Towards Inter-temporal Privacy Metrics
Towards Inter-temporal Privacy Metrics
Stefan Berthold. *Towards Inter-temporal Privacy Metrics*

Licentiate thesis

Karlstad University Studies 2011:25
ISSN 1403-8099
ISBN 978-91-7063-357-7

© The Author

Distribution:
Karlstad University
Faculty of Economic Sciences, Communication and IT
Computer Science
S-651 88 Karlstad
+46 54 700 10 00

www.kau.se

Print: Universitetstryckeriet, Karlstad 2011
Towards Inter-temporal Privacy Metrics

STEFAN BERTHOLD
Department of Computer Science, Karlstad University

Abstract

Informational privacy of individuals has significantly gained importance after information technology has become widely deployed. Data, once digitalised, can be copied and distributed at negligible costs. This has dramatic consequences for individuals that leave traces in the form of personal data whenever they interact with information technology. The right of individuals for informational privacy, in particular to control the flow and use of their personal data, is easily undermined by those controlling the information technology.

The objective of this thesis is to study the measurement of informational privacy with a particular focus on scenarios where an individual discloses personal data to a second party, the data controller, which uses this data for re-identifying the individual within a set of others, the population. Several instances of this scenario are discussed in the appended papers, most notably one which adds a time dimension to the scenario for modelling the effects of the time passed between data disclosure and usage. This extended scenario leads to a new framework for inter-temporal privacy metrics.

The common dilemma of all privacy metrics is their dependence on the information available to the data controller. The same information may or may not be available to the individual and, as a consequence, the individual may be misguided in his decisions due to limited access to the data controller’s information when using privacy metrics. The goal of this thesis is thus not only the specification of new privacy metrics, but also the contribution of ideas for mitigating this dilemma. However, a solution will rather be a combination of technological, economical and legal means than a purely technical solution.

Keywords: privacy, unlinkability, metrics, uncertainty, valuation process, domain-specific language, anonymous communication.
Acknowledgements

First of all, I would like to thank my supervisor Prof. Simone Fischer-Hübner and my co-supervisor Prof. Stefan Lindskog for their steady support, interesting discussions, and advise during my work at Karlstad University.

I am also grateful to Prof. Viiveke Fåk for accepting the role of the opponent in my licentiate thesis defence.

I would like to thank my colleagues in the PriSec group at at Karlstad University, particularly Hans Hedbom, Reine Lundin, Tobias Pulls, and Ge Zhang; and my former colleagues at Technische Universität Dresden, particularly Rainer Böhme, Sebastian Clauß, Thomas Gloe, Matthias Kirchner, Stefan Köpsell, Stefanie Pötzsch, and Antje Winkler for challenging me in inspiring discussions, pointing me to interesting questions, and collaborating with me.

Thanks also to all my colleagues at the Department of Computer Science for providing such a friendly working environment.

A huge thank also to my mum, Erika Berthold, and my sister, Susann Berthold, for always being available; and to all my friends in Karlstad, in Dresden, and in quite a few other places around the world for keeping in touch with me, even over long distances.

Research leading to this work has received funding from the European Community’s Sixth Framework Programme (FP6/2002–2006) under grant agreement no. 507512 and the Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 216483 as well as the Research Council of Norway through the PETweb II project and the German Academic Exchange Service (DAAD).

Karlstad, April 2011

Stefan Berthold
List of Appended Papers


Comments on my Participation

Paper I I was the main author of all parts of the paper and received a number of useful comments from Sebastian Clauß.

Paper II This paper was a joint work of all three authors. I was the driving force behind the first ideas of this paper and later took the responsibility for adding the sections on the attacks and bringing the theoretical ideas together with the study in a coherent paper. Stefan Köpsell carried out the empirical study and drafted the results of it in German. He was also the main author of the sections about anonymity services in a nutshell and the legal background. Rainer Böhme contributed the section about probabilistic intersection attacks and supervised the study.

Paper III This paper was a joint effort with Rainer Böhme. I authored the main part of the paper, i.e., the related work on anonymity and unlinkability as well as Sections 3, 4, 5, and 6. Rainer authored the introduction, the related work on financial methods in information security, and the conclusions; and contributed his knowledge about economics to the main ideas of this paper.

