

The motivation and proposition of a privacy-enhancing architecture for operational databases

Kirsten Wahlstrom & Gerald Quirchmayr

School of Computer and Information Science
University of South Australia
PO Box 2471, Adelaide 5001, South Australia

kirsten.wahlstrom@unisa.edu.au

Abstract

To date, research has focussed on privacy from a wide perspective, enabling organisations to implement various technologies that contribute to privacy protection. However, in such approaches the perspective of the data subject is often obscured in favour of meeting technical design requirements. The privacy architecture proposed in this paper is premised upon a view of privacy as unique to each individual person, changing over time and maintained through the control of personal data. This conceptualisation of privacy is evidenced by the research literature as well as various legislation. This paper establishes a requirement for a Privacy-Enhancing Technology for operational databases through a consideration of the state of practice and the relevant literature. An architecture for such a technology, which acknowledges and supports this understanding of privacy and which is based upon the Use and Disclosure Principle of the Australian privacy regulation framework, is then proposed. The architecture extends its privacy protection capabilities from primary to secondary data processing applications.

Keywords: Privacy, regulation, database architectures.

1 Introduction

Knowledge Discovery over large sets of both operational and warehoused data is widely practiced. Strategists and theorists are increasingly supportive of the role that knowledge plays in the creation of competitive advantage (Davenport, DeLong & Beers 1998) and Burwen's 1998 study on the database and knowledge management market identified a market value of \$US8.8 billion in 1996 and predicted growth to \$US113.5 billion by 2002. Databases terabytes in size are the norm for some application domains (Fayyad, Piatetsky-Shapiro & Smyth 1996; Kohavi 1998) and such databases are a rich source of knowledge, with millions of records describing common phenomena. While providing significant market advantage, these advances have also led to increased

consideration of privacy concerns (Cavoukian 1998; Tavani 1996).

This paper motivates and proposes a privacy-enhancing architecture that enables an approach to protecting the privacy of operational data that can be propagated to datawarehouses. It distinguishes between primary data processing applications, for which opt-in exists, and secondary data processing applications, which were unforeseen at the point of data collection. The capacity of the privacy-enhancing architecture to propagate privacy protection to datawarehouses extends privacy protection to both primary and secondary data processing applications.

Commentators argue for equilibrium between innovation and regulation (Baase 1997; Kizza 1998; Registratiekamer & Information and Privacy Commissioner 1995; Weckert & Adeney 1997). With respect to secondary data processing applications, the equilibrium to be established is a means of exploiting the data resource while simultaneously complying with applicable legislative requirements and safeguarding the data privacy rights of individuals.

The motivation for the architecture proposed here emerges through a consideration of the role that privacy plays in individual peoples' lives (in as much as it is supported by the research literature and various legal systems) and a brief survey of both the research and the state of practice with respect to Privacy-Enhancing Technologies (PETs). The proposed architecture is thus premised on the privacy legislation as well as an acknowledgement of individual citizens' privacy preferences and any further privacy constraints required by organisations or obligated by regulatory bodies.

2 Privacy

Earlier work (Wahlstrom & Roddick 2000) established that privacy is a perception which differs from person to person, changes over time and emerges from a society's communication practices (Gavison 1984; Rachels 1975). Furthermore, people maintain these privacy perceptions by limiting their accessibility to others (Cavoukian 1998; Gavison 1984) and these efforts extend to exercising control over the information that describes them (Tavani 1996).

This view is supported by Kobsa (2001) who summarises the findings of various research projects investigating the actions that users take to protect their privacy during online transactions (see Table 1). It is also supported by

Acquisti and Grossklags (2005) who report that users often have insufficient information for making privacy-related decisions that lead to inadvertent deterioration of privacy. Finally, this view of privacy is acknowledged and supported under various legal systems, notably by the Australian privacy legislation via National Privacy Principle 2: Use and disclosure (Privacy Act 1988).

Privacy-related investigations	Average finding
Extremely of very concerned about divulging personal information online	70.5%
Extremely concerned about being tracked online	65.5%
Leaving web sites that required registration information	41%
Entering fake registration information	32%
Refraining from shopping online due to privacy concerns, or buying less	39%

Table 1: Kobsa's (2001) research findings

2.1 Privacy in legal systems

A discussion of privacy and its legal context is incomplete without a brief historical note. The earliest definitions of privacy were published in the late nineteenth century by Cooley (1888) and Warren & Brandeis (1890) in the United States. Cooley defined privacy as simply “the right to be let alone” and later Warren & Brandeis used prominent British cases to show that the common law supported this right (in Akdeniz, Clarke, Kelman & Oram 1997). This was followed by the reference to privacy in Article 12 of the United Nations’ Universal Declaration on Human Rights (1948) that stated, “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence.”

Over time, electronic record keeping and privacy concerns gradually became more prevalent, and in 1973 the US Department of Health, Education and Welfare published the now widely endorsed Code of Fair Information Practices. The fourth and fifth Practices are of significance to this paper as they further support the view that privacy perceptions and requirements differ from person to person and that people maintain privacy through controlling the information known about them:

1. Any organization creating, maintaining, using, or disseminating records of personally identifiable information must assure the reliability of the data for its intended use and must take precautions to prevent misuse; and
2. There must be a way for an individual to prevent personal information obtained for one purpose from being used for another purpose without his or her consent.

