \varphi_\gamma(x) e^{-\varphi_\gamma(x)a} \mathrm{d}x = \frac{1}{\gamma a} \int_{a(1-\gamma/2)}^{a(1+\gamma/2)} y e^{-y} \mathrm{d}y.$$

After some calculus, this expression can be rewritten as

$$e^{-a}\left(\frac{a+1}{\gamma a}\right)\left(e^{a\gamma/2}-e^{-a\gamma/2}\right) - \frac{e^{-a}}{2}\left(e^{a\gamma/2}+e^{-a\gamma/2}\right).$$

For instance for $\gamma = 2$, we thus find

$$\frac{\mathbb{E}S(N)}{N} \to \frac{1-e^{-2a}-2ae^{-2a}}{2a},$$

which is maximized for $a \approx 0.90$.

We could use an approximation in the spirit of (5.3) to approximate $\mathbb{E}S$. For this model, it takes a straightforward computation to obtain that the Kullback-Leibler distance, as a function of the 'asymmetry parameter' γ equals

$$\kappa = \frac{1}{2\gamma}\left(\left(1+\frac{\gamma}{2}\right)^2 \log\left(1+\frac{\gamma}{2}\right) - \left(1-\frac{\gamma}{2}\right)^2 \log\left(1-\frac{\gamma}{2}\right) - 1\right);$$

Observe that κ is minimal for $\gamma = 0$ (corresponding with the uniform distribution), and symmetric around 0, as could be expected.

5.3 Variance of the number of identifiable objects

This Section considers the variance of the number of singletons. Again, after giving exact expressions and approximations, we study the impact of heterogeneity.

5.3.1 Explicit expressions

As usual, we start with the homogeneous case. Let I_j be the indicator function of the event that there is exactly one ball in bin j (where $j = 1, \dots, N$). It was observed before that

$$\mathbb{P}(I_j = 1) = k \cdot \frac{1}{N}\left(1 - \frac{1}{N}\right)^{k-1},$$

but it is easily verified that for $j_1 \neq j_2$,

$$\mathbb{P}(I_{j_1} = 1, I_{j_2} = 1) = k(k-1) \cdot \frac{1}{N^2}\left(1 - \frac{2}{N}\right)^{k-2}.$$

Observe that $S = I_1 + \cdots I_N$. From

$$\mathbb{V}\mathrm{a}rS = \mathbb{E}S^2 - (\mathbb{E}S)^2 = \sum_{i=1}^{N} \mathbb{E}I_i + \sum_{i \neq j} \mathbb{E}I_i I_j - (\mathbb{E}S)^2,$$

we find (noting that there are $N(N-1)$ terms for which $i \neq j$)

$$\mathbb{V}\mathrm{a}rS = k\left(1 - \frac{1}{N}\right)^{k-1} + k(k-1) \cdot \frac{N-1}{N}\left(1 - \frac{2}{N}\right)^{k-2} - k^2\left(1 - \frac{1}{N}\right)^{2k-4}.$$

Again we can consider the random variable $S(N)$, after scaling $k \equiv aN$. Directly from the previous formula, we obtain

$$\lim_{N \to \infty} \frac{\mathbb{V}\mathrm{a}rS(N)}{N} = ae^{-a} + a^2 \lim_{N \to \infty} N\left(\left(1 - \frac{2}{N}\right)^{aN} - \left(1 - \frac{1}{N}\right)^{2aN}\right).$$

It is clear that

$$\lim_{N \to \infty} N\left(\left(1 - \frac{2}{N}\right)^{aN} - \left(1 - \frac{1}{N}\right)^{2aN}\right) = f'(0),$$

with $f(x) = (1-x)^{2a/x}$. Straightforward calculus yields that $f'(0) = -ae^{-2a}$. In other words, in the homogeneous model,

$$\lim_{N \to \infty} \frac{\mathbb{V}\mathrm{a}rS(N)}{N} = ae^{-a}(1 - a^2 e^{-a}).$$

We now consider the heterogeneous case. Recall the standard relation

$$\mathbb{V}\mathrm{a}r S = \sum_{i=1}^{d} \mathbb{V}\mathrm{a}r S_i + \sum_{i \neq j} \mathbb{C}\mathrm{o}v(S_i, S_j).$$

Let us first compute $\mathbb{V}\mathrm{a}r S_i = \mathbb{E} S_i^2 - (\mathbb{E} S_i)^2$. Observing that we already found $\mathbb{E} S_i$ in (5.2), we now focus on $\mathbb{E} S_i^2$. Conditioning on the number of objects that ends up in group i (which we assumed to have F_i elements, each with probability α_i/N) yields

$$\mathbb{E} S_i^2 = \sum_{j=0}^{k} \binom{k}{j} \left(\frac{\alpha_i F_i}{N}\right)^j \left(1 - \frac{\alpha_i F_i}{N}\right)^{k-j} \cdot \mathbb{E}(S_i^2 \,|\, N_i = j).$$

As earlier,

$$\mathbb{E}(S_i^2 \,|\, N_i = j) = j \left(1 - \frac{1}{F_i}\right)^{j-1} + j(j-1) \cdot \frac{F_i - 1}{F_i} \left(1 - \frac{2}{F_i}\right)^{j-2},$$

so that

$$\mathbb{E} S_i^2 = k F_i \left(1 - \frac{\alpha_i}{N}\right)^{k-1} \frac{\alpha_i}{N} + k(k-1) F_i^2 \left(1 - \frac{2\alpha_i}{N}\right)^{k-2} \left(\frac{\alpha_i}{N}\right)^2. \tag{5.7}$$

We are now left with computing $\mathbb{C}\mathrm{o}v(S_i, S_j) = \mathbb{E} S_i S_j - \mathbb{E} S_i \, \mathbb{E} S_j$ for $i \neq j$. As we already know $\mathbb{E} S_i$, we focus on $\mathbb{E} S_i S_j$. It holds that

$$\begin{aligned}\mathbb{E} S_i S_j = \sum_{\ell_i=0}^{k} \sum_{\ell_j=0}^{k-\ell_i} \binom{k}{\ell_i, \ell_j} \left(\frac{\alpha_i F_i}{N}\right)^{\ell_i} \left(\frac{\alpha_j F_j}{N}\right) \ell_j \\ \left(1 - \frac{\alpha_i F_i}{N} - \frac{\alpha_j F_j}{N}\right)^{k-\ell_i-\ell_j} \cdot \mathbb{E}(S_i S_j \,|\, N_i = \ell_i, N_j = \ell_j),\end{aligned}$$

and in addition a conditional independence argument yields that

$$\mathbb{E}(S_i S_j \,|\, N_i = \ell_i, N_j = \ell_j) = \ell_i \left(1 - \frac{1}{F_i}\right)^{\ell_i - 1} \ell_j \left(1 - \frac{1}{F_j}\right)^{\ell_j - 1}.$$

Standard computations now yield that

$$\mathbb{E} S_i S_j = k(k-1) F_i F_j \left(1 - \frac{\alpha_i}{N} - \frac{\alpha_j}{N}\right)^{k-2} \frac{\alpha_i}{N} \frac{\alpha_j}{N}. \tag{5.8}$$

Now all the above findings can be collected.

Proposition 5.9 *In the heterogeneous model defined above, the variance of the number of singletons equals*

$$\mathbb{V}arS = \sum_{i=1}^{d} \left(\mathbb{E}S_i^2 - (\mathbb{E}S_i)^2\right) + \sum_{i \neq j} \left(\mathbb{E}S_iS_j - \mathbb{E}S_i\,\mathbb{E}S_j\right),$$

with $\mathbb{E}S_i$ *given by* (5.2), $\mathbb{E}S_i^2$ *by* (5.7), *and* $\mathbb{E}S_iS_j$ *by* (5.8).

We now again look at the scaled variant. As before,

$$\frac{\mathbb{V}\mathrm{ar}S_i(N)}{N} \to a\alpha_i f_i e^{-\alpha_i a}\left(1 - a^2\alpha_i f_i e^{-\alpha_i a}\right).$$

Also, due to Lemma 4.1,

$$\frac{\mathbb{C}ov(S_i(N), S_j(N))}{N} \to -a^3\alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i+\alpha_j)a}.$$

We arrive at the following statement.

Proposition 5.10 *In the scaled heterogeneous model defined above, the variance of the number of singletons satisfies, as* $N \to \infty$,

$$\begin{aligned}
\frac{\mathbb{V}arS(N)}{N} &\to \sum_{i=1}^{d} a\alpha_i f_i e^{-\alpha_i a}\left(1 - a^2\alpha_i f_i e^{-\alpha_i a}\right) - \sum_{i \neq j} a^3\alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i+\alpha_j)a} \\
&= a\sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} - a^3 \sum_{i=1}^{d}\sum_{j=1}^{d} \alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i+\alpha_j)a} \\
&= a\sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} - a^3 \left(\sum_{i=1}^{d} \alpha_i^2 f_i e^{-\alpha_i a}\right)^2.
\end{aligned}$$

5.3.2 Impact of heterogeneity; an approximation

We again parameterize $\alpha_i = 1 + \beta_i\varepsilon$. We already observed that

$$a\sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} = ae^{-a}\left(1 + \frac{a}{2}(a-2)\sum_{i=1}^{d} f_i\beta_i^2\varepsilon^2\right) + O(\varepsilon^3),$$

whereas it turns out that

$$a^3\left(\sum_{i=1}^{d} \alpha_i^2 f_i e^{-\alpha_i a}\right)^2 = a^3 e^{-2a}\left(1 + (2-3a)\sum_{i=1}^{d} f_i\beta_i^2\varepsilon^2\right) + O(\varepsilon^3).$$

This leads to the approximation (for the unscaled model)

$$\mathbb{V}\mathrm{ar}S \approx ke^{-k/N}\left(1 - \frac{k}{N}\left(\frac{k}{N} - 2\right)\kappa\right) - \frac{k^3}{N^2}e^{-2k/N}\left(1 + \left(4 - 6\frac{K}{N}\right)\kappa\right).$$

5.3.3 Continuous model

We now consider the continuous model, analogously to Section 5.2.4; the probability of a ball being put in bin i is $\Phi_{i,N}$, equalling the integral over the density $\varphi(\cdot)$ between $(i-1)/N$ and i/N, for $i = 1, \ldots, N$. The proof of following result is similar to the proof of Proposition 5.7.

Proposition 5.11 *In the scaled heterogeneous model defined above, the variance of the number of singletons satisfies, as $N \to \infty$,*

$$\frac{\mathbb{V}arS(N)}{N} \to \int_0^1 a\varphi(x)(1 - a^2\varphi(x))e^{-\varphi(x)a}\,dx.$$

We conjecture that $(S(N) - \mathbb{E}S(N))/\sqrt{\mathbb{V}\mathrm{ar}S(N)/N}$ converges to a standard Normal random variable.

5.4 Probability of at least one singleton

Let $\xi(k, N)$ be the probability of at least one identifiable object, that is, the probability $\mathbb{P}(S > 0)$ of at least one singleton. Particularly if k is large relative to N, this is an interesting anonymity metric. (An example could be: suppose one receives data about the ages of 300 people; is there anyone among these 300 people whose age is unique within that group?). In this Section we develop a recursive scheme to evaluate $\xi(k, N)$.

5.4.1 Recursive scheme

We analyze this probability by computing the probability $\zeta(k, N)$ of its complement (that is, *no* singletons); we start with the homogeneous case. Consider an arbitrary ball that ends up in an arbitrary bin. As there should not be singletons, it means that at least one more ball (out of the remaining $k-1$) should be in that bin as well; realize that the number of balls that are in that bin (apart from the tagged one) follows a binomial distribution with parameters $k-1$ and $1/N$. We thus find

$$\zeta(k, N) = \sum_{j=1}^{k-1} \binom{k-1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{k-1-j} \zeta(k-1-j, N-1).$$

The initialization of this recursion is $\zeta(k, 1) = 1$ and $1 - \zeta(0, N) = \zeta(1, N) = 0$ for any $k = 2, 3 \ldots$ and $N = 1, 2, \ldots$ The first steps can be done easily:

$$\zeta(2, N) = \frac{1}{N}, \quad \zeta(3, N) = \frac{1}{N^2}, \quad \zeta(4, N) = \frac{3N-2}{N^3},$$

and, with a bit more effort,

$$\zeta(5,N)=\frac{10N-9}{N^4},\quad \zeta(6,N)=\frac{15N^2-20N+6}{N^5},$$
$$\zeta(7,N)=\frac{105N^2-259N+155}{N^6}.$$

Table A.1 in Appendix A presents the values of $\zeta(k,N)$ for $k=1,\dots,50$ and $N=1,\dots,20$. It shows that $\zeta(k,N)$ goes to 1 for k large. In addition, for fixed k, $\zeta(k,N)$ decreases with N. A nice sanity check for formulae for $\zeta(k,N)$ is the relation ($k\geq 3$)

$$\zeta(k,2)=1-\frac{k}{2^{k-1}}.$$

5.4.2 Full distribution of number of singletons

The above results immediately lead to the full distribution of the number of singletons S; for ease we restrict ourselves to the uniform case. It is seen that

$$\mathbb{P}(S=j)=\binom{N}{j}k(k-1)\cdots(k-j+1)\left(\frac{1}{N}\right)^j\left(1-\frac{j}{N}\right)^{k-j}\zeta(k-j,N-j);$$

evidently $\mathbb{P}(S=0)=\zeta(k,N)$. Evidently, for $k\leq N$, we already knew from the standard birthday problem that

$$\mathbb{P}(S=k)=\frac{N!/(N-k)!}{N^k}.$$

5.4.3 Heterogeneous case

Once we have computed the numbers $\zeta(i,j)$ (that correspond to the homogeneous case), it is fairly easy to deal with the heterogeneous case:

$$\zeta(k,N)=\sum_{\boldsymbol{j}}\binom{k}{j_1,\dots,j_d}\prod_{i=1}^d\left(\frac{\alpha_iF_i}{N}\right)^{j_i}\zeta_{\rm u}(j_i,F_i),$$

where $\zeta_{\rm u}(\cdot\,,\,\cdot)$ refers to the probability of no singletons in the uniform case, and the summation is over vectors $\boldsymbol{j}\in\{0,1,\dots\}^d$ such that $j_1+\cdots j_d=k$. It is observed that this expression is hard to evaluate, as one has to sum over all vectors $\boldsymbol{j}$ whose entries add up to k, whose number grows explosively in k. This explains the need for approximations. One such approximation relies on the idea of replacing the multinomial distribution by the corresponding Poisson distribution (where the individual components are assumed to be independent). Then one obtains

$$\zeta(k,N)=\prod_{i=1}^d\left(\sum_{j=0}^\infty\left(\exp\left(-\frac{k\alpha_iF_i}{N}\right)\left(\frac{k\alpha_iF_i}{N}\right)^j\Big/ j!\right)\cdot\zeta_{\rm u}(j,F_i)\right).$$

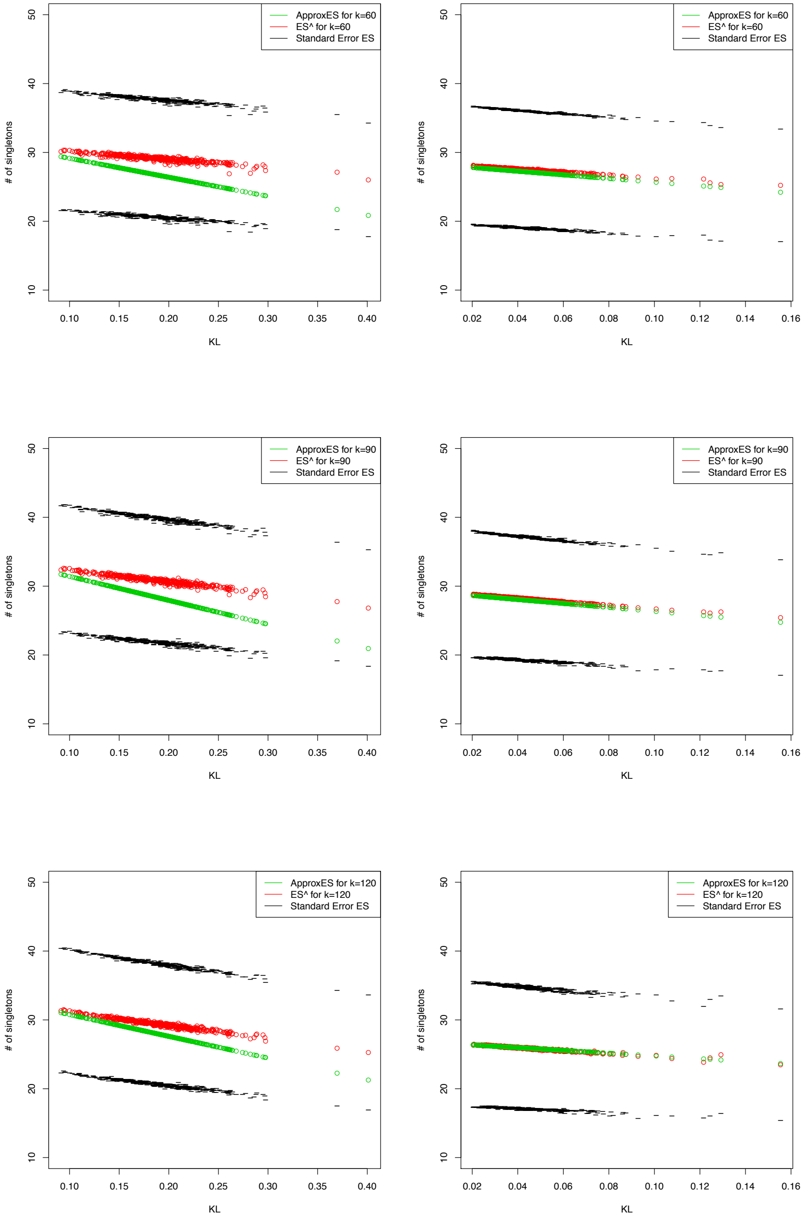


Figure 5.1: Mean number of singletons, as a function of the Kullback-Leibler distance κ. Left panels: full population; right panels: ages 0–79 only. Top to bottom: $k = 60, 90, 120$.