Paper IV This paper was solely authored by myself. During the course of writing the paper, I received many constructive and motivating comments from colleagues and friends which have been acknowledged in the paper.
Selection of Other Peer-reviewed Publications


Contributions to Project Deliverables


Contents

List of Appended Papers vii

INTRODUCTORY SUMMARY 1

1 Introduction 3

2 Background and Related Work 3
   2.1 Anonymity 4
   2.2 Unlinkability 6
   2.3 Information Asymmetry 7
   2.4 Privacy Policy Languages 8

3 Research Questions 8

4 Research Methods 9

5 Contributions 10

6 Summary of Appended Papers 11

7 Conclusions and Future Work 12

PAPER I
Linkability Estimation Between Subjects and Message Contents Using Formal Concepts 18

1 Introduction 21

2 Formalization of Messages and Contents 24
   2.1 Message Lattice 25
   2.2 Data Lattice 26
   2.3 Improve Message Contents by Derivable Data 27
   2.4 Intermediate Results 29

3 Deducing Subject Knowledge 29
   3.1 Subject-Pseudonym Lattice 29
   3.2 Assigning Subjects to Messages 31
   3.3 Contents towards Subject Knowledge 33
   3.4 Intermediate Results 35

4 Composing Contents and Knowledge in One Lattice 35

5 Conclusions 37
Introductory Summary

“O glücklich, wer noch hoffen kann, 
Aus diesem Meer des Irrtums aufzutauchen! 
Was man nicht weiß, das eben brauchte man, 
Und was man weiß, kann man nicht brauchen.”

*Faust*

— *Faust. Der Tragödie erster Teil.* (1808)
Johann Wolfgang von Goethe
1 Introduction

Determining when people first thought about their privacy is probably impossible, but it is reasonable to assume that people ever since sought to assess which actions lead to better privacy and which to worse. The ways of measuring privacy depend on the aspect of privacy, e.g., physical privacy or informational privacy. In this thesis, we focus on the latter, informational privacy.

Informational privacy of individuals has significantly gained importance after information technology has become widely deployed. Data, once digitalised, can be copied and distributed at negligible costs. This has dramatic consequences for individuals that leave traces in the form of personal data whenever they interact with information technology. The right of individuals for informational privacy, in particular to control the flow and use of their personal data, is easily undermined by those controlling the information technology. Privacy metrics can help individuals to evaluate the consequences of disclosing personal data and thus help them to make informed decisions.

The objective of this thesis is to study the measurement of informational privacy with a particular focus on scenarios where the individual discloses personal data to a second party, the data controller, which uses this data for re-identifying the individual within a set of others, the population. Several instances of this scenario are discussed in the appended papers, most notably Paper III which adds a time dimension to the scenario for modelling the effects of the time passed between data disclosure and usage. This extended scenario leads to a new framework for inter-temporal privacy metrics.

The common dilemma of all privacy metrics is their dependence on the information available to the data controller. The same information may or may not be available to the individual and, as a consequence, the individual may be misguided in his decisions due to limited access to the data controller’s information when using privacy metrics. The goal of this thesis is thus not only the specification of new privacy metrics, but also the contribution of ideas for mitigating this dilemma. However, a solution will rather be a combination of technological, economical and legal means than a purely technical solution, as discussed in Paper IV.

The remainder of the introductory summary is structured as follows. Section 2 introduces the basic terminology and concepts we use and refer to in this thesis and discusses related work. Section 3 defines the research questions addressed in this thesis, followed by a brief overview of research methods in Section 4. Section 5 summarises the contributions made in the appended papers to the research field of privacy metrics. Section 6 summarises the appended papers. Section 7 summarises the conclusions and outlines future work.

2 Background and Related Work

This section summarises the basic terms and concepts used in this thesis and refers to the related work. Since the central concept of this thesis is informational privacy, we will focus on this aspect first.
tional privacy, we start the background section with defining it. Though there is a large body of work using the term, there is no consolidated definition for privacy. One of the earliest definitions in modern scientific literature was given by Westin who has defined privacy as

“the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.” [45]

The German Federal Constitutional Court has defined privacy as the right of informational self-determination:

“[The] constitutional right of informational self-determination guarantees the permission of an individual to determine on principle himself on the disclosure and use of his personal information.” [20]

These notions have been widely adopted and we will thus not redefine informational privacy, but rather focus on the facets of it which are discussed in the appended papers.