The Fair Information Practices informed subsequent (and now outdated) US privacy legislation and remain crucial

in that nation, where contemporary and comprehensive privacy legislation does not exist. Most importantly, however, these principles provided a foundation for the development of international privacy guidelines.

In 1980, the Organisation for Economic Cooperation and Development (OECD) issued its Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data. The Guidelines were motivated by a need to prevent incompatible and conflicting data protection laws among OECD member nations. The OECD’s intent was the harmonisation of member nations’ privacy legislation and the provision of a framework to simultaneously support data privacy and prevent disruption to international flows of data and information. Significantly, the Guidelines also acknowledge and support the individual’s prerogative to control of their personal information. The Guidelines have proven influential for data privacy protection, having been cited as the basis for most contemporary international agreements, national legislation and self-regulatory practices (Center For Democracy & Technology 1998).

In 1995, the European Parliament and the Council of the European Union released Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. It provides for the comprehensive and consistent privacy protection for all member state citizens. This consistent approach to privacy protection in turn facilitates the exchange of personal data between member states. Interestingly, in this otherwise rigorous approach to data privacy, data processing may proceed without the explicit consent of the use¹ if it is consistent with the data controller’s legitimate business interests, which is contradictory to the principles underlying the US’s Fair Information Practices and not entirely consistent with the principles outlined in Part Two of the OECD’s Guidelines. However, extensive options for other forms of use control over personal information are provided.

In Australia, the 1988 Privacy Act established the office and role of the Federal Privacy Commissioner. Of special interest to this project are the summaries of its complaint occurrence data provided in its annual reports. The office handled a steadily increasing number of complaints from 2000-2004 (see Figure 1) which may indicate either an increasing awareness of privacy concerns among Australian citizens, an increasing number of privacy violations, or a combination of both. The rising number of complaints provides a further motivation for the work undertaken here.

¹ Fischer-Hübner (2001) introduces the term ‘usee’ to represent a person described by data (frequently referred to as a ‘data subject’ in other work) whereas a user is the person utilising a software system. This convention is appropriate to a paper in which the distinction between the two groups is relevant and the notion of data being subject to the usee is pursued. Thus, this terminology is adopted throughout.

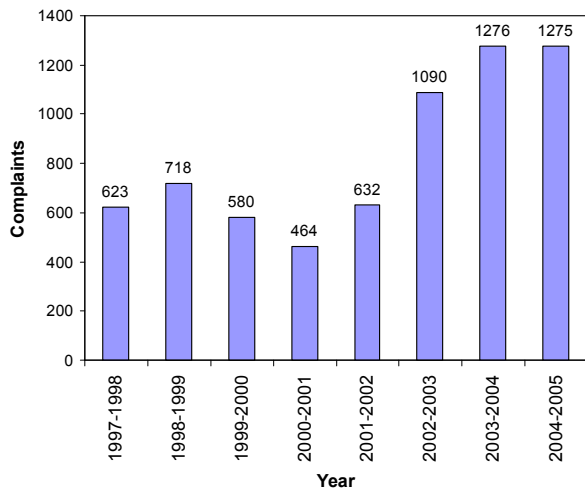


Figure 1: Privacy incidents in Australia

In Australia, the recent amendments (Office of Legislative Drafting 2001) to the Australian Privacy Act (1988) endorse ten National Privacy Principles, which were informed by the OECD's guidelines. This review of the various regulatory and legislative approaches to privacy protection has identified wide acknowledgement and support for the concept of privacy as unique to each person, who then maintains privacy by controlling the exposure of personal information.

Recently, PETs were envisaged as providing means by which the requirements emerging from this view of privacy may be met in technical solutions.

3 Investigating PETs

In 1995, the Registratiekamer of The Netherlands and the Office of the Information and Privacy Commissioner of Ontario published a paper examining the problems of data privacy protection within the context of information technologies. It proposed an approach to systems design which is appropriate to the contemporary legal requirements related to data privacy protection: the Privacy-Enhancing Technology.

It has been stated (Burkert 1997) that an important aspect of a PET is that it demands a return to social innovation: a consideration of identity and trust in the context of data privacy. This nexus of social concerns and systems design is particularly relevant where the user's informed consent is problematic or impossible to ascertain, as in the case of secondary data processing applications.

3.1 The literature

There are approaches to privacy protection which leverage computer security technologies as these are readily applicable to data protection and have seen many years of active research. However, the more recently emerged PETs distinguish conventional data security from data privacy protection (Burkert 1997).

This distinction emerged as security technologies do not protect data on behalf of users and thus they are not ideal tools for protecting privacy. As individuals have quite

distinct privacy preferences, a single data processing application may simultaneously violate one person's privacy and protect another's. Thus, privacy protection techniques are ideally integrated within data processing applications. This requirement has led to the proposal and development of systems which, while still leveraging conventional data security technologies, incorporate approaches to privacy protection in their design.

The definitive paper (Registratiekamer & Information and Privacy Commissioner 1995) investigates the possibility of establishing equilibrium between the needs of users who provide data and the organisations collecting data. It initially provides a broad consideration of information systems and of how identification of data (and by extension, people) is an inherent part of such systems. It shows how existing technologies can be applied to data privacy protection and proposes an approach to systems development that requires the definition of privacy requirements prior to systems design.

PETs can be either technological or organisational (Burkert 1997) and aim to reduce the collection, retention and processing of identifiable personal information without restricting the functionality of the information system (Borking & Raab 2001; Fischer-Hübner 2001).