5.5 Numerical experiments

In this Section we report on numerical experiments that we ran with demographic data of all 428 Dutch municipalities that existed in 2010. For all of them we first determined the distribution of age over the population. Ages were truncated at 94 years, leaving us with 95 bins, viz. 0 up to and including 94). Then we computed on the basis of this data the Kullback-Leibler distance κ, and the mean and variance of the number of singletons. The mean $\mathbb{E}S$, and the mean plus/minus twice the standard error $\mathbb{E}S \pm 2\sqrt{\mathbb{V}\mathrm{ar}S}$ are depicted in the left panels of Fig. 5.1, as a function of the κ — each dot represents one municipality.

We also include Approximation (5.4), which is a linear function of κ. As argued in the derivation, it is supposed to perform well if the distance with respect to the uniform distribution is relatively modest. From the left panels of Fig. 5.1, it is seen that the approximation does not give an accurate prediction. This is mainly due to the fact that the distribution is highly non-uniform for the higher ages (ages above, say, 85 are hardly represented). In the right panels we performed the same experiments, but just for the ages 0 up to and including 79, and there we indeed see an excellent fit.

Although the left panels indicate that Approximation (5.4) does not yield an accurate estimate for the mean number of singletons $\mathbb{E}S$ in case the non-uniformity is too strong, the (nearly) linear shape of the scatter plot does show that knowledge of the Kullback-Leibler distance accurately predicts $\mathbb{E}S$. One could for instance approximate $\mathbb{E}S$ (as a function of κ) by the linear regression $\delta_0 + \delta_1\kappa$, where δ_0 and δ_1 are estimated by a least squares procedure.

The left panels show that the mean number of singletons is highest for $k = 90$, which could be expected from Remark 5.4 (recall that $N = 95$ here). Additional experiments (not reported on here) show that when leaving out the ages 80–94, the mean number of singletons is indeed highest around $k = 80$.

Fig. 5.2 shows a scatter plot of the Kullback-Leibler distance κ and the variance $\mathbb{V}\mathrm{ar}S$. For the full population we observe three decreasing, more or less linear lines; when leaving out the ages 79–94 there is hardly any sensitivity in κ.

5.6 Concluding remarks

This Chapter presented an analysis of the number of singletons in the setting of the generalized birthday problem. Various metrics have been studied. Special attention has been paid to obtaining insight into the impact of heterogeneity on the number of singletons. In Chapter 7 we will discuss applications of the theory developed here. Future research includes extensive testing with demographic data.

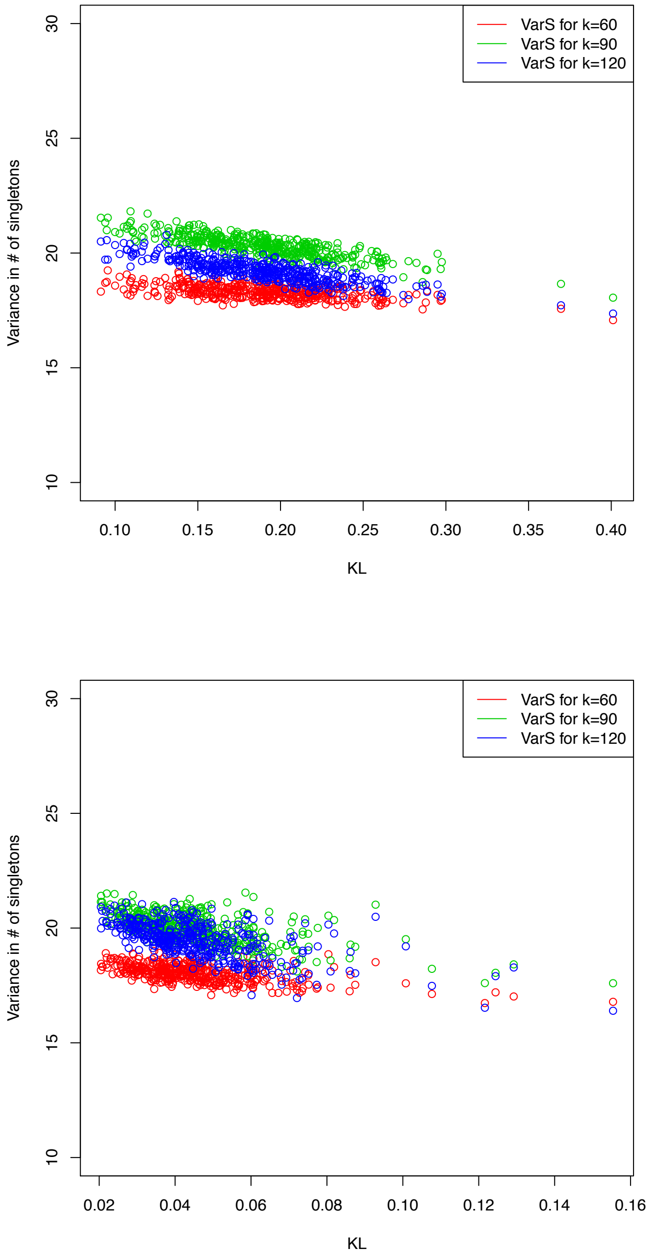


Figure 5.2: Variance of the number of singletons, as a function of the Kullback-Leibler distance κ. Left panel: full population; right panel: ages 0–79 only.

6 Practical guidelines on correlation and aggregation

6.1 Introduction

One objective of this thesis is to quantify to what extent it is possible to unambiguously identify a person from a few pieces of information, such as postal code and age. Recalling Chapter 5, consider the setting in which one is asked to anonymously fill out a questionnaire, at the end of which one is asked to reveal postal code and age. We argued that the above setting gives rise to a set of questions that are mathematically interesting. Considering a group of k individuals that share a postal code: how many of them have an age that is *unique* within that group? Recall from Chapter 4 and Chapter 5 that one can view this question as a generalized birthday problem: one samples k times from a distribution on a finite set (say, $\{1, \ldots, N\}$), and is interested in the distribution of the number of *singletons* S, where singletons are defined as the outcomes that show up precisely once in the sample of size k.

Previous research focused primarily on determining the probability that *all* outcomes are unique (that is, all k people are unambiguously identifiable in the setting that all outcomes are equally probable). There is vast literature on characterization of this quantity; for example, see [26, 31, 40, 41, 53, 62]. However, the scenario in which the N possible outcomes are equally likely to occur is hardly ever met in practice. In addition, focus was on the probability of *all* k individuals corresponding to singletons, and less on the analysis of the *number of singletons* S, for instance in terms of its expectation $\mathbb{E}S$. As argued in Chapter 5, this is clearly a relevant quantity, because a lower number of singletons can be indicative of a higher degree of privacy (we define 'degree of privacy' in terms of the number of persons from which one can't be distinguished using only, in our example, age and postal code). A challenging question is how non-uniformity of the distribution on $\{1, \ldots, N\}$ affects the number of singletons. Note that besides singletons, also doubletons (indistinguishability from one other person), tripletons (indistinguishability from two other persons), etc., may be relevant, as with minor additional effort, these persons can be identified as well.

The primary objective of this Chapter[1] lies in providing practical guidelines for the analysis of the distribution of the number of singletons S and related quantities. In addition, whereas Chapter 5 described the effect of non-uniformity on anonymity, this Chapter asserts the correctness of that description via numerical analysis. The contributions of this Chapter are as follows:

- Section 6.2 recalls the approach to quantifying the effect of non-uniformity on identifiability that was developed in Chapter 5. We advocate the use of an approximation in which the non-uniformity is summarized by a single number, the *Kullback-Leibler distance* [47]. We assess the accuracy of this approach using numerical validation based on real data from Dutch municipalities. Our experiments show that our formulas yield reliable approximations for the metrics under study; in addition, it is shown that estimates that take non-uniformity into account outperform estimates that assume uniformity. These results are presented in Section 6.2.

- Section 6.3 quantifies how aggregation influences identifiability. Consider a questionnaire in which one reveals weight in kilograms. There is a difference between rounding it to the nearest integer and rounding it to the nearest even number. In the former case there will be a lower degree of privacy. In mathematical terms: suppose one is asked to reveal their weight, rounded to a multiple of Δ, what is the impact of Δ on the number of singletons S? Clearly, if Δ is close to zero, then S will be close to k, but how does $\mathbb{E}S$ decrease with Δ? For the case of non-uniform probabilities, we develop an explicit relation between $\mathbb{E}S$ and the aggregation 'interval' Δ. Formulas are derived and tested for the special case of a Normal distribution.

- Section 6.4 quantifies how correlation between variates influences identifiability. Consider a questionnaire in which one is asked to reveal not

[1] This Chapter is based on M. Koot, M. Mandjes, G. van 't Noordende and C. de Laat, *A Probabilistic Perspective on Re-Identifiability*, Mathematical Population Studies, submitted November 2011 [44].

only weight, but also height. The question we address is: to what extent does the correlation between height and weight affect the number of singletons? One would expect that the stronger the correlation between the two variates, the less information the second variate adds to the first variate. Indeed, Section 6.4 confirms that correlated variates yield less singletons than independent variates. Formulas are derived and tested for the special case where the two variates correspond to a two-dimensional Normal distribution.

6.2 Analysis of singletons

In this Section we consider the following setting. Let X be a single-dimensional random variable, defined on a subset of $\mathbb{R}$. We write $F_i(\Delta) := \mathbb{P}(X \in [i\Delta, (i+1)\Delta))$, so that $\sum_i F_i(\Delta) = 1$. We sample k times, independently, from the distribution of X, and wonder how many intervals $[i\Delta, (i+1)\Delta))$ are occupied by just a single observation; in the sequel we refer to these intervals as to *singletons*. S denotes the number of these singletons.

Note that we cover the setting where X is an integer — for instance, if one is asked to fill out age in years, and wants to quantify the identifiability, X lives on $\{0, \ldots, N\}$, where N is some 'practical' upper bound (perhaps 90 or 100); Δ has to be chosen 1 then. Suppose one is asked to round age to a multiple of two, then this corresponds to picking $\Delta = 2$, etc.

6.2.1 General formulas

Due to the fact that S can be written as the sum of the number of singletons in disjoint intervals, we have the following evident expression for the mean number of singletons:

$$\mathbb{E}S = \sum_i \mathbb{E}S_i = \sum_i k\,(1 - F_i(\Delta))^{k-1} \times F_i(\Delta);$$

here the random variable S_i equals 1 if there is a singleton in the interval $[i\Delta, (i+1)\Delta))$ and 0 else, so that $\mathbb{E}S_i$ can be interpreted as the probability that there is a singleton in $[i\Delta, (i+1)\Delta))$.

In a similar way we can express the number of *doubletons* D. Note that we define doubletons as the intervals of the type $[i\Delta, (i+1)\Delta))$ in which two realizations are present; clearly, the number of realizations that corresponds to a doubleton is therefore $2D$. For the expected number of doubletons we have

$$\begin{aligned}\mathbb{E}D &= \sum_i \mathbb{E}D_i = \sum_i \binom{k}{2} (1 - F_i(\Delta)))^{k-2} \times (F_i(\Delta))^2 \\ &= \frac{1}{2}k(k-1) \times \sum_i \left((1 - F_i(\Delta)))^{k-2} \times (F_i(\Delta))^2\right),\end{aligned}$$

where D_i is 1 if there is a doubleton in the interval $[i\Delta, (i+1)\Delta))$ and 0 else. Clearly, tripletons, quadrupletons, etc., can be dealt with similarly. Indeed, let η_j be the mean number of intervals in which j objects are present; then, for $j = 1, \dots, k$,

$$\eta_j = \binom{k}{j} \times \sum_i \left((1 - F_i(\Delta)))^{k-j} \times (F_i(\Delta))^j \right).$$

An elementary computation yields that $\sum_{j=1}^{k} j\eta_j = k$, as to be expected. Let ϕ_j be defined as the fraction of realizations that end up in an interval in a group of size j (that is, with $j-1$ other objects); cf. the concept of k-anonymity, that asserts that in a data set containing de-identified personal data, values for any remaining *quasi-identifying* columns occur at least k times in that data set[1, 73, 77]. From the η_j, we can easily compute the ϕ_j:

$$\phi_j = \frac{j\eta_j}{\left(\sum_{\ell=1}^{k} \ell\eta_\ell \right)} = \binom{k-1}{j-1} \sum_i \left((1 - F_i(\Delta)))^{k-j} \times (F_i(\Delta))^j \right); \qquad (6.1)$$

it is readily verified that we indeed have that the ϕ_j sum to 1.

6.2.2 Formulas for a 'nearly uniform' distribution

We now present more explicit formulas for the special case that X is more or less uniformly distributed, say on $\{1, \dots, N\}$. The probability that X equals i is α_i/N, with $\alpha_i = 1 + \beta_i\varepsilon$ with ε small; evidently, it is required that $\sum_i \beta_i = 0$, as the probabilities should sum up to 1. Let κ be the Kullback-Leibler distance [47] of X with respect to the uniform distribution:

$$\kappa := \sum_{i=1}^{N} \left(\frac{1 + \beta_i\varepsilon}{N} \right) \log \left(\left(\frac{1 + \beta_i\varepsilon}{N} \right) \Big/ \left(\frac{1}{N} \right) \right)$$

Through elementary calculus we obtain that, as $\varepsilon \downarrow 0$,

$$\kappa = \frac{1}{2N} \sum_{i=1}^{N} (\beta_i\varepsilon)^2 + O(\varepsilon^3).$$

The following approximation was derived in Chapter 5:

$$\eta_j \approx Ne^{-k/N} \frac{(k/N)^j}{j!} \left(1 + \left(\frac{k^2}{N^2} + j(j-1) - 2j\frac{k}{N} \right) \kappa \right),$$

and also

$$\phi_j \approx e^{-k/N} \frac{(k/N)^{j-1}}{(j-1)!} \left(1 + \left(\frac{k^2}{N^2} + j(j-1) - 2j\frac{k}{N} \right) \kappa \right). \qquad (6.2)$$

In Chapter 5 we did not yet assess the accuracy of these approximations. In the next subsection we will do so, using demographic data of Dutch municipalities.

6.2.3 Experiments

Consider a questionnaire about a privacy-sensitive topic where respondents do not need to disclose their name, but *are* asked to reveal their postal code and age. As argued in the introduction, a natural question is: to what extent do postal code and age, as a pair, uniquely define a person in the corresponding population? The above formulas can be used to estimate ϕ_1, i.e., the fraction of people that are singletons and thus unambiguously identifiable. Here, k denotes the number of the people sharing a particular postal code and N denotes the number of possible ages. We truncate at 79, so that there are 80 different ages; the reason for this is that we found our formulas to yield less accurate results when considering very low frequency outcomes. Our formulas are, however, applicable in the analysis of privacy for the general population.

For 16 Dutch municipalities[2] we have the date of birth of all inhabitants per postal code. Dutch postal codes are typically shared between 20 to 60 people. In our numerical experiments, we take the following approach. Based on the data of *all* people of age $\leq$ 79 within the municipality, we estimate the probabilities α_i/N (for $i = 0, \ldots, 79$), and the Kullback-Leibler distance κ. Then we use this value of κ to estimate the fraction of people that are singleton in a postal code that is shared between k people, using the formulas for ϕ_1 of the previous subsection; here we evaluate both the exact formula (6.1), and the approximation (6.2) based on κ. In addition to ϕ_1, we also analyze ϕ_2 and ϕ_3 (the fraction of the k people involved that is part of a doubleton, tripleton).

We include here graphs that correspond to a larger city (Amsterdam, about 766k inhabitants) and a smaller municipality (Overbetuwe, 46k inhabitants). These municipalities also differ considerably with respect to the non-uniformity of the population in terms of age; the Kullback-Leibler distances are 0.086 and 0.055 respectively. The graphs of Fig. 6.1 show the estimates: for various values of k, we plot ϕ_1, ϕ_2 and ϕ_3 (both based on (6.1) and (6.2)), the empirical result (which we denoted by ψ), and the result if we would assume all ages 0 up to 79 occur perfectly uniformly (that is, $\kappa = 0$).

The approximations we developed have obvious advantages. Only knowing the age distribution of the municipality facilitates the computation of our identifiability metrics. Approximation (6.2) even needs less information: the non-uniformity of the distribution is summarized in a single number. It is clear, however, that this approach assumes that the Kullback-Leibler distance κ is (more or less) constant across the postal codes within the municipality.

The main conclusions of our experiments are: (i) the approximations perform well, as they are usually just a few percent off; (ii) there is hardly any difference between the curves based on (6.1) and (6.2); (iii) if we would have as-

[2] Data from the municipality of Ameland was received after the empirical study presented in Chapter 3, which lists 15 municipalities, had already been completed. We did, however, use that data for the research presented in the current Chapter.