2.1 Anonymity

Pfitzmann and Hansen [32] were maintaining a terminology overview for research in the informational privacy domain. The continuous updates of this terminology have been widely adopted in other scientific work since its first publication in 2000 [31]. Pfitzmann and Hansen define anonymity as follows:

“Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set.” [32]

The subjects in this definition are the actors in a network that either send or receive messages. If, for instance, the sending subject of a message cannot be determined among a set of other sending subjects, then the sending subject of this message is anonymous and the set of sending subjects in which the sender of the message is hiding is the anonymity set.

The size of the anonymity set is a measurement of anonymity. For sizes greater than 1, the sending subject is considered anonymous. An anonymity set size of exactly 1 means that the sending subject is identified.

Though not explicitly mentioned in [32], the anonymity set notion was most likely motivated by the way anonymity was established in Chaum’s DC network [9]. Chaum motivated his work with an artificial problem which he referred to as the dining cryptographers problem: one message should be sent and unblinded publicly without anyone learning who the sender was. His solution to this problem is the DC network, illustrated in Figure 1: each user, except one, encrypts an empty message with a one-time pad [44]. The remaining user encrypts his non-empty message with his one-time pad key and submits it to the network as well. The keys are chosen such that their bitwise sum is zero. Due to the one-time pad encryption and the specific way
Towards Inter-temporal Privacy Metrics

Figure 1: Illustration of a DC network. Bob sends the message 101 (encrypted with the key 110) which can be read by all three users, but neither Alice nor Charlie learn that Bob was the sender.

of decrypting, i.e., summing up all messages, it is not possible to discern the non-empty message and the empty ones, even after decrypting. Consequently, all users are equally likely the sender of the non-empty message, thus, all users form an anonymity set.

Though originally working in cryptography, Chaum pioneered the research on privacy-enhancing technology (PET). Besides the DC network, he also laid the foundations for another PET, the mix [8]. Mixes are network proxies with some interesting properties, see [15, 17] for recent surveys of different types of mixes and Figure 2 for an illustration of a mix. The term ‘mix’ refers to the mode of operation of these proxies whose main purpose it is to disturb (or ‘mix’) the relation between incoming and outgoing messages. A user who is sending a message to the mix can thus be certain (to the extend security is provided by the methods used, e.g., public-key cryptography [37]) that his message can not only be related to him but equally well to all other users who sent a message to the mix at the same time. The user is hiding in the anonymity set of all senders.

Mix users are not required to send messages continuously over time, i.e., over several mix rounds. This allows a number of attacks and though these attacks do not help revealing a sender-message relation, they are efficient in uncovering sender-recipient relations. The simplest attack is intersecting two or more anonymity sets which have been observed when a specific recipient got a message. This attack and variations of it are known as long-term intersection attacks [6]. If the accuracy of the result is only relevant up to a certain error margin, there is an even more efficient class of attacks, the statistical disclosure attacks [14, 26] which also take the probabilities into account of mix users in

Figure 2: Illustration of a mix. The mix operation makes it impossible to trace an incoming message through the mix and relate it to an outgoing message.
each anonymity set being the sender of a message to a specific recipient. In the mix round illustrated in Figure 2, we see that all senders submitted one message, but recipient 1 received two messages while recipients 2 got only one. The probability of sender A being the sender of a message to recipient 1 is thus \( \frac{2}{3} \) while the probability of A being the sender of a message to recipient 2 is only \( \frac{1}{3} \).

This probability distribution is not represented in the anonymity set notion. A more general notion is the degree of anonymity \([18, 39]\) which uses entropy in Shannon’s \([40]\) sense. The entropy of the probability distribution of individuals within the anonymity set can be interpreted as the average information an attacker would learn when a relation between sender and recipient is revealed to him. When the attacker has yet not discovered this relation, the entropy can be understood as a measure of the effort a successful attack on the mix would require in average. A third interpretation of the entropy is the (logarithm of the) average anonymity set size provided by the mix round.