As noted, approaches to this goal include some existing computer security technologies because of their capacity to protect the identity of users. Another widely investigated approach addresses the data mining stage of the Knowledge Discovery (KD) process.

3.1.1 Privacy preserving data mining

Many approaches to adapting data mining tools so that they provide greater privacy protection have recently emerged. Verykios, Bertino, Fovino, Provenza, Saygin and Theodoridis (2004) survey the state of the art and classify Privacy Preserving Data Mining (PPDM) techniques into three categories based upon the approach adopted: heuristics, cryptography or reconstruction.

Classification, association rule discovery and clustering are all data mining techniques for which heuristic privacy preservation approaches exist (Atallah, Bertino, Elmagarind, Ibrahim & Verykios 1999; Chang & Moskowitz 2000). Such approaches exploit the NP-Hardness of selective data modification and apply it to the data mining technique so that it preserves privacy.

PPDM approaches using cryptography (Clifton, Kantarcioglu, Lin & Zhu 2002; Du & Atallah 2001; Ioannidis, Grama & Atallah 2002; Lindell & Pinkas 2000) provide greater confidentiality of data assets in any data mining context where data is at risk of exposure, providing secure multi-party computation (SMC). Although the benefits of these approaches do coincidentally extend to users, consideration of their privacy preferences is not relevant to the provision of SMC.

Finally, reconstruction approaches (Agrawal & Srikant 2000; Agrawal & Aggarwal 2001; Rizvi & Haritsa 2002) apply perturbation followed by aggregation to provide privacy protection. Again, benefits coincidentally extend

to uses, but their privacy preferences are irrelevant under such a scheme.

Of these three categories, only those approaches based upon heuristics have the potential to incorporate individual's privacy preferences.

3.2 The state of practice

While the literature develops sound arguments to motivate and inform the development of PETs, technologies are also important because they ultimately dictate the degree to which privacy protection is accessible and practiced. Further, as a goal of this work is to motivate and propose a privacy-enhancing architecture for operational databases, an appraisal of available technologies is a useful component of this investigation.

A preliminary, informal survey revealed an extensive and varied range of technologies which protect privacy. While many have strong design correlations to data security technologies, there are also a substantial number without. This review focuses on the latter and categorises the technologies according to the privacy protection paradigm they implement:

- Anonymity;
- Pseudonymity;
- User profile management; and
- Web content filtering.

Many of the available technologies combine some or all of these approaches. The following discussion surveys and describes each type of PET.

3.2.1 Anonymity

In practical terms, anonymity occurs when a user's identity cannot be ascertained (Goldberg, Wagner & Brewer 1997). An example of an anonymous transaction is one in which neither participant recognises or knows anything about the other. Anonymity services have been developed for email (anonymous re-mailers), web browsing and e-commerce (ecash).

A significant disadvantage of anonymity is that accountability becomes problematic and therefore anonymity services are exploitable by those engaged in criminal activities (Clarke 2001; Goldberg, Wagner & Brewer 1997). Examples of anonymity services include Anonymizer, Anonymity 4 Proxy, CGIProxy (Marshall 2002), Freedom, Crowds (Reiter & Rubin 1998), GoProxy, IDzap, iPrivacy, PonoI, Privacy Companion, Siege Surfer and The Cloak.

3.2.2 Pseudonymity

Pseudonymity provides a compromise between anonymity and accountability. A user employing a pseudonym engages in communications and transactions without revealing their identity (Fischer-Hübner 2001). Ideally a trusted third party (usually the pseudonymity service provider) maintains a cross-reference between users and pseudonyms and under extenuating circumstances a user's identity can be ascertained.

When using the web, a user can elect to establish a pseudonym and may endow it with identifying information. The pseudonym is then used on their behalf to conduct online transactions and over time it develops a record of trustworthiness (or otherwise). Examples of pseudonymity services are IBM's idemix (Camenisch & Lysyanskaya 2001), iPrivacy, Rewebber, ZixCharge and the now unavailable Lucent Personalized Web Assistant (Gabber, Gibbons, Matias & Mayer 1997).

3.2.3 Profile managers

Profile managers enable users to explicitly control the sharing of personal data as well as providing pseudonymous transactions. Examples of profile managers are Novell's digitalme and the W3C's Platform for Privacy Preferences Protocol (World Wide Web Consortium 2001).

3.2.4 HTML filters

PETs for filtering HTML remove content from web pages which is likely to violate privacy. The requirement for such technologies is premised upon the often surreptitious web user profiling services which disseminate executable web content to track user clickstreams and to collect other user information. This usually occurs without the user's knowledge of the nature of the information being collected, the primary purpose of the collection and any possible secondary uses to which the information may be put.

This partially violates the Fair Information Principles (US Department of Health, Education and Welfare 1973) and is in direct transgression of the OECD's guidelines (Organisation of Economic Cooperation and Development 1980), the EU's Directive 95/46/EC (European Parliament & the Council of the European Union 1995) and the Australian Privacy Act (1988). Web user profiling is achieved via technologies such as web bugs and cookies.

A web bug is commonly a 1x1 pixel transparent image (in GIF format) placed in a web page or a banner advertisement by a third party marketing organisation. Web bugs surreptitiously execute code to collate profiles of web users accessing the web page. This is to the benefit of third party marketing organisations collecting volumes of personal information for sale to clients, at the nominal and relatively insignificant cost of implementing the web bug. However, it is to the detriment of web users whose personal information is surreptitiously collected and whose privacy is therefore at risk.