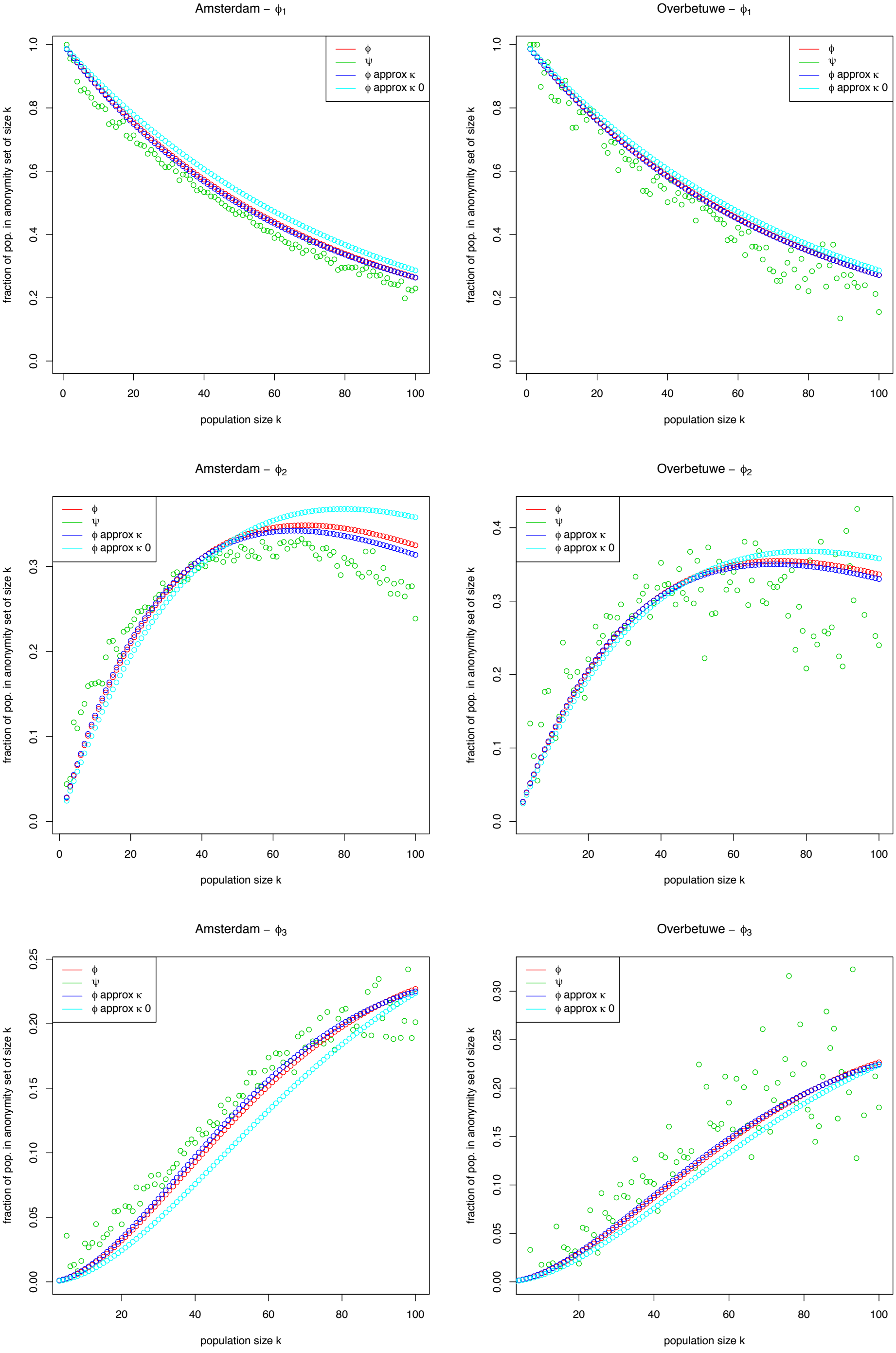


Figure 6.1: ϕ_1, ϕ_2, and ϕ_3 for two municipalities, as a function of the population size of the postal code area.

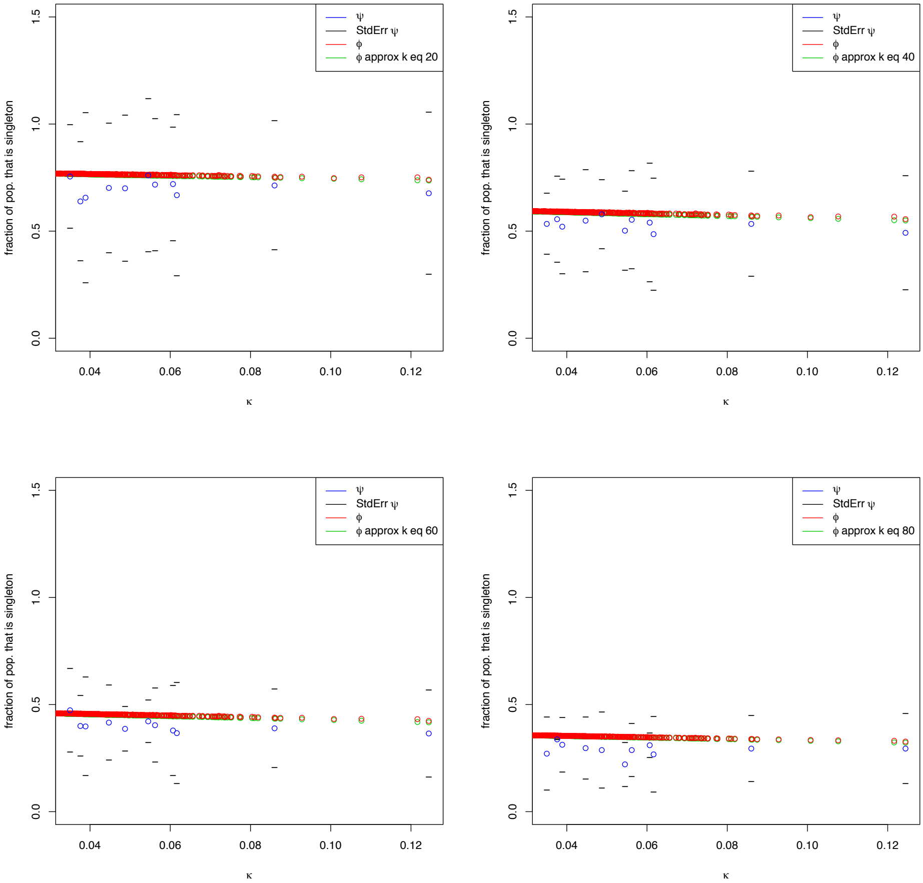


Figure 6.2: ϕ_1 for all municipalities, as a function of the Kullback-Leibler distance κ, for $k = 20, 40, 60, 80$. Notice that the observations (ψ) are accurately predicted (ϕ) by the Kullback-Leibler distance (κ) for various population sizes (k).

sumed uniformity (that is, Kullback-Leibler distance 0), the estimates obtained are systematically worse.

Interestingly, for the number of singletons ϕ_1, we observe that our estimates are typically slightly too high. In other words, in reality there are fewer singletons than what could be expected based on knowledge of the municipality aggregates. This effect can be explained as follows. We observe that the number of singletons decreases in the level of non-uniformity, as captured by the Kullback-Leibler (KL) distance κ. As the estimates of κ are based on the population of the entire municipality, it is likely that within postal code areas there will be a higher discrepancy relative to the uniform distribution (think of a areas with young families, areas with many elderly people); informally: the KL distance per postal code will be higher than the KL distance κ of the entire municipality. Based on this reasoning, one indeed anticipates a smaller number of singletons than what could have been expected based on κ.

In the second series of experiments, we plot, for all Dutch municipalities, the value of ϕ_1 (the fraction of the population that can be unambiguously identified) as a function of the Kullback-Leibler distance κ, again both based on (6.1) and (6.2); we did so for the cases of $k = 20$, $k = 40$, $k = 60$, and $k = 80$ persons in the postal code area, as depicted in Fig. 6.2. For the 16 municipalities for which we have the full data, we can estimate, for the above postal codes sizes, ϕ_1 as well; we have added these estimates and a confidence interval constructed as the estimate ± twice the standard deviation.

6.3 Impact of interval-width

Consider a questionnaire in which one is asked to disclose how much one weighs. Regarding anonymity, it makes quite a difference whether one would be asked to round the weight (in kilograms) to the nearest integer, or to the nearest even number; in the former case there will be a higher level of identifiability. Put in general terms: supposing that one has to reveal their weight, rounded to a multiple of Δ, one would like to quantify the impact of Δ on the number of singletons S. This is the main topic of the present Section.

6.3.1 Theoretical results

In a few special situations (uniform distribution, exponential distribution) the impact of Δ can be examined in an explicit form, in other cases (Normal distribution) approximations need to be developed. In this subsection we cover both these closed-form expressions and approximations.

Uniform distribution. Suppose X is uniformly distributed on $[0, A]$ for some $A > 0$. It is not hard to verify that

$$\mathbb{E}S = k\left(1 - \frac{\Delta}{A}\right)^{k-1}.$$

To study the impact of Δ, we can write, as $\Delta \downarrow 0$,

$$\mathbb{E}S = k\left(1 - (k-1)\frac{\Delta}{A} + \frac{1}{2}(k-1)(k-2)\frac{\Delta^2}{A^2} + O(\Delta^3)\right). \qquad (6.3)$$

Indeed, for $\Delta \downarrow 0$, the mean number of singletons is nearly k, as expected. The formula indicates that for small Δ, $\mathbb{E}S$ decreases roughly linearly in Δ, with slope $k(k-1)/A$.

Exponential distribution. Suppose here that X is exponentially distributed with mean $1/\lambda$. With $L \equiv L_\Delta := e^{-\lambda\Delta}$, it follows that

$$\mathbb{E}S = k\sum_{i=0}^{\infty}\left(1 - L^i + L^{i+1}\right)^{k-1}(L^i - L^{i+1}).$$

This infinite sum can be rewritten to a finite sum, as follows:

$$\begin{aligned}\mathbb{E}S &= k\sum_{i=0}^{\infty}\sum_{j=0}^{k-1}\binom{k-1}{j}\left(-L^i + L^{i+1}\right)^j(L^i - L^{i+1})\\ &= k\sum_{i=0}^{\infty}\sum_{j=0}^{k-1}\binom{k-1}{j}(-1)^j(L^i - L^{i+1})^{j+1}\\ &= k\sum_{j=0}^{k-1}\binom{k-1}{j}(-1)^j\sum_{i=0}^{\infty}(L^{j+1})^i(1-L)^{j+1}\\ &= k\sum_{j=0}^{k-1}\binom{k-1}{j}(-1)^j\frac{(1-L)^{j+1}}{1-L^{j+1}}.\end{aligned}$$

After further computation we obtain, as $\Delta \downarrow 0$,

$$\mathbb{E}S = k\left(1 - \frac{k\Delta\lambda}{2} + \frac{k^2\Delta^2\lambda^2}{6} + O(\Delta^3)\right).$$

We see that this formula has a similar structure as (6.3), and we wonder whether this form holds in general. We now show that this is indeed the case.

General distributions, featuring the Normal distribution. Let $f(\cdot)$ be the density of X, which we assume to be continuous, and to live on $\mathbb{R}$ (if it has only support on just a part of $\mathbb{R}$, the argument below can be adapted in a straightforward manner). Then we have the following obvious approximation, assuming $f(\cdot)$ is differentiable:

$$F_i(\Delta) \approx \Delta \cdot f(i\Delta) + \frac{1}{2}\Delta^2 \cdot f'(i\Delta).$$

This immediately leads to the following expression for the mean number of singletons:

$$\begin{aligned}
\mathbb{E}S &\approx \sum_{i=-\infty}^{\infty} k\left(1-\Delta\, f(i\Delta)-\frac{1}{2}\Delta^2 f'(i\Delta)\right)^{k-1}\times\left(\Delta\, f(i\Delta)+\frac{1}{2}\Delta^2 f'(i\Delta)\right)\\
&\approx \sum_{i=-\infty}^{\infty} k\cdot\left(1-(k-1)\left(\Delta\, f(i\Delta)+\frac{1}{2}\Delta^2 f'(i\Delta)\right)\right)\times\left(\Delta\, f(i\Delta)+\frac{1}{2}\Delta^2 f'(i\Delta)\right)\\
&= k\sum_{i=-\infty}^{\infty}\Delta\, f(i\Delta)-k(k-1)\sum_{i=-\infty}^{\infty}\Delta^2\, f^2(i\Delta)+k\sum_{i=-\infty}^{\infty}\frac{1}{2}\Delta^2\, f'(i\Delta)\\
&\approx k\int_{-\infty}^{\infty} f(x)\mathrm{d}x-\Delta\, k(k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x+\frac{1}{2}\Delta\, k\int_{-\infty}^{\infty} f'(x)\mathrm{d}x\\
&= k-\Delta\cdot\lambda_k,
\end{aligned}$$

where

$$\lambda_k := k(k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x-\frac{1}{2}\,k\int_{-\infty}^{\infty} f'(x)\mathrm{d}x.$$

For various standard distributions including the Normal distribution (but *not* the exponential distribution!), the integral

$$\int_{-\infty}^{\infty} f'(x)\mathrm{d}x=\lim_{x\to\infty} f(x)-\lim_{x\to-\infty} f(x)$$

vanishes; in the sequel we assume this is indeed the case.

The above approximation for $\mathbb{E}S$ intuitively makes sense. First, it shows that if the 'interval' Δ is small, then the mean number of singletons equals the number of realizations k. When Δ grows, there will be more anonymity, as reflected by the fact that $\mathbb{E}S$ decreases; apparently it does so more or less linearly in Δ, with proportionality constant

$$\lambda_k := k(k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x.$$

Such an approximation can be made arbitrarily precise. If we wish to compute the Δ^2 term (that is, a quadratic approximation, in Δ, of $\mathbb{E}S$), we first write (assuming $f(\cdot)$ to have the desired differentiability properties)

$$F_i(\Delta)=\Delta\cdot f(i\Delta)+\frac{1}{2}\Delta^2\cdot f'(i\Delta)+\frac{1}{6}\Delta^3\cdot f''(i\Delta).$$

After considerable calculus we eventually find $\mathbb{E}S=k-\Delta\lambda_k+\Delta^2\bar{\lambda}_k$, with

$$\bar{\lambda}_k := \frac{1}{2}k(k-1)(k-2)\int_{-\infty}^{\infty} f^3(x)\mathrm{d}x-k(k-1)\int_{-\infty}^{\infty} f(x)f'(x)\mathrm{d}x+\frac{k}{6}\int_{-\infty}^{\infty} f''(x)\mathrm{d}x,$$

where it is noticed that integration by parts yields

$$\int_{-\infty}^{\infty} f(x)f'(x)\mathrm{d}x=\frac{1}{2}\left(\lim_{x\to\infty} f^2(x)-\lim_{x\to-\infty} f^2(x)\right).$$

For various distributions, the second and third term vanish, so that we get, as $\Delta \downarrow 0$,

$$\mathbb{E}S = k - \Delta\left(k(k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x\right) + \Delta^2\left(\frac{1}{2}k(k-1)(k-2)\int_{-\infty}^{\infty} f^3(x)\mathrm{d}x\right) + O(\Delta^3).$$

Analogous computations yield

$$\begin{aligned}
\phi_1 &= 1 - \Delta\left((k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x\right) + \Delta^2\left(\frac{1}{2}(k-1)(k-2)\int_{-\infty}^{\infty} f^3(x)\mathrm{d}x\right) + O(\Delta^3),\\
\phi_2 &= \Delta\left((k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x\right) - \Delta^2\left((k-1)(k-2)\int_{-\infty}^{\infty} f^3(x)\mathrm{d}x\right) + O(\Delta^3),\\
\phi_3 &= \Delta^2\left(\frac{1}{2}(k-1)(k-2)\int_{-\infty}^{\infty} f^3(x)\mathrm{d}x\right) + O(\Delta^3);
\end{aligned}$$

in addition we have that $\phi_j = o(\Delta^2)$ for $j = 4, 5, \ldots$.

For the special case that X corresponds to a Normal distribution, the above approximations can be explicitly evaluated. The following lemma is useful. It can be proven by noting that, up to a multiplicative constant, $f^m(\cdot)$ is again a density, $m \in \mathbb{N}$.

Lemma 6.1 *Let X have a normal distribution with mean μ and variance σ^2. Then, with $m \in \mathbb{N}$,*

$$\int_{-\infty}^{\infty} f^m(x)\,dx = \frac{1}{\sqrt{m}}\frac{1}{(\sqrt{2\pi}\sigma)^{m-1}}.$$

Observe that these integrals do not involve μ, as could be expected. Inserting them into our expansion, we thus arrive at an approximation of $\mathbb{E}S$ for X stemming from the Normal distribution:

$$\mathbb{E}S = k - \Delta\,\frac{k(k-1)}{2\sigma\sqrt{\pi}} + \Delta^2\frac{k(k-1)(k-2)}{4\sqrt{3}\sigma^2\pi} + O(\Delta^3).$$

We see that the larger the variance σ^2, the higher the number of singletons, as could have been expected on intuitive grounds; the above relation quantifies this effect.

Remark 6.2 Similar formulas can be derived for the variance of S. Write, as before, $S = \sum_i S_i$, where the random variable S_i equals 1 if there is a singleton in the interval $[i\Delta, (i+1)\Delta))$ and 0 else. It is a standard rule in probability theory that

$$\mathbb{V}\mathrm{ar}\, S = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} \mathbb{C}\mathrm{ov}(S_i, S_j).$$

First observe that

$$\begin{aligned}\sum_{i=-\infty}^{\infty} \mathbb{C}\mathrm{ov}(S_i,S_i) &= \sum_{i=-\infty}^{\infty} \mathbb{V}\mathrm{ar}\, S_i = \sum_{i=-\infty}^{\infty}\Big(\mathbb{E}S_i - (\mathbb{E}S_i)^2\Big)\\ &= k(1-F_i(\Delta))^{k-1}F_i(\Delta) - k^2(1-F_i(\Delta))^{2k-2}(F_i(\Delta))^2\\ &= k - \Delta\, k(k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x - \Delta\, k^2\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x + O(\Delta^2).\end{aligned}$$

It also holds that

$$\sum_{i\neq j}\mathbb{C}\mathrm{ov}(S_i,S_j) = \sum_{i\neq j}\left(\mathbb{E}(S_iS_j) - (\mathbb{E}S_i)(\mathbb{E}S_j)\right),$$

where

$$\begin{aligned}\mathbb{E}(S_iS_j) &= k(k-1)(1-F_i(\Delta)-F_j(\Delta))^{k-2}F_i(\Delta)F_j(\Delta),\\ (\mathbb{E}S_i)(\mathbb{E}S_j) &= k^2(1-F_i(\Delta))^{k-1}(1-F_j(\Delta))^{k-1}F_i(\Delta)F_j(\Delta).\end{aligned}$$

Elementary manipulations now yield that

$$\begin{aligned}\sum_{i\neq j}\mathbb{E}(S_iS_j) &= k(k-1) - 2\Delta\, k(k-1)(k-2)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x + O(\Delta^2),\\ \sum_{i\neq j}(\mathbb{E}S_i\times\mathbb{E}S_j) &= k^2 - 2\Delta\, k^2(k-1)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x + O(\Delta^2).\end{aligned}$$

We eventually find

$$\mathbb{V}\mathrm{ar}\, S = \Delta(2k^2-3k)\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x + O(\Delta^2).$$

We conclude that $\mathbb{V}\mathrm{ar}\, S$ grows essentially linear in Δ, for Δ small. As $\Delta\downarrow 0$, we have that $\mathbb{V}\mathrm{ar}\, S\to 0$, as could be expected from the fact that S approaches k. Formulas for higher moments can be derived in an analogous fashion.

6.3.2 Experiments

In our experiments, we work with the following two data sets:

- One data set containing 25,000 records of human heights and weights [81], obtained in 1993 by a growth survey of 25,000 children from birth to 18 years of age;

- One data set containing all 766,000 birthdays of citizens from the municipality of Amsterdam.