We use a notion similar to the anonymity set in Paper I, simulate the effect of long-term intersection attacks and measure the sizes of the anonymity sets in Paper II, and generalise entropy-based metrics in Paper III.

### 2.2 Unlinkability

Pfitzmann and Hansen define unlinkability as follows:

“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.” \([32]\)

Unlinkability generalises anonymity, i.e., unlinkability describes the non-existence of a relation between two items of interest (possibly including subjects) and anonymity describes the non-existence of a relation between a subject and an item of interest (possibly another subject). Thus, every attack on anonymity is at the same time an attack on unlinkability, but anonymity metrics must be seen as special cases of unlinkability metrics. For instance, Steinbrecher and Köpsell \([41]\) use equivalence class membership relations to model the relations between items of interest. Probability distributions on these membership relations are used to model the degree of unlinkability.

The generalisation from subjects and messages to items of interest allows to abstract away from pure network-based scenarios, e.g., to measuring the linkability of personal data to individuals. Early work motivated by this scenario provided confidentiality metrics for statistical databases \([16, 20, 38, 42]\) and has been adapted \([10]\) to the unlinkability notion later. This new scenario was also the motivation for research on a branch of identity management systems that supports individuals in managing their personal data disclosure with regard to anonymity and unlinkability. The first solutions for such privacy-enhancing identity management \([7, 11]\) lead to a refinement of the
requirements [21] and culminated in several projects [22, 25, 30] funded by the European Community and other institutions [29].

Even though the objective of Paper I is proposing anonymity metrics, it uses several (‘formal scaling’) steps where no subject is involved. These intermediate steps can be understood as unlinkability metrics. In Paper III, we use anonymity metrics as an example which can easily be generalised to unlinkability metrics.

2.3 Information Asymmetry

The purpose of PETs, such as DC networks and mixes, is to minimise the flow of personal data\(^1\) on a technical level. The term ‘data minimisation’ coined in the privacy community and used in legislations such as the German ‘Bundesdatenschutzgesetz’ refers to the strategy of individuals choosing the smallest set of personal data to disclose to a data controller for achieving a specific goal, e.g., selling or purchasing goods of any kind or subscribing and using a service.

The effects of data minimisation can be better explained in the context of contract theory [3]: in a market with asymmetric distribution of information among the parties, the ones with more information have an incentive to use their information against the parties with less information [1]. A data controller sharing personal data of an individual with a third party without informing the individual creates an information asymmetry in which the controller possesses more information than the individual. The individual (benefits or) suffers from the consequences of the distribution of its data, i.e., the data controller creates (positive or) negative externalities [19] for the individual [24, 43]. PETs help to reduce information asymmetries, and thus negative externalities, by reducing the flow of data from the individual to the data controller.

This view suggests that data minimisation is not the only strategy for reducing information asymmetries, but rather has an equivalent sibling, increasing transparency, i.e., increasing the flow of information from the data controller to the individual. Bellotti and Sellen [5] discuss both strategies and the relation between them.

While PETs are effective at the time of the data disclosure, transparency-enhancing tools (TETs) may be effective either as predictive tools (ex-ante) before the data disclosure or as retrospective tools (ex-post) after the fact [23]. Predictive TETs can be used by individuals as decision support tools that anticipate the future consequences of a data disclosure while retrospective TETs inform the individuals about the de facto consequences of their data disclosures. Both types can be mixed so that consequences of past data disclosures can be taken into account for future data disclosures.

\(^1\)Broader definitions of PETs can be found in the literature that also capture technology such as access control, identity management, and even TETs (described later). In this thesis, we restrict the meaning of PETs to data minimisation.
The metrics in Paper I and III can be conceived as predictive TETs as well as the outcome of the experiment in Paper II and contract language in Paper IV.

2.4 Privacy Policy Languages

Privacy policies describe which personal data is collected by a data controller and what is going to happen with the data. The policies can be seen as predictive TETs which inform the individual about the intended usage of its data by the data controller. With this information, the individual can choose among data controllers by choosing the privacy policy that matches the individual’s preferences best. If preferences and policies are machine-processible, the choice of a data controller could even be automated, e.g., as part of an identity management system.