A cookie is a text file of information unique to a web user. Web sites with executable content save cookies on web users' hard drives via their web browsers (Gunning 1997). One cookie may be used by several organisations, each of which receives a copy of the information in the cookie file. Cookies typically record the last web site visited by the user, their time zone and their IP address, as well as system configuration information. Cookies also record any information the user provides in online forms (for example, their email address or name). The cookie

may be returned to the web site immediately, later, or left on the user's hard drive indefinitely.

Contemporary web browsers permit users to opt out of accepting cookies. However, making an informed choice is difficult, as web site operators do not provide notice as to what information a cookie collects, the purpose of the collection, nor any secondary purposes to which the collected information will be put. Significantly, the user is restricted in their decision-making in those cases where web sites are un-viewable without a cookie being set.

As cookies are set by banner advertisements most privacy-enhancing HTML filters focus on eliminating such advertising. Examples include AdDelete, AdSubtract Pro, Guidescope and Internet Junkbuster Proxy.

There are also many tools for managing cookies once they are installed on the user's hard drive. Examples include Anonymity Proxy 4, CGIProxy (Marshall 2002), GoProxy, IE Clean and Naviscope.

3.3 Findings

This brief investigation of the research literature and the current state of the technology establishes a requirement for, and motivates, a privacy-enhancing architecture for databases which acknowledges uses' diverse privacy preferences and which is also premised upon applicable privacy legislation.

4 Privacy requirements analysis

Five of the ten National Privacy Principles are relevant to secondary data processing applications and they are now considered in terms of scope for meeting the goals of this paper.

4.1 Collection

The Collection Principle defines the requirements to be satisfied by an organisation collecting data from individuals. Specifically, that the purpose of the data collection and the use's right to access personal data are specified and that organisations must not collect data in an unreasonably intrusive way.

In the case of operational data used in KD applications, the requirement for organisations to provide notice of the purpose of collection presents an intractable problem. This is because its purpose is not known until KD is complete (Cavoukian 1998) and KD cannot be stated as the purpose of collection as its function is the generation (collection) of additional use information.

Adequate notice of the use's right to access personal data is readily addressed with improved business practices. However, a PET that exploits existing email systems to send uses a copy of the data that an organisation holds in order for the use to make amendments or deletions could offer value in terms of improving data access for uses.

Furthermore, a privacy-enhancing approach may be adopted to improve organisations' capacities to fulfil the requirement that they must not collect data in an

unreasonably intrusive way. As noted in the section *Privacy*, uses have a diversity of privacy preferences. Consider a use whose privacy preferences led them to deliberately withhold information subsequently inferred in a KD application. In such a case, the inferred information will have been collected in an unreasonably intrusive way.

However, if a use's privacy preferences are collected at the same time as their data, then a privacy-enhancing architecture may ensure their data is used according to these preferences. For example, a use may withhold consent for their data to be used in secondary data processing applications. In such a case, their data may be omitted from such applications, inferred information would not apply to this use and their privacy would therefore be protected.

This suggests that uses' privacy preferences may be leveraged to facilitate the organisation of operational data so that a use's capacity to control access to their data is extended. Personal data designated as highly private by uses may easily be omitted from secondary data processing applications.

4.2 Use and Disclosure

The Use and Disclosure Principle states the circumstances under which the use and disclosure of personal information are prohibited: firstly, in those circumstances where an organisation has not requested and received prior consent and, secondly, use and disclosure are prohibited for sensitive information. A privacy-enhancing approach can be envisaged which satisfies both of these requirements.

As outlined above, if uses' consent and disclosure preferences are known then data may be organised according to these preferences. This then enables informed use and disclosure of private data which, in turn, facilitates more comprehensive compliance with the legislation and with uses' privacy preferences.

If operational data have known consent preferences, then a privacy-enhancing architecture that leverages these preferences could prevent non-consensual use of personal data. This can be achieved by sorting operational data into two sets: the first set with consent for secondary purposes granted and the second set without consent. Then only those data in the first set may be used for secondary data processing applications.

Furthermore, where legacy data is merged with operational data the consent preferences for the operational data may be analysed. This would enable consent preferences consistent with those in the operational data to be retrospectively applied to the legacy data.

Use data could also be organised according to uses' disclosure preferences where data that the use designates as 'not private' may be freely disclosed and data designated 'highly private' may not be disclosed. In addition, if disclosure preferences for operational data are known, it follows that disclosure preferences for aggregate data, KD patterns and the output of other

secondary data processing applications are calculable. Thus, such outputs are also protected under the disclosure prevention approach described here.

4.3 Data Quality

The requirement for data quality as outlined by the Data Quality Principle is that an organisation must take reasonable measures to make sure the data it collects, uses or discloses is accurate and complete.

This requirement is consistent with the widely practiced KD process that incorporates a stage of data pre-processing to remove noise and to ensure that missing data are managed consistently. Other enhancements to data quality are typically applied at data collection. Thus, a privacy-enhancing approach would not offer benefit with regard to data quality.

4.4 Data Security

The Data Security Principle requires that organisations take reasonable steps to protect personal information from misuse and loss as well as from breaches of computer security (for example, unauthorised access or modification). It also requires that personal information no longer used for the purposes stated at data collection be permanently anonymised or destroyed.

Approaches to protecting personal data from security threats include encryption, access control and (to a limited extent) auditing. Loss prevention, anonymisation and destruction of data are best realised via appropriate systems administration and security practices.