QQ-plots reveal that weight and height in the first data sets are accurately approximated by the Normal distribution; for weight, the estimated standard deviation is 5.289 kg; for height it is 4.830 cm. Also, the birthdays in the second data set are nearly uniformly distributed over the 365 days of the year (leap years are ignored).

We sampled 10,000 times k persons from both data sets for height, length and birthday. Next, we estimated the mean number of singletons $\mathbb{E}S$ in these groups of size k, for different granularities Δ. In the tables below these estimates are in roman, and the corresponding approximations in italics. For weight and height, these approximations are based on the Normal distribution; more specifically, the $O(\Delta)$-approximation is

$$\mathbb{E}S \approx k - \Delta\,\frac{k(k-1)}{2\sigma\sqrt{\pi}}$$

and the $O(\Delta^2)$-approximation

$$\mathbb{E}S \approx k - \Delta\,\frac{k(k-1)}{2\sigma\sqrt{\pi}} + \Delta^2\,\frac{k(k-1)(k-2)}{4\sqrt{3}\sigma^2\pi};$$

for birthdays we use the counterparts of these formulae based on the uniform distribution, as given through (6.3).

The main conclusions from the tables are the following. (i) The approximations are highly accurate for relatively small Δ and k. Its performance degrades for larger Δ and k, but for quite a large set of parameters the fit remains reasonable. (ii) The $O(\Delta^2)$-approximation performs substantially better than the $O(\Delta)$-approximation (where it is noted that, obviously, adding an $O(\Delta^3)$-term would improve the approximation even more).

6.4 Multivariate distributions

The previous Section considered identifiability in the case where one reveals a specific single-dimensional attribute. In this Section, we study the case of multidimensional data. Consider a questionnaire in which one is asked to reveal their weight, but in addition also height. It is clear that there is a positive correlation between weight and height, and the question that arises is to what extent this correlation affects the identifiability, measured in terms of the mean number of singletons.

One would expect that the stronger the correlation between the two variates, the less information the second variate adds to the first variate, thus less increasing identifiability in terms of the number of singletons. The main finding of this Section is that this intuition indeed holds; in the special case the two variates stem from a two-dimensional Normal distribution, we derive explicit formulas that quantify this effect. The formulas are tested using real data.

6.4.1 Theoretical results

In this Section we consider the case of (X, Y) having a bivariate Normal distribution; the joint density $f(x, y)$ is given by

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}}\exp\left(-\frac{1}{2(1-\varrho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}-\frac{2\varrho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right).$$

Here μ_X and μ_Y are the means of X and Y, respectively, σ_X^2 and σ_Y^2 are the corresponding variances, and ϱ is the correlation between X and Y (whose effect we study in this Section), that is, $\mathbb{C}ov(X, Y) = \varrho\,\sigma_X\sigma_Y$.

In our experiments, the intervals for both coordinates are given by Δ_X and Δ_Y, respectively. Relying on

$$\begin{aligned} F_{i,j}(\Delta_X, \Delta_Y) &:= \mathbb{P}(X \in [i\Delta_X, (i+1)\Delta_X), Y \in [j\Delta_Y, (j+1)\Delta_Y) \\ &= \Delta_X\Delta_Y \cdot f(i\Delta_X, j\Delta_Y) + G(\Delta_X, \Delta_Y), \end{aligned}$$

where $G(\Delta_X, \Delta_Y)$ contains higher order terms, we obtain, in precisely the same way as in the single-dimensional case (see Section 6.3)

$$\mathbb{E}S = k - \Delta_X\Delta_Y \cdot k(k-1)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^2(x, y)\mathrm{d}x\mathrm{d}y + O((\Delta_X\Delta_Y)^2).$$

Using the following lemma, the double integral can be evaluated explicitly. Its proof is very similar to that of Lemma 6.1.

Lemma 6.3 *Let (X, Y) have a bivariate normal distribution with means (μ_X, μ_Y), variances (σ_X^2, σ_Y^2) and correlation ϱ. Then*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^m(x, y)\,dx\,dy = \frac{1}{m}\frac{1}{(2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2})^{m-1}}.$$

We thus obtain the following approximation:

$$\mathbb{E}S = k - \Delta_X\Delta_Y \cdot k(k-1)\frac{1}{4\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}} + O((\Delta_X\Delta_Y)^2).$$

As before, we observe that the larger the variances σ_X^2 and σ_Y^2, the higher the number of singletons. In addition, the formula shows that the more the variates X and Y are correlated (that is, the closer ϱ, in absolute value, is to 1), the lower the number of singletons. This is consistent with our intuition: a combination of two correlated variates can be less identifying than a combination of two non-correlated variates. If the correlation is 0, then no information on Y is captured in X, and as a result the mean number of singletons is relatively high.

6.4.2 Experiments

We again work with the data set containing 25,000 records of human heights and weights available from [81]. Estimation of the (Pearson-)correlation between height and length yields $\varrho = 0.5028$. As before, we sampled 10,000 times k people from the data set of 25,000 people, who now have to reveal both weight and height, and we count the number of unique samples.The intervals Δ_W for weight and Δ_H for height are varied, as indicated in the caption below Fig. 6.4.

The graphs of Fig. 6.4 show that the approximation works excellently for small intervals Δ_W and Δ_H, and k relatively small (so that, as a consequence, $\mathbb{E}S$ is close to k), and still fine for moderate values of the intervals and k. Evidently, the fit can be improved by adding the $(\Delta_W \Delta_H)^2$-term.

In Fig. 6.5 we keep (in the left panel) Δ_H fixed (at 1 cm) and vary Δ_W, and (in the right panel) we keep Δ_W fixed (at 1 kg) and vary Δ_H. As expected from Section 6.3, the approximation matches the simulation-based estimates well for small Δ_W (left panel) and small Δ_H (right panel). In these experiments we chose $k = 10$.

6.5 Discussion

This Chapter focused on probabilistic analysis of the number of singletons. The contribution of this Chapter is threefold: we address the effect of non-uniformity, quantify the effect of aggregation and assess the impact of correlation between variates.

Regarding the first issue, we have empirically validated approximations that we developed in Chapter 5; it was concluded that our technique to estimate the mean number of singletons, doubletons, tripletons, etc. yields reliable estimates. In our experiments, we estimate the Kullback-Leibler (KL) distance by using data from the entire population, and then approximate the mean number of singletons (that is, unambiguously identifiable individuals) among k people sharing the same postal code. The fit of the approximations can probably improved by not estimating the KL distance based on the entire population, but just on the part of the city the specific postal code is in.

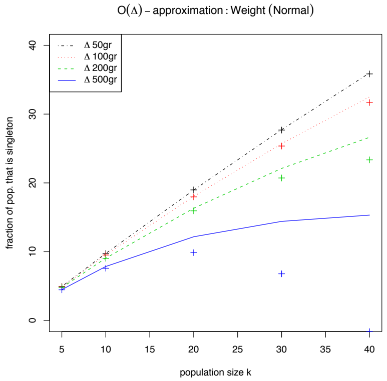
Regarding the second issue, impact of the interval Δ, we showed that the mean number of singletons S can be accurately approximated by polynomial in Δ; the linear approximation is $\mathbb{E}S = k - \lambda_k\Delta$.

Also, the accuracy of these approximations decreases for events of low probability; in our framework it remains an open question how those should be handled. Depending on the practical context, a questionnaire maker could decide not to ask respondents to reveal their precise age if it is higher than, for example, 79 — allowing the respondent to skip the question or check "79 or higher".

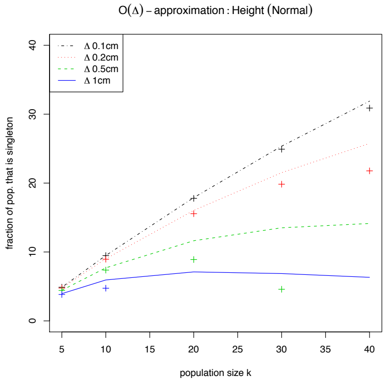
Regarding the third issue, we extend the setting of the second issue, that was a single non-categorical variable, to multiple non-categorical variables. We show explicitly the effect of the correlation between the variates. As can be intuitively understood, the higher the correlation, the higher the privacy level. Our analysis does not cover the impact of correlation between categorical data, or correlation between categorical and non-categorical data; think of for instance gender and income, or civil status and age.

The accuracy of the latter two approximations can be made arbitrarily high by adding more terms of the polynomial expansion. The formula for the mean number of singletons allows various easy estimates. Suppose, for instance, that X corresponds to weight rounded to multiples of 500 grams, and for $k = 10$ we observe that the mean number of singletons is about 8. Then a small computation tells us that σ is about 6.6. Doubling Δ (to multiples of 1 kilogram) increases the anonymity, in that the mean number of singletons will be reduced to roughly 6; halving Δ leads to $\mathbb{E}S$ equalling roughly 9. We propose that this can be used as a (rough) rule of thumb.

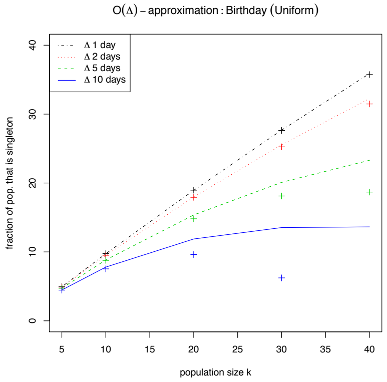
Chapter 7 will discuss applications of the theory developed here.



k	0.05kg	0.1kg	0.2kg	0.5kg	1.0kg	2.0kg
5	4.95	4.90	4.79	4.49	4.02	3.25
	4.95	*4.89*	*4.79*	*4.47*	*3.93*	*2.87*
	4.94	*4.89*	*4.79*	*4.49*	*4.03*	*3.26*
10	9.76	9.54	9.07	7.86	6.21	3.94
	9.76	*9.52*	*9.04*	*7.60*	*5.20*	*0.40*
	9.76	*9.53*	*9.09*	*7.90*	*6.38*	*5.13*
20	18.99	18.06	16.33	12.16	7.70	3.62
	18.99	*17.97*	*15.95*	*9.87*	*-0.27*	*- 20.53*
	19.01	*18.09*	*16.39*	*12.68*	*10.97*	*24.40*
30	27.75	25.72	22.10	14.39	7.63	3.09
	27.68	*25.36*	*20.72*	*6.80*	*- 16.40*	*-62.80*
	27.78	*25.76*	*22.32*	*16.80*	*23.61*	*97.24*
40	35.98	32.49	26.63	15.28	7.14	2.81
	35.84	*31.68*	*23.36*	*- 1.60*	*- 43.20*	*- 126.40*
	36.08	*32.65*	*27.25*	*22.74*	*54.17*	*263.07*



k	0.1cm	0.2cm	0.5cm	1.0cm	2.0cm	5.0cm
5	4.88	4.76	4.43	3.94	3.10	1.54
	4.88	*4.77*	*4.42*	*3.83*	*2.66*	*- 0.84*
	4.88	*4.77*	*4.45*	*3.95*	*3.14*	*2.11*
10	9.48	8.97	7.68	5.98	3.68	1.21
	9.47	*8.95*	*7.37*	*4.74*	*-0.51*	*-16.28*
	9.49	*9.01*	*7.73*	*6.16*	*5.16*	*19.17*
20	17.90	16.05	11.72	7.151	3.17	0.90
	17.78	*15.56*	*8.90*	*- 2.19*	*-24.39*	*- 90.97*
	17.92	*16.10*	*12.27*	*11.28*	*29.50*	*245.82*
30	25.35	21.53	13.50	6.90	2.70	0.83
	24.92	*19.84*	*4.59*	*- 20.81*	*- 71.62*	*-224.05*
	25.40	*21.76*	*16.59*	*27.17*	*120.29*	*975.38*
40	31.94	25.67	14.13	6.37	2.42	0.82
	30.89	*21.78*	*- 5.55*	*- 51.11*	*- 142.22*	*- 415.55*
	32.06	*26.45*	*23.63*	*65.64*	*324.80*	*2503.29*



k	1 day	2 days	5 days	10 days	20 days	30 days
5	4.94	4.89	4.73	4.46	3.97	3.55
	4.95	*4.89*	*4.73*	*4.45*	*3.91*	*3.36*
	4.95	*4.89*	*4.73*	*4.48*	*4.00*	*3.56*
10	9.73	9.50	8.84	7.80	6.05	4.67
	9.75	*9.51*	*8.77*	*7.54*	*5.08*	*2.62*
	9.76	*9.52*	*8.84*	*7.81*	*6.16*	*5.04*
20	18.94	17.97	15.37	11.88	6.97	4.05
	18.96	*17.92*	*14.81*	*9.62*	*-0.77*	*-11.15*
	18.99	*18.03*	*15.45*	*12.17*	*9.45*	*11.83*
30	27.68	25.55	20.09	13.53	6.03	2.74
	27.62	*25.25*	*18.11*	*6.23*	*-17.54*	*-41.31*
	27.71	*25.61*	*20.39*	*15.32*	*18.83*	*40.52*
40	35.89	32.25	23.30	14.63	4.71	1.71
	35.74	*31.48*	*18.69*	*-2.62*	*-45.25*	*-87.87*
	35.96	*32.36*	*24.22*	*19.50*	*43.26*	*111.27*

Figure 6.3: Graphical illustration of accuracy of the $O(\Delta)$-approximation; $\mathbb{E}S$ as a function of k for height, weight and birthday. The lines correspond to the estimates resulting from simulation, and the '+' with the $O(\Delta)$-approximation. Tables show mean number of singletons for various values of k.

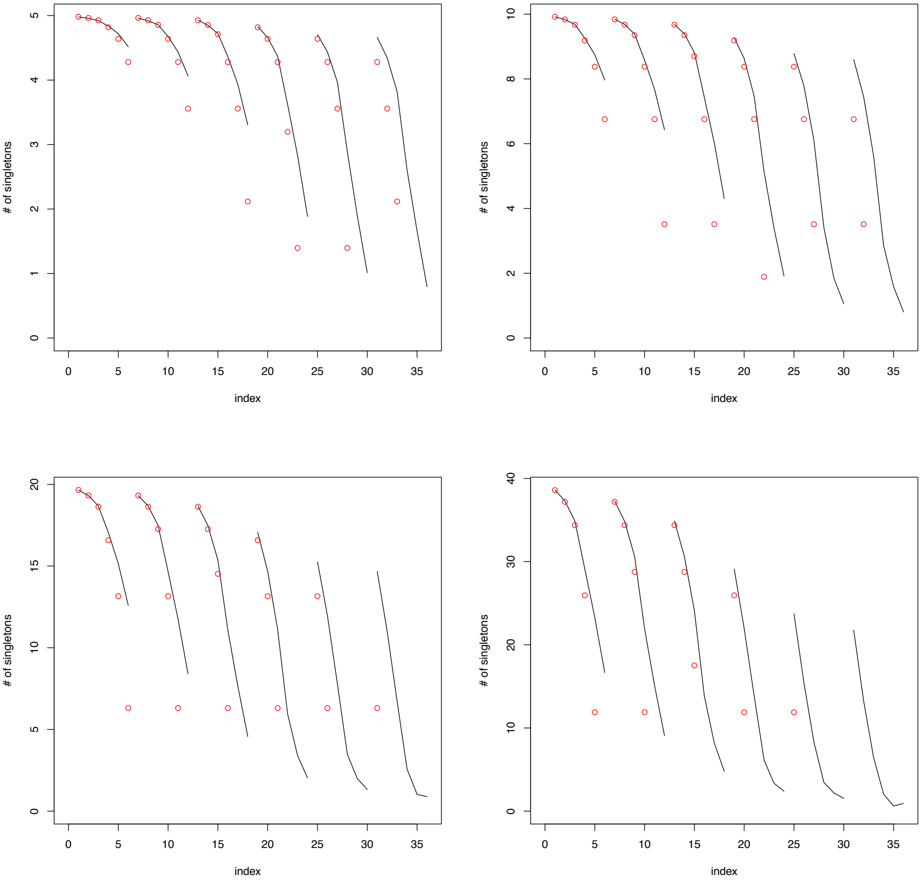


Figure 6.4: Expected number of singletons, for $k = 5, 10, 20, 40$, respectively ($k = 30$ is skipped due to page layout). The solid lines are the simulation-based estimates, the dots are the approximations based on the formulas derived in this Section. Per picture, the first 6 data points correspond to $\Delta_H = 0.5$ cm, the second 6 data points to $\Delta_H = 1.0$ cm, the third set of 6 data points to $\Delta_H = 2.0$ cm, the fourth set of 6 data points to $\Delta_H = 5.0$ cm, the fifth set of 6 data points to $\Delta_H = 10.0$ cm, and the last set of 6 data points to $\Delta_H = 20.0$ cm. Within each group of 6 data points, these correspond to $\Delta_W = 0.5, 1.0, 2.0, 5.0, 10, 20$ kg.

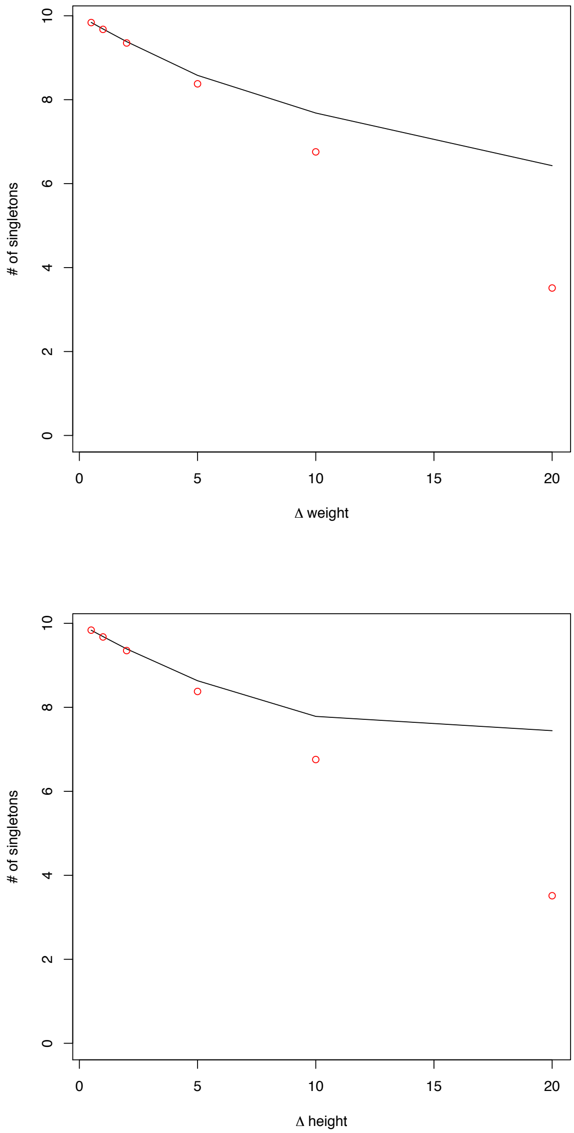


Figure 6.5: Left panel: effect of Δ_W for Δ_H fixed; right panel: effect of Δ_H for Δ_W fixed.