Several specifications of formal languages for privacy policies and preferences have been proposed. Among the first privacy policy languages and the first that became a standard was the platform for privacy preferences (P3P) [13, 36]. It defines a number of assertions, partly formalised and partly free text assertions, for the content category of the collected data, purpose of the data collection, whether the data is used in a personally identifiable way, and which parties receive the data [36]. In conjunction with P3P, ‘a P3P preference exchange language’ (APPEL) [12] has been proposed as the corresponding privacy preference language. The current working draft of the language allows to specify the P3P assertions acceptable for the individual by stating simple expressions.

Another privacy policy language is the enterprise privacy authorization language (EPAL) [4] which contributed a new concept to privacy policies, the obligation. According to the specification, obligations are additional steps to be taken when a certain action is performed [4].

EPAL has been discontinued and remained a submission for standardisation. A competing language for privacy policies, the extensible access control markup language (XACML), [27, 28] has become a standard earlier and has been found more flexible than EPAL [2].

As opposed to P3P, both EPAL and XACML are access control languages. This line has been continued in the PrimeLife policy language (PPL) [35] which extends XACML and specifies a number of subdialects, including an equivalent for privacy preferences. The access control rules in all of these three languages are optimised for efficient decisions on access rights, but are considerably less suited for the anticipation of access requests, i.e., the anticipation of what data is requested and when. Their use as a predictive TET is therefore limited.

In Paper IV, we propose a language for privacy contracts which is better suited as a predictive TET and provides a more general model for obligations than EPAL and XACML.

3 Research Questions

The following research questions are addressed in this licentiate thesis.
1. How can possible information asymmetries between data subjects and data controllers be reduced?

Possible solutions to this question are discussed in all four papers, mainly in the form of transparency-enhancing technologies or tools (TETs). Paper II, III, and IV provide predictive TETs and Paper I outlines a solution which can work as predictive as well as a retrospective TET. In addition, the canonical form of privacy option contracts in Paper IV can be seen as tool for data minimisation, thus as a PET.

2. How can we formally model the effect of uncertainty about the validity of personal data?

In Paper I and II, we make the assumption that privacy once lost cannot be regained. In reality, however, the longer the duration of storing data without validation or update, the greater is the likelihood of storing now invalid data. For the data controller, this means that the data becomes increasingly useless and the data subject thus implicitly regains its privacy. A mathematical model of this process gives interesting insights into how the expected usefulness of the data for the data controller and the expected infringement on the privacy of the data subject can evolve over time. We discuss a first approach of such a mathematical model in Paper III.

4 Research Methods

The main research method used in this thesis is the mathematical method [33]. Besides that, we used in Paper II the scientific method [34]. Both methods have a lot in common, but differ in their details. We briefly discuss them in this section.

Both methods consist of four steps. The first step in the mathematical method is understanding the problem. It is similar to the first step, characterisation of the problem, in the scientific method. In both methods, this step is to identify the unknown, the data and conditions, and to formulate them in suitable notation.

The second step in the mathematical method is devising a plan (also referred to as analysis), and in the scientific method, it is the formulation of the hypothesis. In both cases, the second step can be summarised as finding a connection between the data and the unknown. While the hypothesis in the scientific method usually is a mathematical model, the plan in the mathematical method can also be an analogy to a similar (or related) problem or a proof idea.

The third step in both methods can be different. In the mathematical method, it is carrying out the plan (synthesis), i.e., specifying the analogy or carrying out the proof. It can also be a deduction as the third step in the scientific method is, i.e., instantiating the hypothesis to predict the outcome of laboratory experiments or observations in the nature.

The forth step in the mathematical method is looking back. Its main purpose is to review the results from the previous steps, i.e., validating or verifying
the results, possibly simplifying the plan, and look for other applications of the same solution, i.e., generalising the solution. In the scientific method, the fourth step is mainly concerned with validating the predictions from the previous step in experiments.

In Paper I, III, and IV, we applied the mathematical method. The approach used in Paper I includes the reformulation of the problem, i.e., analysing the relation of disclosed data in terms of a similar problem, Formal Concept Analysis, for which a solution is known. In Paper III, the problem is to understand and model the uncertainty in privacy measurement and the solution is an analogy to the uncertainty model in option pricing. Once the analogy is exposed, the model could be adapted with small adjustments. In Paper IV, the same analogy is used to adapt the syntax and semantics of a language for specifying financial option contracts in order to specify a language for terms and conditions of data disclosures. It is even proved that contracts of that language converge to a canonical form.