However, a privacy-enhancing architecture would contribute to preventing misuse of personal information. As noted above, if useses' consent and disclosure preferences are known then it is possible to organise data according to these preferences. This contributes to the prevention of misuse by controlling both disclosure to users and the use of data in secondary data processing applications.

4.5 Openness

The requirements for openness as specified by the Openness Principle are that an organisation informs individuals of what kind of information it holds, for what purposes the information is held and how it is collected, retained, used and disclosed. When secondary data processing applications are conducted over personal information, this Principle requires organisations to inform individuals of new information relating to them, which is a cumbersome and expensive undertaking at best.

While strongly related to business practices, it may be possible to minimise the impact of this requirement with the privacy-enhancing approach envisaged above. If useses indicate their disclosure preferences at data collection, then disclosure preferences for any information inferred by subsequent secondary data applications would be calculable. Should such an application infer highly private new information, organisations may opt to preclude it from further use.

Such preclusion minimises the impact of the Openness Principle.

4.6 Privacy requirement identification

This review of the privacy requirements identifies the Use and Disclosure Principle as offering the greatest potential for enabling this project's goals. Thus, this is the privacy requirement that the architecture is based on. However, we envisage that some of the privacy requirements related to the Collection, Data Security and Openness principles will be met coincidentally.

5 The envisaged architecture

A privacy-enhancing architecture to address the requirements of the Use and Disclosure Principle may be based on privacy preferences provided by useses. This obliges useses to provide consent for data usage² and more importantly, it also obliges useses to indicate at data collection which data they consider private and also to what extent these data are private (for example, highly, moderately or not private). While further burdening the individual these obligations, once satisfied, would facilitate better-informed and selective use and disclosure of personal data.

A complexity so far overlooked in the research literature is that useses' privacy preferences may change over time. To date, both technical and theoretical approaches to privacy protection assume either that privacy preferences remain stable or rely upon the usee for ad hoc amendment of their preferences. These approaches lead to inconsistencies in the correctness of privacy preferences and thus also to inconsistencies in their applicability. Clearly, approaches to privacy protection that use AI or statistical approaches for adapting data to apparent changes in accumulated privacy preferences will have an inherent error. Nevertheless, such adaptability is more realistic, potentially more reliable in terms of correctness and applicability, and therefore remains desirable.

Thus, a rigorous approach to privacy protection would require that useses indicate their privacy preferences at data collection and would incorporate a technical approach to automatically adapting these preferences to apparent trends in privacy preferences over time. Any approach that does not engage useses in the consideration of their privacy preferences imposes privacy assumptions on useses and thus fails to provide adequate adaptability to usee contexts.

Thus far, the architecture has two types of usee-defined privacy preferences: disclosure and consent. Disclosure preferences defined by useses will be leveraged by operational databases and secondary data processing applications to control the disclosure of private data. In a similar way, useses' consent preferences will be leveraged to organise data so that non-consensual use is prevented. The approach suggested here constrains operational querying and secondary data processing applications

² This obligation is already imposed by the amended Privacy Act 1988.

according to the disclosure and consent preferences specified by users. From this point, precise terminology is used. In order to reflect this semantic difference, the phrase *disclosure and consent preferences* is employed in user contexts and *disclosure and consent constraints* is used in system contexts.

The disclosure and consent constraints may be specified in many ways. For example, they could apply to individual datum or to the collection of data that a single user provides. Further investigation of an appropriate granularity for these constraints is conducted concurrently to proposing conceptual models for use and disclosure in the forthcoming sections.

In addition to users' privacy preferences, disclosure and consent constraints may also be derived from legislative requirements, organisations' policies and any other regulatory body's requirements (for example, a nation's privacy commissioner). Firstly, legislative requirements may be used to organise operational data. For instance, the confidentiality of medical records is specifically guaranteed by various legislation in Australia (examples include the Australian Capital Territory's Health Records (Privacy and Access) Act 1997; New South Wales's Health Records and Information Privacy Act 2002 and Privacy and Personal Information Protection Act 1998; and Victoria's Health Records Act 2001 and Information Privacy Act 2000). Secondly, disclosure and consent constraints may be derived from individual organisations' privacy policies. This is of special relevance to those organisations implementing privacy policies that extend the privacy requirements specified in the legislation. For example, an organisation may elect to define all personal data relating to online sales as highly private. Finally, disclosure and consent constraints may be derived from the requirements of any other relevant regulatory body. For example, the Office of the Privacy Commissioner may find it necessary to require a specific organisation to define all personal data relating to insurance records as highly private.

5.1 Conceptual model for use

The Use and Disclosure Principle requires that data be omitted from any secondary data processing applications not consented to by the user. Users' consent preferences may be gathered at data collection and then used as consent constraints within the architecture. The use of private data may then be regulated: data may be organised according to consent constraints and data without consent for secondary data processing applications may be omitted from such applications. The only useful values for the consent constraint are binary indications of whether the user consents to their data being used for secondary purposes or not.

This model is useful in situations where legacy data is merged with operational data. In such a case, the operational data's consent constraints may be analysed to identify any inconsistencies in applicability. For example, there may be a consistency such that an overwhelming majority of users above a specific age withhold consent. On the strength of such a consistency, the organisation

holding the data may determine to withhold consent for all merged data meeting the same age criterion.

5.2 Conceptual model for disclosure

When disclosure constraints are known, operational data may be organised according to these constraints so that all data with any one constraint value is in the same set. Thus, if three disclosure constraint values exist (for example not private, moderately private and highly private) then three sets of operational data can be envisaged.