7 Practical applications

In this Chapter we will share preliminary ideas on applying in real life the techniques developed in this thesis. We provide a conceptual framework for the application of the distribution-informed prediction of anonymity properties via Kullback-Leibler distances (KL-distances) as developed in Chapter 4 and Chapter 5. The KL-based predictions can be applied to quasi-identifiers consisting of any combination of numerical variables (e.g. { *age + height* }) and categorical variables (e.g. { *gender* }).

In addition, we discuss application of the techniques developed in Chapter 5 and Chapter 6, which only apply to numerical variables, such as the analysis of the effect of interval-widths on identifiability (see Chapter 6). The latter enables pollsters, for example, to protect the anonymity of respondents by deciding *beforehand*, based on quantifications, whether to ask respondents to reveal, say, their exact age or rather the age *group* to which they belong — instead of collecting exact ages and having respondents trust their unknown pollster that she will make the data less precise *afterwards*. Quantifications remove some of the pollster's uncertainty that may otherwise have led the pollster to choose an overly wide interval (while it may be beneficial for the analysis to have more specific information), or perhaps to simply ignore the issue and ask for exact data that puts the respondent at risk (or at least, leave them with a feeling of unease).

The remainder of this Chapter is organized as follows: Section 7.1 will introduce our preliminary model; Section 7.2 will discuss ‘non-functional’ aspects crucial to real-life application of the model; Section 7.3 will describe steps to take toward implementing a real-life application; Sections 7.4 and 7.5 will discuss various practical aspects that need to be taken into account, including the limitations of our work; and Section 7.6 will conclude this Chapter. For further inspiration we refer to the example analysis of anonymity in Appendix B, that considers a questionnaire observed in real life.

Remark 7.1 *Measuring unidentifiability is measuring identifiability. Our techniques are intended for privacy protection but can be used directly for purposes of identifiability as well, such as in marketing and forensics. Our perspective, however, is that of privacy protection.*

7.1 Preliminary model

We now introduce a preliminary conceptual framework for applying distribution-informed prediction techniques. Figure 7.1 shows the framework, distinguishing a *repository* (that stores Kullback-Leibler distances), *policy* (decisions about what data (not) to disclose, collect and share), *data holder*(s) (anyone with access to personal data) and *policy maker*(s) (anyone deciding about the processing of personal data, notably including the subjects themselves). We distinguish four tasks, chronologically ordered: *publish*, *query*, *analyze* and *decide*. These will be explained below.

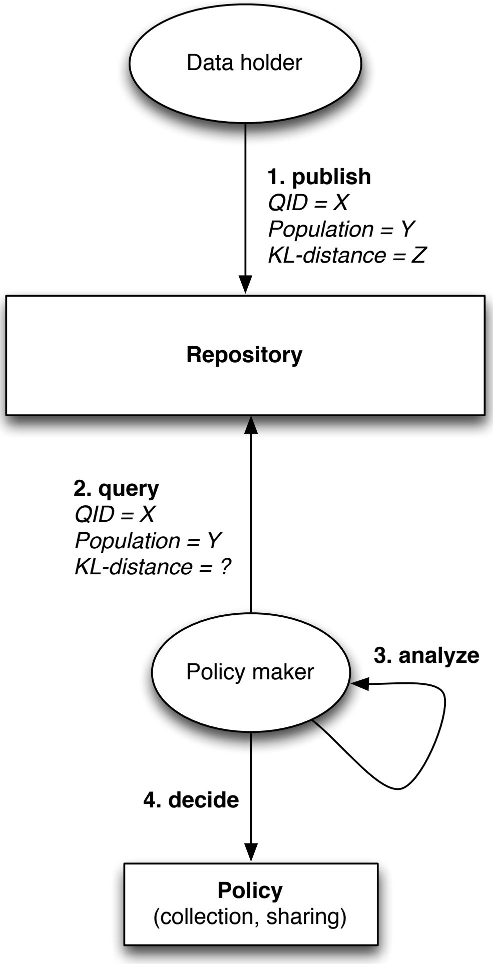


Figure 7.1: Preliminary model for applying distribution-informed privacy predictions as part of privacy policy making.

7.1.1 Data holder

The *data holder* collects and stores personal information about individuals. Although entities that are *legally* assigned the role *data controller* or *data processor* can act as data holder, not only they can. For example, an individual can be data holder of his/her own personal data. To prevent confusion with the legal domain we use the label "data holder", consistent with Solove [74] (see Section 1.1). In our model, anyone having access to a collection of personal data can act as data holder.

The disclosure of information by a person to a data holder establishes a context conform Nissenbaum [61], including context-relative informational norms, compliancy to which constitutes contextual integrity; i.e., that the disclosed information does not end up in a situation where presence of it constitutes a privacy violation (as perceived by data subjects or wider society, but not necessarily made explicit in laws; also, note that translating implicit, subjective and changing contextual roles into a disclosure policy is non-trivial). The publication of a statistic about a population of which that person is part is unlikely to violate privacy law. In the Netherlands, for example, privacy law only applies to the processing of data that can be traced to individuals without considerable effort. The practical meaning of 'considerable effort' remains unclear to us; from a privacy perspective we hope it means no less than 'disproportionate to potential gain'. Such publication might, however, violate a context-relative informational norm, for example when the person does not agree with their data being part of a openly published statistic (such as in our model). Additional work is needed to assess the moral and legal risk in openly publishing statistics computed from existing collections of personal data.

Task:

- **publish**: submit to repository one or more Kullback-Leibler distances, accompanied by specification of the QID and population.

7.1.2 Policy maker

Policy maker assesses privacy risk and decides what data (not) to collect and what data (not) to share. The decision is influenced by legal norms and, if data holder and policy maker are the same entity, the context-relative informational norms between data holder and the persons about whom the data holder stores data. As a special case, policy maker can be a self-assessing individual that wants to decide what (combined) information not to disclose during, for example, an anonymous questionnaire.

Task:

- **query**: request from repository the Kullback-Leibler distance (*KL-distance*), given a specification of the QID and population;
- **analyze**: apply our methodology to analyze QIDs;
- **decide**: decide what data (not) to collect and what data (not) to share.

Presumably, these tasks will be part of a more comprehensive privacy risk management process that takes into account existing information collection and sharing that might influence the privacy risk involved in the per-instance context to which our methodology is most relevant.

7.1.3 Repository

The repository is a publication facility that allows the data holder to submit KL-distances, and the policy maker to query KL-distances. The repository stores three-tuples consisting of a description of the QID, a description of the population and the KL-distance. While KL-distance is a number, the description of QID and population will be less trivial. For clarity of exposition, **we will assume that policy maker and data holder use the same ontologies and data structures**; i.e., they have a shared vocabulary. Under that assumption, the QID and population can be defined in terms of that vocabulary. Consider a shared ontological concept and data structure *Citizen* that contains, among others, the attributes *PostalCode*, *Gender*, and *BirthYear*; second, that population can be specified in terms of *City*; and third, that data holder stores this data for *all* citizens of Amsterdam. To publish the KL-distance that applies to citizens of Amsterdam regarding *QID* = {*PostalCode, Gender, BirthYear*}, data holder first computes the KL-distance and sends to the repository the following message:

```
QID =   {PostalCode, Gender,  BirthYear}
population = {City=Amsterdam}
KL-distance = ...
```

Data holder may also publish KL-distances for less specific QIDs. For example, leaving out *Gender*:

```
QID =   {PostalCode, BirthYear}
population = {City=Amsterdam}
KL-distance = ...
```

In addition, data holder may also publish KL-distances for subpopulations. For example, only including persons that own a car:

```
QID =   {PostalCode, BirthYear}
population = {City=Amsterdam, CarOwner=yes}
KL-distance = ...
```

The use and applicability of various QIDs and various subpopulations will depend on the (existence of) information collection and sharing to which the involved persons are exposed.

7.2 Issues

7.2.1 Assess privacy risk of KL-repository itself

The purpose of our model is privacy protection, but possibly, the model poses privacy risk *itself.* At this point, we do not know if or how a public repository of KL-distances might be abused. As we noted, measuring unidentifiability is equivalent to measuring identifiability, and our techniques might be applied for purposes of identifiability rather than *un*identifiability. We expect that smallness of populations, as determined by the amount of information in the QID, will be a key issue in deciding what KL-distances (not) to publish. A tradeoff exists between accuracy of prediction (more specific population = more accurate prediction) and protecting against use for identifiability (less specific population = less usable for identifiability).

7.2.2 Disputes

In a perfect world, there are no errors in the data from which KL-distances are computed, and the data covers complete populations. In real life, errors do occur and coverage is often incomplete. The mileage will vary between different data holders. The Dutch municipal registry offices, for example, will tend to cover the complete population within their municipality, while a corporate data set might only cover the consumers within that population; and not only within the municipality where they are located, but consumers from any location.

In the occasion of multiple data holders submitting different KL-distances for the same QID and population, a decision must be made which information to use. We consider this beyond the scope of our thesis.

7.2.3 Incentives

For policy makers, the legal obligation to comply with privacy law may be incentive to publish KL-distances: e.g., if the policy maker wants to legally avoid collecting personal data, then, under Dutch privacy law, the data collected must not be traceable to individuals without effort that is disproportionate to the risk associated with such disclosure. The policy maker may be motivated to apply our methods In order to know to what extent data is traceable to individuals. For some policy makers, there may be a moral or marketing-inspired desire to comply with context-relative informational norms. For the special case where an individual person maps to the policy maker actor, the desire to know what (combined) information is (quasi-)identifying to a certain extent will be sufficient incentive.

7.3 What steps to take next

The following subsections describe the steps that need to be taken next to apply our model in real life, after the issues described above have been (sufficiently) resolved.

7.3.1 Make an inventory of data holders and their data

An inventory is needed of data holders and their data. Specifically, for each relevant data set, a list of columns and description of the population about which data is present need to be established. In the Netherlands, a potential starting point is the Dutch Data Protection Agency (CBP), that maintains a registry of data protection officers and a registry of (reported) processing of personal data. Other pointers for the Dutch can be found in the report ‘Onze digitale schaduw’ (2009) [70] commissioned by the Dutch Data Protection Agency and in the report ‘iOverheid’ (2011) established by the Dutch Scientific Council for Government Policy (WRR) [27]. For the UK, pointers can be found in the report ‘Database State’ (2009) [3] commissioned by the Joseph Rowntree Reform Trust.

Then, a second inventory is needed: a list of the (combined) information that pollsters ask for during anonymous questionnaires (whether online or offline), and the information that is shared in contexts of science and policy research. In the Netherlands, both Statistics Netherlands and the KNAW/NWO Data Archiving and Networked Services (DANS) institute[1] may provide pointers. This second inventory can be jump-started through a simple brainstorm process.

By matching both inventories, and taking into account privacy risks of the KL-repository itself (see Section 7.2.1) and the desired scope of the repository, it needs to be decided which QIDs and populations to include/exclude. Theoretically, the scope of the repository could be unlimited: one could attempt to establish a single nation-wide or even global repository that contains KL-distances for every possible QID and every possible population. Practically, the scope is limited to data holders and policy makers that (are able to) share a data vocabulary (see Section 7.1.3) and are also willing to participate. It is

[1] Website: `http://www.dans.knaw.nl/`

probably sensible to limit a first attempt at a repository to common QIDs and common populations; both of which can be established via the inventories and common sense.

7.3.2 Build software tools

Information technology will need to be built. First, a repository needs to be set up. Second, software for KL-analysis and publication to the repository needs to be engineered and distributed to data holders (tools for computing KL-distances from data stored in MySQL, MSSQL, etc.). Third, software for performing distribution-informed analysis is needed (possibly local, possibly remote). Fourth, the system needs to be maintained, and policy maker and data holder should be able to get support when needed.

7.4 Other aspects

Apart from above considerations, several other aspects need to be taken into account. Two factors that play an important role in quasi-identifier analysis are the granularity, or interval width, of variables in a quasi-identifier; and the correlation between variables in a quasi-identifier. Regarding the former, obviously the more fine-grained the data is, the more identifying a quasi-identifier will tend to be. The methodology of Section 6.3 can be applied to quantify this effect. Regarding the latter, the technique developed in Section 6.4 can be used to predict identifiability in case there is substantial correlation between multiple non-categorical, numerical variables within the quasi-identifier. In principle the same KL-based technique can be used as in the single-variate case, as long as the correlation between the variables is taken care of adequately as demonstrated in Section 6.4.

Lastly, the repository needs to be protected from misinformation (e.g. unintentionally incorrect KL-distances being submitted) and disinformation (intentionally incorrect KL-distances, e.g. to make privacy risk appear less than it really is). We consider this to be outside the scope of our thesis, but emphasize that it must be addressed when working toward a real-life implementation.

7.5 Limitations and future work

Our distribution-informed prediction techniques require that Kullback-Leibler distances can be computed between the Uniform distribution and the actual distribution. To know the actual distribution, access is required to (personal) data. That data must be representative, in terms of the quasi-identifier variables under consideration, for the population for which the quasi-identifier analysis is performed. As mentioned in Section 7.2.2, the data ideally provides full coverage of that population, and has few errors. If it is not possible to get access to that data, or the data contains too many errors in the variables present in the quasi-identifier, the distribution-informed techniques cannot be applied. Also, note that while an individual may use these techniques to get an *on-the-average estimate* of identifiability of members of the population to which he/she belongs, that individual may self have outlier values and be more identifiable than the estimate suggests. In a population where nearly everyone has blue eyes or brown eyes, disclosing that one has green eyes is obviously more revealing than on the average.

Furthermore, in the analysis of singletons outlined in Chapter 5, notably Figure 5.1, it is observed that our approximations become inaccurate in presence of strong outliers. In our example of age distributions, our approximations showed accurate for the range 0-79; we were unable to obtain accurate results when including ages above 79. Clearly, this implies that our methodology is insufficient by itself for performing exhaustive privacy analysis. We propose that other methods and techniques that *do* sufficiently take outliers into account are applied together with ours.

In Chapter 6, note that the $O(\delta)$ approximation techniques for determining the effects of interval width, as Figure 6.3 shows for height, width and birthday, become less accurate when predicting outcomes for large interval widths. In addition, the techniques we proposed for taking into account the effects of correlation between variables have only been examined for the setting where the variables have a bivariate normal distribution. These aspects must be considered when applying these techniques in practice.

7.6 Conclusion

Although distribution-informed prediction is not the only method we developed throughout Chapter 4, 5 and 6, it is our most innovative result. In this Chapter, we primarily focused on that result and share preliminary ideas about how to apply it in practice. Additional work is needed: first, the privacy risk of a public repository of Kullback-Leibler distances computed from sets of personal data needs to be assessed. Second, inventory needs to be made of (candidate) data holders, and of the QIDs and populations that are most relevant to be subjected to privacy-analysis. We provided pointers to information sources that we believe are useful during these activities.

8 Conclusions and future work

In our increasingly computer-networked world, more and more personal data is collected, linked and shared. This raises questions about privacy — i.e. about the feeling and reality of enjoying a private life in terms of being able to exercise control over the disclosure of information about oneself. In attempt to provide privacy, databases containing personal data are sometimes de-identified, meaning that obvious identifiers such as Social Security Numbers, names, addresses and phone numbers are removed. In microdata, where each record maps to a single individual, de-identification might however leave variables that, combined, can be used to re-identify the de-identified data.

To establish the case for quantified privacy analysis, we first performed an empirical study on the identifiability of nameless hospital intake data and welfare fraud data about Dutch citizens, using large amounts of personal data collected from municipal registry offices. We showed, through quantifications, the possibility of large differences in actual privacy of citizens depending on the municipality where they live.

We developed a range of novel techniques for predicting aspects of anonymity, building on probability theory, and specifically birthday problem theory and large deviations theory. We empirically validated our formulas using public data insofar possible, and using our privately collected data insofar necessary to ensure coherence of research.

In the final Chapter we gave preliminary ideas for applying our techniques in real life. We feel these are suitable and useful input to the privacy debate; practical application will depend on competence and willingness of data holders and policy makers to correctly identify quasi-identifiers. In the end, it remains a matter of policy what value of k can be considered *sufficiently strong* anonymity for particular personal information.

We propose three directions for future research:

- Our formulas may have uses outside the context of data anonymity, such as in the context of communication anonymity. KL-distance based prediction, for example, might show to be useful in contexts handling distributions related to aspects of packets or network flows that are relevant to anonymity of communication. We do not know whether this is the case for onion routing (e.g. Tor), garlic routing, Crowds, MUTE, I2P or any other existing system for anonymous communication. Possibly, our methods allow creation of a new system, or have a function under environmental assumptions different from those under which existing systems are designed, operated and used;

- Our formulas may have uses outside the context of privacy altogether: notably, forensics and marketing. In forensics, for example, the question might be raised how probable it is that some piece of evidence is unique to a person. Similarly, a marketeer might wonder how probable it is that some piece of information is unique to a person. Especially the formulas developed in Chapter 4 and Chapter 5 may be relevant to those contexts. Whether this is true, and whether other parts of our work have application outside privacy, needs further research;

- Study is needed to show what sort of background information is easy to obtain, and what the impact is on re-identifiability. What possibilities do various types of adversaries — corporate, government, individual — have to obtain information? How does this vary between adversaries targeting specific individuals and adversaries targeting anyone who's data they are able to obtain?

We hope others will be inspired to build forth on our work, as we too built forth on the work of others.