In Paper II, we use the scientific method to carry out an experiment with data gathered from an operational anonymity service. The experiment is a simulation of two attacks, the intersection and the cross-section attack which are likely to be mounted on similar the data. Simulation has been chosen over collecting more detailed data that would have allowed real measurement in order to raise the acceptance among possible participants of the experiment.

5 Contributions

This section summarises the contributions made in this thesis.

1. Proposal of an ontological approach to analysing the linkability of individuals’ data. Paper I discusses how individuals (possibly with support of the data controller) can keep track of what personal data has been sent to data controllers. An ontological approach, Formal Concept Analysis, is used for modelling explicit and implicit data disclosure, i.e., also modelling data that can be derived from disclosed data. Keeping track of data disclosures is, besides minimising data disclosure, a way of increasing transparency and therefore reducing information asymmetries (Research Question 1).

2. Identification of a new realistic attacker model and analysis of its the effects on an operative anonymity service. In Paper II, a study about an attacker’s knowledge gain to be expected from the information provided by a law-abiding data controller in the context of Germany’s first implementation² of the European Data Retention Directive (2006/24/EC) is performed. The results suggest that a random attacker would greatly benefit from the retained data while the service would still provide enough anonymity to be of no considerable use in court when only

²The law has been enacted by the German government in December 2007 and nullified by the Federal Constitutional Court of Germany in March 2010.
relying on single requests to the anonymity service. The attacker model and the study contribute to Research Question 1 by informing users of anonymity services about the data that will be retained by the service provider and about the utility of that data with respect to different attacker models.

3. Identification of the analogy between data disclosure and option contracts known from economics. Paper III discusses the analogy of factors relevant for valuating the monetary value of financial options and the informational value of disclosed data. The analogy to financial options is called Privacy Options. This finding indirectly contributes to Research Question 2 by introducing a thought pattern that is used for making the next contribution.

4. Proposal of a model of the uncertainty in anticipating the future consequences of personal data disclosure today, i.e., for valuating informational privacy. Paper III uses the analogy between financial options and data disclosure and adapt the mathematical methods for accounting for the uncertainty in option pricing for modelling the uncertainty in valuating informational privacy. This finding directly contributes to Research Question 2. The anticipation of the future consequences can also be understood as enhancing the transparency and therefore reducing the information asymmetry (Research Question 1).

5. Specification of a privacy contract language which generalises the notion of time and events in privacy policies, defines permissions in terms of obligations instead of artificially separating these terms, and provides a canonical form for avoiding unintentionally creating covert channels by exploiting degrees of freedom in the language syntax. The contract language specified in Paper IV establishes the connection between privacy policy languages and Privacy Options and outlines a framework in which the uncertainty model introduced in the previous contribution can be applied. The contracts specify what is going to happen with the disclosed data, thus form a predictive TET and contribute therefore directly to Research Question 1. Initiating new risks to privacy is avoided, e.g., by specifying a canonical form. This can be seen as one aspect of minimising the disclosed data and thus reducing information asymmetries as a PET.

6 Summary of Appended Papers

Paper I – Linkability Estimation Between Subjects and Message Contents Using Formal Concepts

This paper presents an ontological approach to analysing linkability of data disclosure. The analysis uses a directed graph in which nodes represent data items, messages, pseudonyms of the data subject, and the data controllers.
Edges represent the relations between the node entities. These graphs are created from incidence matrices by means of Formal Concept Analysis.

**Paper II – Data Retention and Anonymity Services**

This paper presents empirical results on the anonymity that can be provided within the terms of the Germany’s first implementation of the European Data Retention Directive (2006/24/EC). This implementation required that, e.g., internet service providers retain traffic data and thus specified a model for data collection. It also specified an attacker model by regulating the conditions under which the data could be accessed and by whom. The study is performed using an operative anonymity service and real user data. The degree of anonymity has been examined under the attacker model specified in the implementation of the European Data Retention Directive and under less restrictive assumptions about the attacker.