Consider a primary data processing application executed over operational data enhanced with disclosure constraints. The calculation of correlated disclosure constraints for any output is trivial. Should the output have a disclosure constraint value of 'highly private' its disclosure to users can be prevented if the system has a disclosure threshold which is set so that only moderately private information is disclosed. On the other hand, if the output has a 'not private' disclosure constraint and the disclosure threshold is set to permit moderately private information, then the output will be disclosed to users. Thus, output is protected from disclosure when its disclosure constraints are higher than that prescribed by the disclosure threshold.

5.2.1 A limitation and amendment

The example immediately above illustrates a disclosure model that requires the users' disclosure preferences to be known. A limitation of this approach is that it provides no adaptability for different users. If data is withheld for one user, it is withheld for all users. This is ideal from a privacy protection perspective, but it limits the functionality available to the organisation and thus does not achieve the equilibrium between user and organisation aspired to in this paper. Thus, an extension to this model is necessary so that greater functionality can be leveraged by the organisation.

An informative approach may be observed in Multi-Level Secure database architectures in which users are authorised to view data at a specified security level (Denning 1986; Elmasri & Navathe 1994; Pfleeger 1997). This approach is readily adapted to a privacy context so that individual end users are authorised to view private data at or below a disclosure threshold.

Under this architecture, applications execute over the data and output is delivered in sets corresponding to the disclosure constraints. Users are assigned disclosure privileges that correspond to the disclosure constraints. Very few users are privileged to view the most private data while all users are privileged to view data that is not private. Thus, individual users are enabled to view data at or below a specific disclosure threshold.

One of the strengths of this architecture is that it can be envisaged with any number of constraint values. For clarity, so far we have discussed three constraints and they have been labelled Highly private, Private and Not private, but any number of constraints may be leveraged in a given operational context. Indeed, the optimal

number of constraints may only be defined with reference to the relevant operational context.

Thus, this model may be envisaged with 2..N disclosure constraint values, however technical and human cognition limitations restrict the specification of constraint values to a much more modest range. Thus, the discussion has so far focussed on a range of three constraints. The preferences indicated by uses may be stored as integers for computational efficiency.

5.2.2 Disclosure constraint benefits

Disclosure constraints offer two key benefits. Firstly, trends in disclosure preferences can be used to update the disclosure constraints. As new data and their disclosure constraints are added to the operational database, emerging trends in the constraints will be identifiable.³ Secondly, when a KD process (or some other secondary data processing application) is executed over the data, disclosure constraints for new patterns (or any other product) are calculable. Any subsequent disclosure will then be regulated by the disclosure constraints of the new patterns.

6 Emerging problems

This architecture poses two major problems. Firstly, this approach is useful in any context where uses' privacy is of relevance as it empowers individuals to define their own privacy preferences based on their specific contextual requirements. However, while offering great benefits in terms of adaptability to uses' privacy contexts, the provision of these preferences by uses at data collection may prove onerous and are therefore susceptible to criticism. On the other hand, if disclosure and consent constraints do not exist for the operational data, the products of secondary data processing applications would have no constraints except for those which can be derived from the legislation and from organisations' privacy policies. Given the diversity of changing privacy preferences in the usee population, the collection of their privacy preferences remains desirable.

Secondly, as outlined above, when secondary data processing applications are executed over operational data with known disclosure constraints, then similar constraints for the outputs of these applications may be calculated. However, using disclosure constraints in this way establishes a problematic situation in which any

inaccurate constraint values in the operational data will cause inaccuracies in any calculated constraints.

To illustrate, consider a situation in which legacy data are merged with operational data and KD is then executed over the entire data set. The disclosure constraints of the patterns generated by the KD process will have been calculated from the constraints present in the merged data. Should the legacy data have had inaccurate constraint values, then the system would have calculated constraints for the KD patterns which cannot be accurate. If inaccurate disclosure constraints exist for any product of secondary data processing applications, private information is at risk of inappropriate use and disclosure.

That derived disclosure constraints may be problematic in this way gives rise to a requirement for high levels of transparency and proactive system management so that the risk of inadvertent disclosure of private information and, conversely, the risk of privacy scope creep are minimised.

7 Architecture summary

To summarise, uses provide sensitive data and have diversities of changing privacy preferences. It is proposed here that they could also specify disclosure and consent preferences relevant to the data they provide. To ensure adequate privacy protection under both primary and secondary data processing applications, these data and preferences should be integrated with operational databases. The privacy preferences specified by uses are utilised in the architecture as disclosure and consent constraints.

Consent constraints can be leveraged to create a subset of operational data with consent for unforeseen secondary data processing applications granted. Such applications may then be freely executed over this subset of operational data.

Similarly, disclosure constraints are leveraged to create subsets of operational data for disclosure to privileged users. Disclosure constraints for any products of secondary data processing applications (for example, any new patterns produced by a KD application) may be calculated from the disclosure constraints of operational data.

The architecture is best realised with accurate disclosure and consent constraints, as this will facilitate accurate use and disclosure of operational data, providing higher levels of integrity with regard to the privacy preferences of uses. However, if accurate measures of uses' disclosure and consent preferences are omitted, the architecture still provides an effective technique for integrating legislative and organisations' privacy requirements with an operational database and therefore, by extension, with primary and secondary data processing applications.