A $\zeta(k, N)$ for $k = 1 \ldots, 50$ and $N = 1, \ldots, 20$

k/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.500	0.333	0.250	0.200	0.167	0.143	0.125	0.111	0.100	0.091	0.083	0.077	0.071	0.067	0.063	0.059	0.056	0.053	0.050
3	1.000	0.250	0.111	0.063	0.040	0.028	0.020	0.016	0.012	0.010	0.008	0.007	0.006	0.005	0.004	0.004	0.003	0.003	0.003	0.003
4	1.000	0.500	0.259	0.156	0.104	0.074	0.055	0.043	0.034	0.028	0.023	0.020	0.017	0.015	0.013	0.011	0.010	0.009	0.008	0.007
5	1.000	0.688	0.259	0.121	0.066	0.039	0.025	0.017	0.012	0.009	0.007	0.005	0.004	0.003	0.003	0.002	0.002	0.002	0.001	0.001
6	1.000	0.813	0.333	0.162	0.090	0.055	0.036	0.025	0.018	0.013	0.010	0.008	0.006	0.005	0.004	0.003	0.003	0.002	0.002	0.002
7	1.000	0.891	0.443	0.195	0.095	0.051	0.030	0.018	0.012	0.008	0.006	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001
8	1.000	0.938	0.557	0.240	0.114	0.060	0.034	0.021	0.013	0.009	0.006	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001
9	1.000	0.965	0.660	0.302	0.139	0.069	0.036	0.021	0.012	0.008	0.005	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000
10	1.000	0.980	0.744	0.378	0.171	0.081	0.042	0.023	0.013	0.008	0.005	0.003	0.002	0.002	0.001	0.001	0.001	0.000	0.000	0.000
11	1.000	0.989	0.811	0.460	0.211	0.099	0.049	0.026	0.014	0.008	0.005	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
12	1.000	0.994	0.862	0.541	0.261	0.121	0.058	0.029	0.016	0.009	0.005	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
13	1.000	0.997	0.900	0.617	0.319	0.148	0.070	0.035	0.018	0.010	0.006	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
14	1.000	0.998	0.928	0.684	0.382	0.182	0.085	0.041	0.021	0.011	0.006	0.004	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
15	1.000	0.999	0.949	0.742	0.446	0.222	0.104	0.050	0.025	0.013	0.007	0.004	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
16	1.000	1.000	0.963	0.792	0.510	0.267	0.127	0.060	0.029	0.015	0.008	0.004	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
17	1.000	1.000	0.974	0.833	0.572	0.317	0.155	0.073	0.035	0.017	0.009	0.005	0.003	0.001	0.001	0.001	0.000	0.000	0.000	0.000
18	1.000	1.000	0.982	0.866	0.629	0.369	0.187	0.089	0.043	0.021	0.010	0.005	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000
19	1.000	1.000	0.987	0.894	0.681	0.423	0.223	0.108	0.052	0.025	0.012	0.006	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000
20	1.000	1.000	0.991	0.916	0.727	0.476	0.263	0.131	0.063	0.030	0.015	0.007	0.004	0.002	0.001	0.001	0.000	0.000	0.000	0.000
21	1.000	1.000	0.994	0.934	0.768	0.528	0.306	0.157	0.076	0.036	0.018	0.009	0.004	0.002	0.001	0.001	0.000	0.000	0.000	0.000
22	1.000	1.000	0.996	0.948	0.804	0.578	0.350	0.186	0.092	0.044	0.021	0.010	0.005	0.003	0.001	0.001	0.000	0.000	0.000	0.000
23	1.000	1.000	0.997	0.959	0.835	0.625	0.396	0.218	0.110	0.053	0.026	0.013	0.006	0.003	0.002	0.001	0.000	0.000	0.000	0.000
24	1.000	1.000	0.998	0.968	0.861	0.668	0.442	0.253	0.131	0.064	0.031	0.015	0.007	0.004	0.002	0.001	0.001	0.000	0.000	0.000
25	1.000	1.000	0.999	0.975	0.884	0.707	0.487	0.291	0.155	0.077	0.038	0.018	0.009	0.004	0.002	0.001	0.001	0.000	0.000	0.000
26	1.000	1.000	0.999	0.980	0.903	0.743	0.531	0.329	0.181	0.092	0.045	0.022	0.011	0.005	0.003	0.001	0.001	0.000	0.000	0.000
27	1.000	1.000	0.999	0.985	0.919	0.776	0.574	0.369	0.210	0.110	0.054	0.026	0.013	0.006	0.003	0.002	0.001	0.000	0.000	0.000
28	1.000	1.000	1.000	0.988	0.933	0.804	0.614	0.409	0.241	0.129	0.065	0.032	0.015	0.008	0.004	0.002	0.001	0.000	0.000	0.000
29	1.000	1.000	1.000	0.991	0.944	0.830	0.651	0.449	0.274	0.151	0.077	0.038	0.019	0.009	0.004	0.002	0.001	0.001	0.000	0.000
30	1.000	1.000	1.000	0.993	0.954	0.853	0.686	0.489	0.308	0.174	0.092	0.046	0.022	0.011	0.005	0.003	0.001	0.001	0.000	0.000
31	1.000	1.000	1.000	0.994	0.962	0.872	0.719	0.527	0.343	0.200	0.107	0.055	0.027	0.013	0.006	0.003	0.002	0.001	0.000	0.000
32	1.000	1.000	1.000	0.996	0.968	0.890	0.748	0.564	0.378	0.227	0.125	0.065	0.032	0.016	0.008	0.004	0.002	0.001	0.000	0.000
33	1.000	1.000	1.000	0.997	0.974	0.905	0.775	0.599	0.414	0.256	0.145	0.076	0.039	0.019	0.009	0.005	0.002	0.001	0.001	0.000
34	1.000	1.000	1.000	0.997	0.978	0.918	0.800	0.633	0.449	0.286	0.166	0.089	0.046	0.023	0.011	0.005	0.003	0.001	0.001	0.000
35	1.000	1.000	1.000	0.998	0.982	0.930	0.822	0.664	0.484	0.317	0.189	0.104	0.054	0.027	0.013	0.007	0.003	0.002	0.001	0.000
36	1.000	1.000	1.000	0.998	0.985	0.940	0.842	0.694	0.518	0.349	0.213	0.120	0.064	0.032	0.016	0.008	0.004	0.002	0.001	0.000
37	1.000	1.000	1.000	0.999	0.988	0.948	0.860	0.722	0.551	0.381	0.239	0.138	0.074	0.038	0.019	0.009	0.005	0.002	0.001	0.001
38	1.000	1.000	1.000	0.999	0.990	0.956	0.877	0.747	0.583	0.413	0.266	0.157	0.086	0.045	0.023	0.011	0.006	0.003	0.001	0.001
39	1.000	1.000	1.000	0.999	0.992	0.962	0.891	0.771	0.614	0.445	0.293	0.178	0.100	0.053	0.027	0.013	0.007	0.003	0.002	0.001
40	1.000	1.000	1.000	0.999	0.993	0.967	0.904	0.793	0.643	0.476	0.322	0.199	0.114	0.062	0.032	0.016	0.008	0.004	0.002	0.001
41	1.000	1.000	1.000	1.000	0.995	0.972	0.915	0.813	0.670	0.507	0.351	0.222	0.130	0.072	0.038	0.019	0.010	0.005	0.002	0.001
42	1.000	1.000	1.000	1.000	0.996	0.976	0.925	0.831	0.696	0.537	0.380	0.246	0.148	0.083	0.044	0.023	0.011	0.006	0.003	0.001
43	1.000	1.000	1.000	1.000	0.996	0.980	0.934	0.848	0.721	0.566	0.409	0.271	0.166	0.095	0.051	0.027	0.013	0.007	0.003	0.002
44	1.000	1.000	1.000	1.000	0.997	0.983	0.942	0.863	0.743	0.594	0.438	0.297	0.186	0.108	0.060	0.031	0.016	0.008	0.004	0.002
45	1.000	1.000	1.000	1.000	0.998	0.985	0.949	0.877	0.765	0.621	0.466	0.323	0.206	0.123	0.069	0.037	0.019	0.010	0.005	0.002
46	1.000	1.000	1.000	1.000	0.998	0.987	0.956	0.890	0.785	0.647	0.494	0.349	0.228	0.138	0.079	0.043	0.022	0.011	0.006	0.003
47	1.000	1.000	1.000	1.000	0.998	0.989	0.961	0.901	0.803	0.671	0.522	0.376	0.250	0.155	0.090	0.050	0.026	0.013	0.007	0.003
48	1.000	1.000	1.000	1.000	0.999	0.991	0.966	0.911	0.820	0.695	0.549	0.402	0.273	0.173	0.102	0.057	0.031	0.016	0.008	0.004
49	1.000	1.000	1.000	1.000	0.999	0.992	0.970	0.921	0.836	0.717	0.575	0.429	0.297	0.191	0.115	0.066	0.036	0.019	0.009	0.005
50	1.000	1.000	1.000	1.000	0.999	0.993	0.974	0.929	0.850	0.738	0.600	0.455	0.321	0.211	0.129	0.075	0.041	0.022	0.011	0.006

Table A.1: $\zeta(k,N)$ for $k = 1\ldots,50$ and $N = 1,\ldots,20$.

B Example analysis: questionnaire

This Appendix discusses an internet-based questionnaire that was observed in real life and asks anonymous respondents to reveal various demographics. The questionnaire was held in June 2010 by the Concertgebouw (the famous concert hall in Amsterdam) and concerned non-sensitive topics. We use it here as a toy example. We will show what information respondents are asked to reveal, and analyze how anonymity decreases by each piece of information the respondent reveals.

Of course, if we were to assume that the pollster does not try to trace survey data to named individuals and that the survey data is not sold or compromised, this analysis would not be needed: there would simply not be any threat of identification to protect against. But we choose to assume, more diligently, that the pollster might try to trace survey data to named individuals, that the data might get sold and that the data might get compromised. Under those assumptions, analysis of anonymity is needed.

First, as shown in Figure B.1, the respondent is asked to reveal full postal code ('PC6' postal code: four digits and two letters), gender (choice between male and female), and Year of Birth ('YoB', four digits). Based on empirical data of 2,777,953 Dutch citizens obtained from 16 municipalities (see Chapter 3 and Chapter 6), Table B.1 shows per anonymity set size $1 \leq k \leq 10$: the number of citizens that are in an anonymity set of size k, and: their percentage of the total sample population. Results: 1,733,282 citizens, ~62% of our sample population, are unambiguously identifiable by this data alone; another 646,566, 23% of the total, are identifiable up to a group of two persons. In total, ∼99.2% of our sample population has an anonymity set of size $1 \leq k \leq 10$. In other words, the questions observed in this first screen already pretty much put the respondent at risk of perfect identifiability. In contrast, if the pollster would have asked to reveal not the full 'PC6' postal code but only the four-digit 'PC4' postal code, the numbers look significantly different: see Table B.2. In that case, most respondents would at this point in the questionnaire still have had much stronger anonymity; only 4,164 citizens would still have been unambiguously identifiable; and 5,066 would have been identifiable up to a group of two persons. In total, only ∼2.6% of our sample population would have been in an anonymity set of size $1 \leq k \leq 10$. Reversely, ∼97.4% would have been in an anonymity set of size $k > 10$, which may still be sufficient for a non-sensitive questionnaire.

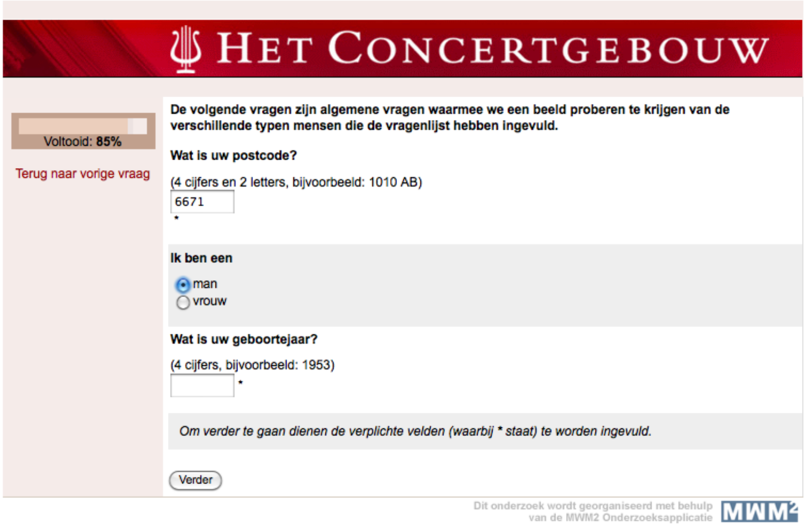
To perform such analysis for the total Dutch population without requiring that the anonymity analyst him/herself has access to microdata of all Dutch citizens, our distribution-informed predictions could be applied; see Chapter 4 and Chapter 5. This requires cooperation between the analyst and the data holder(s), as described in Chapter 7.

For the remainder of the questionnaire we do not have the relevant microdata and therefore cannot determine anonymity set sizes by simple counting. We can, however, estimate upper bounds of anonymity set sizes by looking at the most common value per demographic. The number of citizens sharing that value is the upper bound anonymity set size. What the most common value is and how many citizens share that value can in many cases be looked up from a public statistics repository such as Statline[1]. However, many of the possible answers observed in this particular questionnaire cannot be directly linked to statistics published in Statline; we necessarily permit ourselves some creative freedom in making estimations based on our best judgement. We think that it suffices for the illustrative purpose of this Appendix; real life applications may require more diligence.

We now reset our anonymity analysis and start off with the maximum anonymity set size for this questionnaire, which is the total Dutch population: 16 million citizens. At the end of this Appendix we will consider again the gender, YoB and PC4 postal code.

Remark B.1 *From here on, numbers will indicate 'orders of magnitude'-effects. Higher precision, more elaborate analysis requires additional input data that enables the use of methods such as the distribution-informed prediction developed in Chapter 4, Chapter 5 and Chapter 7.*

[1] Website: `http://statline.cbs.nl/`



De volgende vragen zijn algemene vragen waarmee we een beeld proberen te krijgen van de verschillende typen mensen die de vragenlijst hebben ingevuld.

Wat is uw postcode?

(4 cijfers en 2 letters, bijvoorbeeld: 1010 AB)

6671

*

Ik ben een

● man
○ vrouw

Wat is uw geboortejaar?

(4 cijfers, bijvoorbeeld: 1953)

*

Om verder te gaan dienen de verplichte velden (waarbij * staat) te worden ingevuld.

Dit onderzoek wordt georganiseerd met behulp van de MWM2 Onderzoeksapplicatie

Figure B.1: Revealing demographics: questionnaire screen 1.

Table B.1: Results for $1 \leq k \leq 10$; $QID=\{PC6 + gender + YoB\}$

k	# of citizens	% of total
1	1,733,282	62.4%
2	646,566	23.3%
3	210,963	7.6%
4	79,504	2.9%
5	36,370	1.3%
6	19,200	0.7%
7	11,844	0.4%
8	8,432	0.3%
9	5,490	0.2%
10	4,260	0.2%
TOTAL:	2,755,911	99.2%

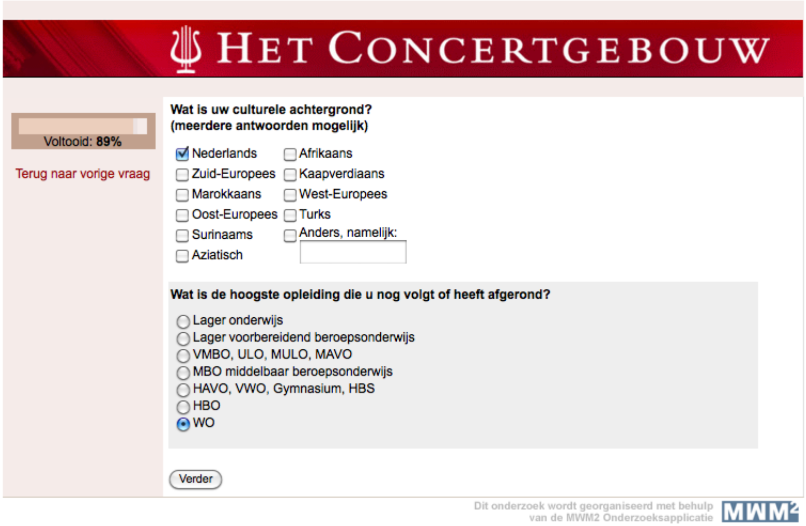
Table B.2: Results for $1 \leq k \leq 10$; $QID=\{PC4 + gender + YoB\}$

k	# of citizens	% of total
1	4,164	0.2%
2	5,066	0.2%
3	5,691	0.2%
4	6,372	0.2%
5	6,925	0.3%
6	7,848	0.3%
7	7,742	0.3%
8	8,392	0.3%
9	9,450	0.3%
10	10,310	0.4%
TOTAL:	71,960	2.6%

The next screen of the questionnaire is shown in Figure B.2. The respondent is asked to reveal cultural background (zero or more answers can be given: Dutch, Southern European, Moroccan, Eastern European, Surinamese, Asian, African, Cape Verdean, Western European, Turkish, and/or 'Other, please specify') and level of completed or current education (one answer must be given: primary education, pre-vocational, secondary general education, middle vocational, higher secondary education or pre-university secondary education, higher vocational, or academic university). In the Netherlands, the largest cohort in education level is middle vocational: $\sim$30% has middle vocational education (i.e., 'MBO' at levels 2, 3 and 4 combined; alas, no statistic was present about MBO 1 or MBO 1-4 combined). If the respondent's educational level is vocational, revealing that decreases his/her anonymity by a factor of $100/30 \approx 3.33$. The anonymity set of 16 million Dutch citizens is hence divided by 3.33 and reduced to 4.8 million citizens. For all other educational levels, the decrease in anonymity is larger. For self-perceived cultural background, we could not find public statistics. However, Statline does contain statistics about non-immigrants and immigrants (citizens known to have at least one parent or grandparent of non-Dutch nationality are counted as immigrant). The largest cohort is non-immigrants: $\sim$79%. If being non-immigrant, revealing that decreases anonymity by a factor of $100/79 \approx 1.26$. Hence, the anonymity set is reduced to 3.8 million citizens. (Revealing that one is immigrant decreases anonymity by a factor of $100/21 \approx 4.76$, and would have reduced the anonymity set to 1 million citizens.)

In the next screen, shown in Figure B.3, the respondent is asked to reveal his/her living situation (one answer must be given: adult(s) with children living at home; two or more adults without children; living at home or with caretakers; single or LAT-relationship; student home; or 'Other, please specify') and the number of children (zero or more answers can be given: no children; number of children aged 1-3; aged 4-7; aged 8-12; aged 13-18; aged 18+). For living situation, the largest cohort is the multiple-person household with children: $\sim$33%. If being in a multiple-person household with children, revealing that decreases anonymity by a factor of $100/33 \approx 3$. Hence, the anonymity set is reduced to 1.2 million citizens. For numbers of children per age group we were not confident about a way to link the questionnaire answers to statistics present in Statline. Alas, we must skip this question.