**Paper III – Valuating Privacy with Option Pricing Theory**

This paper presents a probabilistic framework for measuring informational privacy. The actual measurement of informational privacy can be done by any reasonable metric that deals with attribute frequencies. The framework adds the time dimension, i.e., a model for uncertainty about the individual’s attributes induced by the development of the individual and the population of all individuals over time. This uncertainty model builds on the analogy to uncertainty models in option pricing known from economics. The framework is thus called Privacy Option.

**Paper IV – Towards a Formal Language for Privacy Options**

This paper presents a formally specified language, POL, for Privacy Options. It particularly focuses on the aspects of time in Privacy Options. It is similar to privacy policy languages such as P3P and PrimeLife’s PPL as both types of languages specify the rights and obligations of data controllers, but differs in its design as a contract model rather than an access control model. In POL, rights are defined as obligation contracts with the freedom to nullify them, i.e., rights and obligations are expressed with the same vocabulary, which allows defining obligations in a more flexible way than known from privacy policy languages. Besides, we present a canonical form of POL contracts which can be used to normalise the syntax of contracts and thus avoid unintentionally creating covert channels which could create new privacy risks otherwise.

**7 Conclusions and Future Work**

In this thesis, we have shown that privacy metrics are an important research area with challenging research questions. We have also presented several novel
approaches for privacy metrics. The most important contribution was to underpin the uncertainty about the validity of collected data induced by time with a mathematical model, the Privacy Option. Various parameters of this model can be specified in terms of contracts in the Privacy Option Language and privacy measurement can be defined as semantics of these contracts. Other contract semantics are conceivable, e.g., the data management semantics defined in Paper IV. We are convinced that future contract languages for privacy-enhancing identity management have to provide and cope (at least) with these two semantics equally well.

In Paper II, we have measured privacy with empirical data and with regard to a realistic (law-abiding) attacker model and an operational anonymity service. The conclusion is that, despite the fears in public, privacy can be preserved quite well against the law-abiding attacker. However, attacks on the same data beyond the restrictions of the law-abiding attacker may have more severe consequences in terms of privacy decrease.

Paper I extends the anonymity set notion: data is either related to an individual or not. One of the most interesting challenges is to generalise the methods such that imperfect knowledge can be modelled in all relations, i.e., imperfect knowledge about deducible data items as well as imperfect knowledge about the linkability of data to an individual. Paper III can be seen as a follow-up in this regard, since modelling imperfect knowledge is the main subject of this paper. In order to present the metrics in Paper III in a compact way, we limited our model of personal data to the most basic case, i.e., one attribute with two attribute values. In future work, we will generalise this data model. Paper IV outlines the plan for this work. Another opportunity for future work is the validation of the privacy metrics presented in Paper III with empirical data.

Future work with regard to the contract language specified in Paper IV may focus on new languages that wrap the contracts and put them in a context, e.g., specifying the contract parties. New languages that can be wrapped by the contract language are also conceivable, e.g., expression languages that refine the data model so that contracts can be made that allow, disallow, or enforce specific operations on the data.

References


Towards Inter-temporal Privacy Metrics


Towards Inter-temporal Privacy Metrics

Informational privacy of individuals has significantly gained importance after information technology has become widely deployed. Data, once digitalised, can be copied and distributed at negligible costs. This has dramatic consequences for individuals that leave traces in the form of personal data whenever they interact with information technology. The right of individuals for informational privacy, in particular to control the flow and use of their personal data, is easily undermined by those controlling the information technology.

The objective of this thesis is to study the measurement of informational privacy with a particular focus on scenarios where an individual discloses personal data to a second party, the data controller, which uses this data for re-identifying the individual within a set of others, the population. Several instances of this scenario are discussed in the appended papers, most notably one which adds a time dimension to the scenario for modelling the effects of the time passed between data disclosure and usage. This extended scenario leads to a new framework for inter-temporal privacy metrics.

The common dilemma of all privacy metrics is their dependence on the information available to the data controller. The same information may or may not be available to the individual and, as a consequence, the individual may be misguided in his decisions due to limited access to the data controller’s information when using privacy metrics. The goal of this thesis is thus not only the specification of new privacy metrics, but also the contribution of ideas for mitigating this dilemma. However, a solution will rather be a combination of technological, economical and legal means than a purely technical solution.