8 Conclusion

This paper established an understanding of privacy as relevant to individual people: privacy differs from person to person, it changes over time and people establish and maintain their privacy by controlling which and how

³ To illustrate, consider a specific type of data (for example, religious affiliation) which has historically had an average disclosure constraint value ranging from not private to moderately private. In newer data, religious affiliation may have an average disclosure constraint value of highly private. Clearly, a trend has begun to emerge in the disclosure constraints associated with uses' religious affiliations. In such a case, it may be appropriate to amend the organisation's policy so that all religious affiliation data is prevented from being disclosed and to update all disclosure constraints accordingly.

much of the data that describes their lives is known. It presented evidence from the research literature and various legislative approaches in support of this view.

Through an examination of the PET research literature and the state of practice, it established a requirement for a PET which provides sufficient adaptability so that it may be used to represent the diversity of privacy preferences inherent to the understanding of privacy on which this work is premised. It then proposed a privacy-enhancing architecture which acknowledges and supports this view of privacy. The architecture incorporates user-defined privacy preferences into operational databases in such a way that the privacy protection it offers is extended to primary and secondary data processing applications. It is also premised on the Use and Disclosure Principle which underlies the Australian Privacy Protection Framework. Future work will attempt to demonstrate the viability of the architecture through a proof-of-concept prototype.

9 References

- Acquisti, A. & Grossklags, J. (2005): Privacy and rationality in individual decision making. *IEEE Security and Privacy*, **3**(1):26-33.
- Agrawal, D. & Aggarwal, C. C. (2001): On the design and quantification of privacy preserving data mining algorithms. *Proc. of the 20th ACM Symposium on Principles of Database Systems*. Santa Barbara, California, US, 247-255, ACM Press.
- Agrawal, R. & Srikant, R. (2000): Privacy-preserving data mining. *Proc. of the ACM SIGMOD Conference on Management of Data*. Dallas, Texas, US, 439-450, ACM Press.
- Akdeniz, Y., Clarke, O., Kelman, A. & Oram, A. (1997): Cryptography and Liberty: Can the Trusted Third Parties be Trusted? A Critique of the Recent UK Proposals. *The Journal of Information, Law and Technology* (2).
- Atallah, M. J., Bertino, E., Elmagarmid, A. K., Ibrahim, M. & Verykois, V. S. (1999): Disclosure limitation of sensitive rules. *Proc. of the IEEE Knowledge and Data Engineering Workshop*. Chicago, Illinois, US, 45-52, IEEE Computer Society Press
- ACT legislation register (1997): Health Records (Privacy and Access) Act. <http://www.legislation.act.gov.au/>. Accessed 20 Oct 2006.
- Baase, S. (1997): *A gift of fire: social, legal, and ethical issues in computing*, New Jersey, Prentice-Hall.
- Borking, J. & Raab, C. (2001): Laws, PETs and other technologies for privacy protection. *The Journal of Information, Law and Technology* (1).
- Burkert, H. (1997): Privacy-Enhancing Technologies: typology, critique, vision. In *Technology and privacy: the new landscape*. 125-142. P. E. Agre & M. Rotenberg (eds). The MIT Press, Massachusetts, US.
- Burwen, M. P. (1998): Database solutions: a white paper covering market, competitive and user trends and issues for data warehousing, decision support, business intelligence, knowledge management. Palo Alto Management Group, Inc.
- Camenisch, J. & Lysyanskaya, A. (2001): An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation. *EUROCRYPT 2001: Proc. of the International Conference on the Theory and Application of Cryptographic Techniques*. Innsbruck, Austria. 93-118.
- Cavoukian, A.: Data Mining: staking a claim on your privacy. <http://www.ipc.on.ca/>. Accessed 27 Apr 2000.
- Center for Democracy and Technology: Privacy basics: the OECD guidelines. <http://www.cdt.org/>. Accessed 6 February 2002.
- Chang, L. & Moskowitz, I. S. (2000): An integrated framework for database inference and privacy protection. *Data and Applications Security, IFIP Working Group 11.3*. School, The Netherlands, 161-172, Kluwer Academic.
- Clarke, R. (2001): Introducing PITs and PETs: Technologies affecting privacy. *Privacy law and policy reporter* **7**(9):181-183.
- Clifton, C., Kantarcioglu, M., Lin, X. & Zhu, M. Y. (2002): Tools for privacy preserving distributed data mining. *SIGKDD Explorations* **4**(2).
- Cooley, T. (1888): *A Treatise on the Law of Torts or the Wrongs which arise independent of contract*, 2nd ed. Callaghan, Chicago, USA.
- Davenport, T. H., DeLong, D. W. & Beers, M. C. (1998): Successful knowledge management projects. *Sloan Management Review* **39**(2):43-57.
- Denning, D. (1986): An Intrusion-Detection Model. *IEEE Transactions on Software Engineering* **SE-13**(2):222-232.
- Du, W. & Atallah, M. J. (2001): Secure multi-party computation problems and their applications: A review and open problems. *Proc. of the 2001 Workshop on New Security Paradigms*. Cloudcroft, New Mexico, 13-22, ACM Press.
- Elmasri, R. & Navathe, S. B. (1994): *Fundamentals of database systems*, 2nd edn. Benjamin Cummings Publishing Company, Inc, California, US.
- European Parliament & the Council of the European Union. (1995): Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995. *Official Journal of the European Communities* (L. 281):31-39.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996): From Data Mining to Knowledge Discovery in databases. *AI magazine* **17**(3):37-54.
- Fischer-Hübner, S. (2001): *IT-security and privacy: design and use of privacy-enhancing security mechanisms*, Springer-Verlag, Berlin, Germany.
- Gabber, E., Gibbons, P., Matias, Y. & A. Mayer, A. (1997): How to Make Personalized Web Browsing Simple, Secure, and Anonymous. *Proc. of Financial*