Lastly, in Figure B.4, the respondent is asked to reveal the category or categories his/her profession belongs to (zero or more answers can be given: high school student; student; pensioner; unemployed; government; education or science; non-profit; cultural sector; media or journalism; commercial; healthcare; musician or singer; self-employed), and gross household income (one answer must be given: less than € 23,000; € 23,000 to € 34,000; € 34,000 to € 56,000; more than € 56,000; or 'I would rather not say'). The largest professional cohort is 'corporate': $\sim$37%. If employed in the corporate sector, revealing that



Wat is uw culturele achtergrond? (meerdere antwoorden mogelijk)

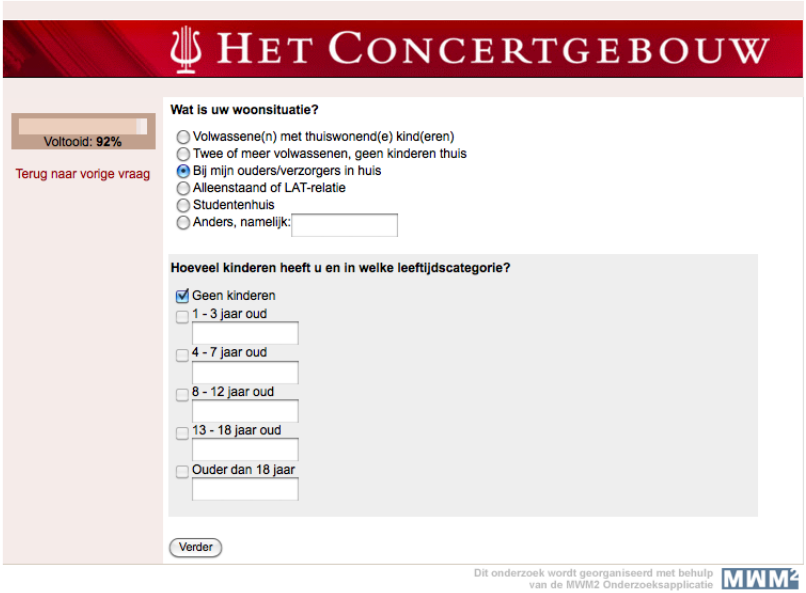
☑ Nederlands
□ Zuid-Europees
□ Marokkaans
□ Oost-Europees
□ Surinaams
□ Aziatisch
□ Afrikaans
□ Kaapverdiaans
□ West-Europees
□ Turks
□ Anders, namelijk:

Wat is de hoogste opleiding die u nog volgt of heeft afgerond?

○ Lager onderwijs
○ Lager voorbereidend beroepsonderwijs
○ VMBO, ULO, MULO, MAVO
○ MBO middelbaar beroepsonderwijs
○ HAVO, VWO, Gymnasium, HBS
○ HBO
● WO

Dit onderzoek wordt georganiseerd met behulp
van de MWM2 Onderzoeksapplicatie

Figure B.2: Revealing demographics: questionnaire screen 2.



Wat is uw woonsituatie?

○ Volwassene(n) met thuiswonend(e) kind(eren)
○ Twee of meer volwassenen, geen kinderen thuis
● Bij mijn ouders/verzorgers in huis
○ Alleenstaand of LAT-relatie
○ Studentenhuis
○ Anders, namelijk:

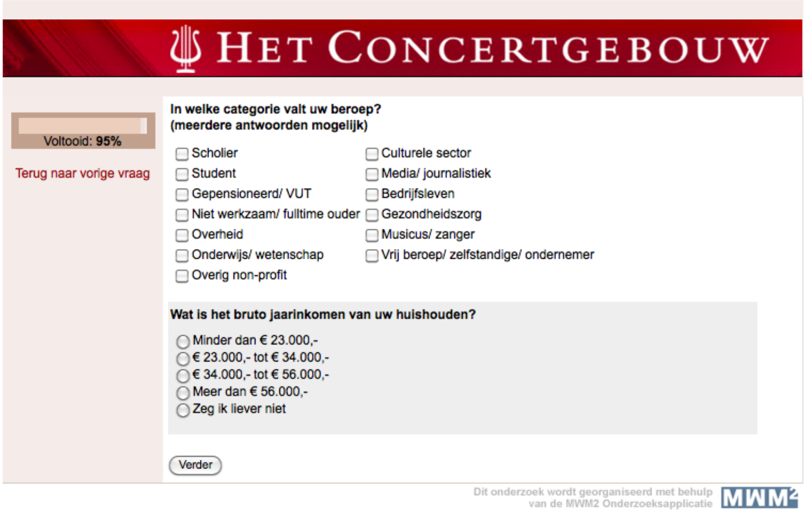
Hoeveel kinderen heeft u en in welke leeftijdscategorie?

☑ Geen kinderen
□ 1 - 3 jaar oud
□ 4 - 7 jaar oud
□ 8 - 12 jaar oud
□ 13 - 18 jaar oud
□ Ouder dan 18 jaar

Dit onderzoek wordt georganiseerd met behulp
van de MWM2 Onderzoeksapplicatie

Figure B.3: Revealing demographics: questionnaire screen 3.

decreases anonymity by $100/37 \approx 2.7$. Hence, the anonymity set is reduced to 470k citizens. For gross household income, 'more than 56,000' is the largest cohort: $\sim$44%. If having a gross household income of more than € 56,000, revealing that decreases anonymity by a factor of $100/44 \approx 2.3$. Hence, the anonymity set size is reduced to 94k citizens. Here, analysis of interval width, as developed in Chapter 6, might have been of help during the development of the questionnaire, to establish income intervals that are useful to the pollster but also not needlessly identifying from the respondent's point of view.



HET CONCERTGEBOUW

Voltooid: **95%**

Terug naar vorige vraag

In welke categorie valt uw beroep?
(meerdere antwoorden mogelijk)

□ Scholier
□ Student
□ Gepensioneerd/ VUT
□ Niet werkzaam/ fulltime ouder
□ Overheid
□ Onderwijs/ wetenschap
□ Overig non-profit
□ Culturele sector
□ Media/ journalistiek
□ Bedrijfsleven
□ Gezondheidszorg
□ Musicus/ zanger
□ Vrij beroep/ zelfstandige/ ondernemer

Wat is het bruto jaarinkomen van uw huishouden?

○ Minder dan € 23.000,-
○ € 23.000,- tot € 34.000,-
○ € 34.000,- tot € 56.000,-
○ Meer dan € 56.000,-
○ Zeg ik liever niet

Verder

Dit onderzoek wordt georganiseerd met behulp van de MWM2 Onderzoeksapplicatie

MWM2

Figure B.4: Revealing demographics: questionnaire screen 4.

For the fictional $QID = \{PC4 + gender + YoB\}$, the most common value in our sample population is $\{1056 + \text{F} + 1981\}$: $\sim$0.002% of the total Dutch population. Revealing that information decreases anonymity by a factor of $100/0.002 = 50{,}000$. Hence, the anonymity set is reduced to four citizens: see Table B.3. In conclusion, anonymous respondents should expect that their answers can be traced down to a group of four or less individuals.

Note, however, that we explicitly treated the questions as if they were independent from each other. In real life, variates such as income and YoB might be correlated. Revealing one variate then also partially reveals the other. And hence, revealing the other adds less new information than if both were not correlated. Such effects may result in a larger anonymity set, and thus in a more optimistic outlook than the expectation stated above; i.e., that respondents' answers can be traced down to a group of four or less individuals. The work developed in Chapter 6 may be helpful in examining such effects.

Table B.3: Estimated decrease in anonymity per question

Demographic	Largest cohort	Decrease k	Possible identities
-	-	-	16,000,000
education	'vocational'	3.33	4,804,804
+ cultural background	'non-immigrant'	1.26	3,813,337
+ living situation	'1+ household w/children'	3	1,271,112
+ work	'corporate sector'	2.7	470,782
+ gross income	'more than € 56,000'	2.3	204,347
+ $\{PC4{+}gender{+}YoB\}$	'1056 + F + 1981'	50,000	**4**

In reality, probably hardly anyone belongs to the largest cohort in every question. The proper way to interpret the result of this (partial and rough) analysis is to say: "at best, a respondent honestly answering all questions in this questionnaire is indistinguishable from three other persons; but most respondents will belong to a smaller anonymity set". Additional analysis is needed to determine what anonymity remains if instead of disclosing PC4, gender and YoB, the respondent would only disclose, say, municipality, gender and YoB. Of course, anyone attempting to trace the survey data to individuals would also need to have access to identified microdata containing all these columns. Despite attempts to make an inventory of data collections throughout society [3, 27, 70], there is no complete picture about what microdata is processed and by whom. When collecting data about sensitive topics such as politics, health and sex habits, it probably makes sense to assume the worst-case scenario: i.e., that somewhere, an identified table exists that contains all columns, all filled with truthful values (as many governments seek to create). When collecting data about non-sensitive surveys, more optimistic assumptions might be justified; however, it should not be disregarded that leaks of non-sensitive survey microdata may itself help accomplish that worst-case scenario.

C Publications

Scientific

- (journal) **Matthijs R. Koot**, Michel R.H. Mandjes, Guido J. van 't Noordende, Cees Th.A.M. de Laat: "A Probabilistic Perspective on Re-Identifiability", Mathematical Population Studies, Submitted November 2011

- (conference) **Matthijs R. Koot**, Michel R.H. Mandjes, Guido J. van 't Noordende, Cees Th.A.M. de Laat: "Efficient Probabilistic Estimation of Quasi-Identifier Uniqueness", Proceedings of NWO ICT.Open 2011, November 2011

- (journal) **Matthijs R. Koot**, Michel R.H. Mandjes: "The analysis of singletons in generalized birthday problems", Probability in the Engineering and Informational Sciences. April 2012

- (conference) **Matthijs R. Koot**, Guido J. van 't Noordende, and Cees Th.A.M. de Laat: "A Study on the Re-Identifiability of Dutch Citizens", HotPETS track of the Privacy-Enhancing Technology Symposium 2010 (PETS2010), July 2010

- (journal) Guido J. van 't Noordende, Silvia D. Olabarriaga, **Matthijs R. Koot**, Cees Th.A.M. de Laat: "A Trusted Data Storage Infrastructure for Grid-based Medical Applications", International Journal of Grid and High Performance Computing, 1(2), 1-14; April-June 2009

- (conference) Guido van 't Noordende, Silvia D. Olabarriaga, **Matthijs R. Koot**, Cees Th.A.M. de Laat: "Privacy and Trust for Grid-based Medical Applications", The 8th IEEE International Symposium on Cluster Computing and the Grid (CCGRID2008); May 2008

Non-scientific

C.0.1 Professional

- "Vertrouwen in vrije digitale X.509 certificaten", Informatiebeveiliging, July 2008
- "Interview with Amit Jasuja (Oracle)", Informatiebeveiliging, September 2007
- "E-DRM: Bescherming van informatie binnen en buiten het bedrijf", Informatiebeveiliging, November 2007

C.0.2 Press

- "UvA-wetenschapper wijst op risico's Google Profiles", University of Amsterdam press release, June 2011
- "35 Million Google Profiles Captured In Database", Information Week, May 2011
- "Google Profiles: Is Easy Aggregation An Invasion Of Privacy?", Forbes blog (Kashmir Hill), May 2011
- "35 Million Google Profiles Collected", Slashdot, May 2011
- "35m Google Profiles dumped into private database", The Register, May 2011

Bibliography

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 901–909, 2005.

[2] I. Altman. *The environment and social behavior: privacy, personal space, territory, crowding.* Brooks/Cole Pub. Co., 1975.

[3] R. Anderson, I. Brown, T. Dowty, P. Inglesant, W. Heath, and A. Sasse. Database State, March 2009.

[4] J. C. M. Baeten. A brief history of process algebra. *Theoretical Computer Science*, 335:131–146, May 2005.

[5] M. Bangemann. *Europe and the global Information Society (Bangemann report)*, 1994.

[6] J. Bergstra and J. Klop. Algebra of communicating processes with abstractions. In *Theoretical Computer Science*, volume 33, pages 77–121, Netherlands, 1985.

[7] M. Bhargava and C. Palamidessi. *Probabilistic anonymity*, pages 171–185. Springer-Verlag, London, UK, 2005.

[8] D. Boyd. *Taken Out of Context: American Teen Sociality in Networked Publics.* PhD thesis, University of California, Berkeley, Berkeley, CA, USA, 2008.

[9] M. Camarri and J. Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electronic Journal of Probability*, 5:1–18, 2000.

[10] A. Campan, T. M. Truta, and N. Cooper. p-Sensitive k-Anonymity with generalization constraints. *Trans. Data Privacy*, 3:65–89, August 2010.

[11] CBS. *Documentatierapport Landelijke Medische Registratie (LMR) 2005V1*, March 2007.

[12] CBS. *Documentatierapport Bijstandsfraudestatistiek (BFS) 200901-06V1*, November 2009.

[13] CBS. *Documentatierapport Landelijke Medische Registratie (LMR) 2007V1*, July 2009.

[14] CBS. Website: Cbs - gemeentelijke indeling op 1 januari 2009, 2009. `http://www.cbs.nl/`.

[15] CBS. Website: Cbs - ziekenhuisopnamen - dataverzameling, 2009. `http://www.cbs.nl/`.

[16] K. Chatzikokolakis and C. Palamidessi. Probable innocence revisited. In T. Dimitrakos, F. Martinelli, P. Ryan, and S. Schneider, editors, *Formal Aspects in Security and Trust*, volume 3866 of *Lecture Notes in Computer Science*, pages 142–157. Springer Berlin / Heidelberg, 2006.

[17] D. Chaum. The dining cryptographers problem: Unconditional sender and recipient untraceability. *Journal of Cryptology*, 1:65–75, 1988.

[18] D. L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun. ACM*, 24(2):84–90, Feb. 1981.

[19] T. Chothia, S. Orzan, J. Pang, and M. T. Dashti. A framework for automatically checking anonymity with μCRL. In *Proceedings of the 2nd international conference on Trustworthy global computing*, TGC'06, pages 301–318, Berlin, Heidelberg, 2007. Springer-Verlag.

[20] S. Claußand S. Schiffner. Structuring anonymity metrics. In *Digital Identity Management*, pages 55–62, 2006.

[21] T. Dalenius. Finding a needle in a haystack-or identifying anonymous census record. *Journal of Official Statistics*, 2:329–336, 1986.

[22] J. DeCew. Privacy. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, CA, USA, fall 2008 edition, 2008.

[23] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications 2nd ed.,*. Springer Verlag, New York, 1998.

[24] Y. Deng, C. Palamidessi, and J. Pang. Weak probabilistic anonymity. *Electronical Notes in Theoretical Computer Science*, 180:55–76, June 2007.

[25] Y. Deng, J. Pang, and P. Wu. Measuring anonymity with relative entropy. In *Proceedings of the 4th international conference on Formal aspects in security and trust*, FAST'06, pages 65–79, Berlin, Heidelberg, 2007. Springer-Verlag.

[26] P. Diaconis and F. Mosteller. Methods for studying coincidences. *Journal of the American Statistical Association*, 84:853–861, 1989.

[27] Dutch Scientific Council for Government Policy (WRR). *iOverheid*, 2011.

[28] W. Feller. *An Introduction to Probability Theory and its Applications, 3rd Edition.* Wiley, New York, NY, United States, 1968.

[29] L. Forer. *A Chilling Effect: The Mounting Threat of Libel and Invasion of Privacy Actions to the First Amendment.* Norton, 1989.

[30] A. Fujioka, T. Okamoto, and K. Ohta. A practical secret voting scheme for large scale elections. In *Proceedings of the Workshop on the Theory and Application of Cryptographic Techniques: Advances in Cryptology*, ASIACRYPT '92, pages 244–251, London, UK, 1993. Springer-Verlag.

[31] M. Gail, G. Weiss, N. Mantel, and S. O'Brien. A solution to the generalized birthday problem with application to allozyme screening for cell culture contamination. *Journal of Applied Probability*, 16:242–251, 1979.

[32] P. Golle. Revisiting the uniqueness of simple demographics in the us population. In *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 77–80, New York, NY, USA, 2006. ACM.

[33] J. Y. Halpern and K. R. O'Neill. Anonymity and information hiding in multiagent systems. *Journal of Computer Security*, 2004.

[34] J. Y. Halpern and K. R. O'Neill. Secrecy in multiagent systems. *ACM Transactions on Information and System Security*, 12:5:1–5:47, October 2008.

[35] A. Harel, A. Shabtai, L. Rokach, and Y. Elovici. m-score: estimating the potential damage of data leakage incident by assigning misuseability weight. In *Proceedings of the 2010 ACM workshop on Insider threats*, Insider Threats '10, pages 13–20, 2010.

[36] I. Hasuo and Y. Kawabe. Probabilistic anonymity via coalgebraic simulations. In *Proceedings of the 16th European conference on Programming*, ESOP'07, pages 379–394, Berlin, Heidelberg, 2007. Springer-Verlag.

[37] C. A. R. Hoare. Communicating sequential processes. *Commun. ACM*, 21(8):666–677, Aug. 1978.

[38] J. Holvast. *Op weg naar een risicoloze maatschappij? De vrijheid van de mens in de informatie-samenleving.* Leiden, 1986.

[39] G. Horn. Online searches and offline challenges: The chilling effect, anonymity and the new fbi guidelines. *New York University Annual Survey of American Law 60 N.Y.U.*, 735, 2004-2005.

[40] K. Joag-Dev and F. Proschan. The birthday problem with unlike probabilities. *American Mathematical Monthly*, 99:10–12, 1992.

[41] J. Klotz. The birthday problem with unequal probabilities. *Technical Report No. 59, Department of Statistics, University of Wisconsin*, 1979.

[42] M. Koot and M. Mandjes. The analysis of singletons in generalized birthday problems. *Probability in the Engineering and Information Sciences (to appear)*, 2011.

[43] M. Koot, M. Mandjes, G. van 't Noordende, and C. de Laat. Efficient probabilistic estimation of quasi-identifier uniqueness. In *Proceedings of NWO ICT.Open 2011*, November 2011.