- Cryptography* 97. Anguilla, UK, 17-31, Springer-Verlag.
- Gavison, R. (1984): Privacy and the limits of law. *Yale law journal* **89**: 421-71.
- Goldberg, I., Wagner, D. & Brewer, E. (1997); Privacy-enhancing Technologies for the Internet. *Proc. of 42nd IEEE International Computer Conference: Hot Systems, Cool Software*, San Jose, California, US, 103-109, IEEE Computer Society Press.
- Gunning, P. (1997): Evaluating privacy for Internet users and service providers. *Privacy law & policy reporter* **4**(4):67-70.
- Ioannidis, I., Grama, A. & Atallah, M. (2002): A secure protocol for computing dot-products in clustered and distributed environments. *Proc. of the International Conference on Parallel Processing*. Vancouver, British Columbia, Canada, 379-384, IEEE Computer Society Press.
- Kantarcioglu, M. & Clifton, C. (2002): Privacy preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering* **16**(9):1026-1037.
- Kizza, J. M. (1998): *Ethical and social issues in the information age*, Springer-Verlag Inc, New York, US.
- Kobsa, A. (2001): Tailoring privacy to users' needs. *Lecture Notes in Artificial Intelligence 2109*, eds M. Bauer, P. J. Gmytrasiewicz & J. Vassileva, Springer-Verlag, 303-313.
- Kohavi, R. (1998): Crossing the chasm: from academic machine learning to commercial Data Mining. *Invited talk at ICML-98*.
- Lindell, Y. & Pinkas, B. (2002): Privacy preserving data mining. *Journal of Cryptology* **15**(3):177-206.
- Marshall, J.: CGIProxy. <http://www.jmarshall.com/>. Accessed 24 April 2002.
- NSW Lawlink (2002): Health Records and Information Privacy Act. <http://www.lawlink.nsw.gov.au/>. Accessed 20 Oct 2006.
- NSW Lawlink (1998): Privacy and Personal Information Protection Act. <http://www.lawlink.nsw.gov.au/>. Accessed 20 Oct 2006.
- Office of Legislative Drafting, 2001, The national privacy principles in the privacy amendment (private sector) Act 2000 as at 10/01/2001, Human Rights and Equal Opportunity Commission.
- Organisation of Economic Cooperation and Development: Guidelines governing the protection of privacy and transborder flows of personal data. <http://www.oecd.org/>. Accessed 1 May 2000.
- Pfleeger, C. P. (1997): *Security in computing*, Prentice-Hall, New Jersey, US.
- Privacy Act 1988: Act No. 119 of 1988 as amended. <http://scaleplus.law.gov.au/>. Accessed 19 Dec 2001.
- Rachels, J. (1975): Why privacy is important. *Philosophy and public affairs* **4**(4):323-333
- Registratiekamer (Netherlands) & Information and Privacy Commissioner (Ontario, Canada): Privacy-Enhancing Technologies: the path to anonymity, vol 1. <http://www.ipc.on.ca/>. Accessed 19 February 2002
- Reiter, M. K. & Rubin A. D. (1998): Crowds: anonymity for web transactions. *ACM transactions on information and system security (TISSEC)* **1**(1):66-92.
- Rizvi, S. J. & Haritsa, J. R. (2002): Maintaining data privacy in association rule mining. *Proc. of the 28th International Conference on Very Large Databases*. Hong Kong, China, 682-693, Morgan Kaufmann Publishers.
- Tavani, H. (1996); Computer matching and personal privacy: Can they be compatible? In *Proc of the Symposium on Computers and the Quality of Life*. Philadelphia, Pennsylvania, US, 97-101, ACM Press.
- United Nations: Universal Declaration on Human Rights. <http://www3.itu.int/>. Accessed 30 July 2002.
- US Department of Health, Education and Welfare: Records, computers and the rights of citizens: report of the secretary's advisory committee on automated personal data systems. <http://aspe.os.dhhs.gov/>. Accessed 6 February 2002.
- Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y. & Theodoridis, Y. (2004): State-of-the-art in Privacy Preserving Data Mining. *SIGMOD Record* **33**(1):50-57.
- Victorian Law Today (2001): Health Records Act. <http://www.dms.dpc.vic.gov.au/>. Accessed 20 Oct 2006.
- Victorian Consolidated Legislation (2000): Information Privacy Act. <http://www.austlii.edu.au/>. Accessed 20 Oct 2006.
- Wahlstrom, K. & Roddick, J. (2000): On the impact of Knowledge Discovery and Data Mining. *Conferences in research and practice in information technology: Second Australian Institute of Computer Ethics Conference (AICE2000)*. **1**:22-27.
- Warren, S. D. & Brandeis, L. D. (1890): The right to privacy. *Harvard Law Review* **4**(5):193-220.
- Weckert, J. & Adeney, D. (1997): *Computer and information ethics*, Greenwood Press, Connecticut, US.
- World Wide Web Consortium: The Platform for Privacy Preferences 1.0 (P3P1.0) specification: W3C working draft 28 September 2001. <http://www.w3.org/>. Accessed 21 January 2002.