[44] M. Koot, M. Mandjes, G. van 't Noordende, and C. de Laat. A probabilistic perspective on re-identifiability. *Mathematical Population Studies (submitted)*, 2011.

[45] M. Koot, G. van 't Noordende, and C. de Laat. A study on the re-identifiability of Dutch citizens. In *Electronic Proceedings of HotPETS 2010*, July 2010.

[46] S. Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

[47] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[48] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[49] S. Lodha and D. Thomas. Probabilistic anonymity. In *Proceedings of the 1st ACM SIGKDD international conference on Privacy, security, and trust in KDD*, PinKDD'07, pages 56–79, 2008.

[50] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-Diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1, March 2007.

[51] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37:179–192, 2004.

[52] M. Mandjes. *Large Deviations for Gaussian Queues.* Wiley, Chichester, 2007.

[53] M. Mandjes. Generalized birthday problems in the large-deviations regime. *Submitted.*, 2011.

[54] R. A. Merkt, A. L. Mchose, and A. Chiappone. The "new jersey self-defense law". *Assembly No. 159, State of New Jersey, 213th Legislature*, 2008.

[55] R. Milner. *A Calculus of Communicating Systems.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

[56] R. Milner, J. Parrow, and D. Walker. A calculus of mobile processes, i. *Inf. Comput.*, 100:1–40, September 1992.

[57] F. Mosteller. Understanding the birthday problem. In S. Fienberg and D. Hoaglin, editors, *Selected Papers of Frederick Mosteller*, Springer Series in Statistics, pages 349–353. Springer New York, 2006.

[58] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125, Washington, DC, USA, 2008. IEEE Computer Society.

[59] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 665–676, 2007.

[60] A. Nicolaï. Kst99754: Modernisering gemeentelijke basisadministratie persoonsgegevens, 2006.

[61] H. Nissenbaum. *Privacy in context: technology, policy, and the integrity of social life.* Stanford Law Books, 2010.

[62] T. S. Nunnikhoven. A birthday problem solution for nonuniform birth frequencies. *The American Statistician*, 46(4):pp. 270–274, 1992.

[63] NVVB. *Schema voor schriftelijke verzoeken om gegevensverstrekking uit de GBA*, January 2010.

[64] A. Pfitzmann and M. Hansen. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf, Aug. 2010. v0.34.

[65] B. Pierce. *Foundational Calculi for Programming Languages*, pages –. CRC Press, Boca Raton, FL, 1997.

[66] W. L. Prosser. Privacy. *California Law Review*, 48(3), 1960.

[67] M. Reiter and A. Rubin. Crowds: Anonymity for web transactions. *ACM Transactions on Information and System Security*, 1(1), June 1998.

[68] M. A. Rothstein. Is deidentification sufficient to protect health privacy in research? *The American Journal of Bioethics*, 10(9):3–11, 2010.

[69] P. Rust. The effect of leap years and seasonal trends on the birthday problem. *The American Statistician*, 30:197–198, 1976.

[70] B. W. Schermer and T. Wagemans. Onze digitale schaduw, Jan. 2009.

[71] S. Schneider and A. Sidiropoulos. Csp and anonymity. In E. Bertino, H. Kurth, G. Martella, and E. Montolivo, editors, *ESORICS*, volume 1146 of *Lecture Notes in Computer Science*, pages 198–218. Springer, 1996.

[72] A. Serjantov and G. Danezis. Towards an information theoretic metric for anonymity. In R. Dingledine and P. Syverson, editors, *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002)*. Springer-Verlag, LNCS 2482, April 2002.

[73] A. Solanas, F. Sebé, and J. Domingo-Ferrer. Micro-aggregation-based heuristics for p-sensitive k-anonymity: one step beyond. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, PAIS '08, pages 61–69, New York, NY, USA, 2008. ACM.

[74] D. J. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3):477–560, Jan. 2005.

[75] L. Sweeney. Uniqueness of simple demographics in the u.s. population. *LIDAP-WP4 Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000*, 2000.

[76] L. Sweeney. *Computational disclosure control: a primer on data privacy protection.* PhD thesis, Massachusetts Institute of Technology, 2001. Supervisor: Abelson, Hal.

[77] L. Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 10:557–570, 2002.

[78] P. F. Syverson and S. G. Stubblebine. Group principals and the formalization of anonymity. In *Proceedings of the World Congress on Formal Methods (1)*, pages 814–833, 1999.

[79] Tieto Netherlands Healthcare BV. *Landelijke Medische Registratie (LMR) Gebruikershandleiding*, 2009.

[80] G. Tóth, Z. Hornák, and F. Vajda. Measuring anonymity revisited. In S. Liimatainen and T. Virtanen, editors, *Proceedings of the Ninth Nordic Workshop on Secure IT Systems*, pages 85–90, Espoo, Finland, November 2004.

[81] University of California, Los Angeles. *SOCR: Statistics Online Computational Resource, Data Dinov 020108 HeightsWeights*, 1993.

[82] J. van Eijck and S. Orzan. Epistemic verification of anonymity. *Electron. Notes Theor. Comput. Sci.*, 168:159–174, February 2007.

[83] R. von Mises. Über Aufteilungs- und Besetzungswahrscheinlichkeiten. *Revue de la Faculté des Sciences de L'Université d'Istanbul*, 4:145–163, 1938.

[84] P. Wang, P. Ning, and D. S. Reeves. A k-anonymous communication protocol for overlay networks. In *Proceedings of the 2nd ACM symposium on Information, computer and communications security*, ASIACCS '07, pages 45–56, 2007.

[85] S. D. Warren and L. D. Brandeis. The right to privacy. *Harvard Law Review*, IV, December 1890.

[86] D. Webb. *Privacy and Solitude in the Middle Ages.* Hambledon Continuum, London, 2007.

[87] A. Westin. *Privacy and freedom.* Atheneum, New York, 1970.

[88] L. Willenborg and T. de Waal. *Statistical Disclosure Control in Practice*, volume 111 of *Lecture Notes in Statistics.* Springer, 1996. ISBN: 978-0-387-94722-8.

[89] X. Wu and E. Bertino. Achieving k-anonymity in mobile ad hoc networks. In *Proceedings of the First international conference on Secure network protocols*, NPSEC'05, pages 37–42, 2005.

[90] X. Xiao and Y. Tao. m-invariance: towards privacy preserving republication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 689–700, 2007.

List of Figures

1.1 Privacy in 'functions' and 'states', according to Westin [87]. 5

1.2 Taxonomy of privacy violations according to Solove [74]. 6

2.1 Degrees of anonymity according to Reiter and Rubin [67] 14

2.2 Linking to re-identify data [76] 24

3.1 Box-and-whisker plot showing anonymity set sizes k_A, per municipality. Whiskers denote the minimum and maximum values; the boxes are defined by lower and upper quartiles and the median value is shown. 37

3.2 Box-and-whisker plot showing anonymity set sizes k_B, per municipality. Whiskers denote min-max values. 38

4.1 For all Dutch municipalities: the Kullback-Leibler distance and the estimated uniqueness probability, when revealing age. 48

4.2 For all Dutch municipalities: the Kullback-Leibler distance and the estimated uniqueness probability, when revealing age and gender. . 48

4.3 For two Dutch municipalities: the uniqueness probability as a function of the group size k; also the curve under uniformity has been added. 51

4.4 For all Dutch municipalities: the effect of aggregated (age) statistics on the KL-distance. 51

5.1 Mean number of singletons, as a function of the Kullback-Leibler distance κ. Left panels: full population; right panels: ages 0–79 only. Top to bottom: $k = 60, 90, 120$. 67

5.2 Variance of the number of singletons, as a function of the Kullback-Leibler distance κ. Left panel: full population; right panel: ages 0–79 only. 69

6.1 ϕ_1, ϕ_2, and ϕ_3 for two municipalities, as a function of the population size of the postal code area. 76

6.2 ϕ_1 for all municipalities, as a function of the Kullback-Leibler distance κ, for $k = 20, 40, 60, 80$. Notice that the observations (ψ) are accurately predicted (ϕ) by the Kullback-Leibler distance (κ) for various population sizes (k). 77

6.3 Graphical illustration of accuracy of the $O(\Delta)$-approximation; $\mathbb{E}S$ as a function of k for height, weight and birthday. The lines correspond to the estimates resulting from simulation, and the '+' with the $O(\Delta)$-approximation. Tables show mean number of singletons for various values of k. 87

6.4 Expected number of singletons, for $k = 5, 10, 20, 40$, respectively ($k = 30$ is skipped due to page layout). The solid lines are the simulation-based estimates, the dots are the approximations based on the formulas derived in this Section. Per picture, the first 6 data points correspond to $\Delta_H = 0.5$ cm, the second 6 data points to $\Delta_H = 1.0$ cm, the third set of 6 data points to $\Delta_H = 2.0$ cm, the fourth set of 6 data points to $\Delta_H = 5.0$ cm, the fifth set of 6 data points to $\Delta_H = 10.0$ cm, and the last set of 6 data points to $\Delta_H = 20.0$ cm. Within each group of 6 data points, these correspond to $\Delta_W = 0.5, 1.0, 2.0, 5.0, 10, 20$ kg. 88

6.5 Left panel: effect of Δ_W for Δ_H fixed; right panel: effect of Δ_H for Δ_W fixed. 89

7.1 Preliminary model for applying distribution-informed privacy predictions as part of privacy policy making. 93

B.1 Revealing demographics: questionnaire screen 1. 107

B.2 Revealing demographics: questionnaire screen 2. 109

B.3 Revealing demographics: questionnaire screen 3. 109

B.4 Revealing demographics: questionnaire screen 4. 110

List of Tables

3.1 Municipalities included in our study (ordered by number of citizens) 31
3.2 Anonymity set size k for various (potential) quasi-identifiers 33
3.3 Number of Dutch citizens per anonymity set size, for various quasi-identifiers 34
3.4 Statistical summary of k_A, divided by municipality (ordered by median) 36
3.5 Statistical summary of k_B, divided by municipality (ordered by median) 38

A.1 $\zeta(k,N)$ for $k = 1\ldots,50$ and $N = 1,\ldots,20$. 104

B.1 Results for $1 \leq k \leq 10$; *QID=*$\{$*PC6 + gender + YoB*$\}$ 107
B.2 Results for $1 \leq k \leq 10$; *QID=*$\{$*PC4 + gender + YoB*$\}$ 107
B.3 Estimated decrease in anonymity per question 111

Abstract (English)

In our increasingly computer-networked world, more and more personal data is collected, linked and shared. This raises questions about privacy — i.e. about the feeling and reality of enjoying a private life in terms of being able to exercise control over the disclosure of information about oneself. In attempt to provide privacy, databases containing personal data are sometimes de-identified, meaning that obvious identifiers such as Social Security Numbers, names, addresses and phone numbers are removed. In microdata, where each record maps to a single individual, de-identification might however leave columns that, combined, can be used to re-identify the de-identified data. Such combinations of columns are commonly referred to as Quasi-IDentifiers (QIDs).

Sweeney's model of k-anonymity addresses this problem by requiring that each QID value, i.e., a combination of values of multiple columns, present in a data set must occur at least k times in that data set, asserting that each record in that set maps to at least k individuals, hence making records and individuals unlinkable. Many extensions have been proposed to k-anonymity, but always address the situation in which data has already been collected and must be de-identified afterwards. The question remains: can we predict what information will turn out to be identifiable, so that we may decide what (not) to collect beforehand?

To build a case we first inquired into the (re-)identifiability of hospital intake data and welfare fraud data about Dutch citizens, using large amounts of data collected from municipal registry offices. We show the large differences in (empirical) privacy, depending on where a person lives. Next, we develop a range of novel techniques to predict aspects of anonymity, building on probabilistic theory, and specifically birthday-problem theory and large-deviations theory.

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. We present an approach to estimate anonymity for the more realistic case where the variables composing a QID follow a non-uniform distribution. We present an efficient and accurate approximation of the uniqueness probability using the group size and a measure of heterogeneity called the Kullback-Leibler distance. The approach is thoroughly validated by comparing the approximation with results from a simulation using the real demographic information we collected in the Netherlands.

We further describe novel techniques for characterizing the number of singletons, i.e., the number of persons have 1-anonymity and are unambiguously (re-)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. 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 our results to be replicated/reproduced by others).

Next, we study in detail three specific issues in singletons analysis. First, we assess the effect on identifiability of non-uniformity of the possible outcomes. Suppose one has the ages of the members of the group; what is the effect on the 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. 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.

Lastly, we give preliminary ideas for applying our techniques in real life. We hope these are suitable and useful input to the privacy debate; practical application will depend on competence and willingness of data holders and policy makers to correctly identify quasi-identifiers. In the end, it remains a matter of policy what value of k can be considered *sufficiently strong* anonymity for particular personal information.

Abstract (Dutch)

In onze steeds verdergaand verbonden wereld worden meer en meer persoonsgegevens verzameld, gekoppeld en gedeeld. Hierdoor dringen zich vragen op over privacy — over het gevoel en de realiteit van de persoonlijke levenssfeer en het invloed kunnen uitoefenen over verspreiding van persoonlijke informatie. Omwille van privacy worden databases soms gedeïdentificeerd, dat wil zeggen: ontdaan van evident identificerende informatie zoals Burger Service Nummers, namen, adressen en telefoonnummers. Echter, in microdata, waarbij records informatie bevatten op individueel niveau, kunnen na deïdentificatie kolommen achterblijven die in combinatie zouden kunnen worden gebruikt om de gedeïdentificeerde data te heridentificeren. Zulke combinaties van kolommen worden 'Quasi-IDentifiers' (QIDs) genoemd.

Sweeney's model van k-anonimiteit adresseert dat probleem door te waarborgen dat elke QID-waarde in een tabel ten minste k keren in die tabel voorkomt, waardoor elk record in de tabel niet valt te herleiden tot minder dan k verschillende personen en dus onlinkbaarheid ontstaat. Er zijn diverse uitbreidingen voorgesteld van k-anonimiteit, maar die zijn alleen bruikbaar in een situatie waarin vooraf gegevens zijn verzameld en er achteraf wordt geïdentificeerd. De vraag blijft: valt te voorspellen welke gegevens quasi-identificerend zullen zijn, zodat we vooraf kunnen besluiten die gegevens niet, of op minder fijnkorrelig niveau, te verzamelen?

Ter onderbouwing van het probleem is eerst onderzoek gedaan naar heridentificeerbaarheid van Nederlandse persoonsgegevens over ziekenhuisopnames en bijstandsfraude, gebruikmakend van een grote hoeveelheid gegevens uit Gemeentelijke Basis Administraties. We tonen aan dat er in deze voorbeelden grote verschillen bestaan in privacy, afhankelijk van de gemeente waar iemand woont. Vervolgens zijn nieuwe technieken ontwikkeld om eigenschappen van anonimiteit te voorspellen, voortbouwend op kansrekening en in het bijzonder de ‘birthday paradox’ en ‘large deviations theory’.

Anonimiteit kan worden gekwantificeerd als de kans dat elk lid van een groep uniek kan worden geïdentificeerd via een QID. Het schatten van deze uniciteitskans is eenvoudig wanneer alle mogelijke QID-waarden even waarschijnlijk zijn, dus, wanneer de onderliggende verdeling homogeen is. Dit werk presenteert een nieuwe aanpak voor het schatten van anonimiteit voor het meer realistische scenario waarin de verdeling van QID-waarden heterogeen is. Een efficiënte en accurate benadering van de uniciteitskans wordt gepresenteerd, gebruikmakend van groepsgroottes en Kullback-Leibler afstanden (een maat van heterogeniteit). Het gepresenteerde wordt grondig gevalideerd door de benadering te vergelijken met uitkomsten van een simulatie gebaseerd op echte demografische gegevens die in Nederland zijn verzameld.

Verder worden nieuwe technieken beschreven om het aantal ‘singletons’ te karakteriseren, dat wil zeggen, het aantal personen dat 1-anonimiteit heeft en dus ondubbelzinnig (her)identificeerbaar is, in het ‘generalized birthday problem’. Dat wil zeggen, het ‘birthday problem’ waarbij geboortedagen niet-uniform over het jaar zijn verdeeld. Benaderingen voor het gemiddelde en de variantie worden gepresenteerd die een expliciete indicatie geven van de impact die heterogeniteit op anonimiteit heeft, in termen van de Kullback-Leibler afstand ten opzichte van de homogene verdeling. Een iteratief schema wordt gepresenteerd om de verdeling van het aantal singletons te bepalen. De formules zijn experimenteel gevalideerd via demografische gegevens die openbaar beschikbaar zijn.

Vervolgens worden drie specifieke aspecten van de analyse van singletons in detail bestudeerd. Ten eerste is het effect bestudeerd dat niet-uniformiteit van een verdeling heeft op de mogelijke uitkomsten. Stel dat men de leeftijden van alle leden van een groep kent: wat is het effect op identificeerbaarheid dat sommige leeftijden vaker voorkomen dan andere? Opnieuw blijkt dat de heterogeniteit goed kan worden beschreven via één enkel getal, de Kullback-Leibler afstand, en dat de uitkomsten van de formules accuraat zijn. Ten tweede is het effect van fijnkorreligheid van gegevens op identificeerbaarheid bestudeerd. Het is duidelijk dat een leeftijd in maanden meer identificerend is dan een leeftijd in jaren. Een techniek wordt gepresenteerd om dit effect expliciet te kwantificeren in termen van intervalbreedtes. Ten derde is het effect van correlatie tussen numerieke variabelen bestudeerd met als leidend voorbeeld lengte en gewicht, die positief gecorreleerd zijn. Voor de benadering van het niveau van identificeerbaarheid wordt een expliciete formule gepresenteerd die gebruik maakt van de correlatiecoëfficiënt. De formules zijn experimenteel gevalideerd via openbaar beschikbare gegevens en via niet-openbare gegevens over Nederlandse burgers die aan het begin van deze studie zijn verzameld.

Ten slotte geven we preliminaire ideeën voor toepassing van de technieken in de echte wereld. Deze zijn bedoeld als stof voor discussie in het privacydebat: praktische toepassing is afhankelijk van de competentie en bereidheid van gegevenshouders en beleidsmakers om op QIDs te letten. Welke waarde van k als *voldoende sterke* anonimiteit wordt beschouwd voor bepaalde persoonsgegevens, blijft een beleidskwestie.