Controlling Inference: Avoiding *P*-level Reduction During Analysis

Adepele Williams and Ken Barker

Computer Science Department
University of Calgary, 2500 University Drive
Calgary, Alberta, Canada, T2N 1N4
{awilliam, barker}@cpsc.ucalgary.ca

This paper presents a concept hierarchy-based approach to privacy preserving data collection for data mining called the P-*level model. The* P-*level model allows data providers to divulge information at any chosen privacy level (*P-*level), for any attribute. Data collected at a high* P-*level signifies divulgence at a higher conceptual level and thus ensures more privacy. Providing anonymity guarantees, such as* k-*anonymity, prior to release can further protect the collected data set from privacy breaches due to linking the released data set with external data sets. However, the data mining process, which involves the integration of various data values, can constitute a privacy breach if combinations of attributes at certain* P-*levels result in the inference of knowledge that exists at a lower* P-*level. This paper describes the* P-*level reduction phenomenon and proposes methods to identify and control the occurrence of this privacy breach.*

Keywords: Privacy preserving data mining, data collection, concept hierarchies, inference-control.

ACM Classification:K.4.1 Public Policy Issues, Privacy

1. INTRODUCTION

There is a growing concern in the field of privacy preserving data mining to increase the data provider's control over privacy (Aggarwal *et al*, 2004; Jutla and Bodorik, 2005). Research shows that the higher the perception of control a data provider has over issues of privacy through the use of user-intervention-type tools such as user encryption, cookie crushers, anonymizers and pseudoanymizers, the more trusting the individual is (Jutla and Bodorik, 2005). The success of data mining, which is dependent on the availability of large amounts of truthful data, can be jeopardized if data providers are unwilling to share their data or if they provide incorrect data (Yang *et al*, 2005). Studies in information privacy predict that the future of privacy preserving applications lie with solutions that give the data provider some control over information divulgence (Aggarwal *et al*, 2004; Agrawal and Aggarwal, 2001). Various techniques have been proposed to provide user control during data collection for data mining. Such techniques suggest the use of cryptography for anonymous collection (Yang *et al*, 2005), randomized responses (Clifton *et al*, 2002), agent-guided privacy decision-making (Ackerman and Cranor, 1999), and masks that enable data providers to divulge information anonymously as groups (Isitani *et al*, 2003).

Manuscript received: 25 June 2008
Communicating Editor: Ljiljana Brankovic

Controlling Inference: Avoiding *P*-level Reduction During Analysis

Research into the privacy preferences and behaviours of data providers reveals that users can generally be classified into privacy fundamentalists, the pragmatic majority, and marginally concerned users, depending on the level of concern they have about privacy (Ackerman *et al*, 1999). Furthermore, within each group, depending on the data requested, there are various levels of information sensitivity. For example, though a user may be willing to share their specific age, they may not be comfortable divulging their personal phone number. Thus, the notion of what is sensitive/private differs with culture, context, and individual. Existing approaches to privacy preserving collection tend to offer "one-size-fits-all" or "all-or-nothing" solutions that restrict the control data providers actually have. There is a need for privacy solutions that provide flexible privacy options since there are critical differences among groups of people which vary with situations and information particulars.

One possible data collection approach is to support privacy hierarchies/levels. For example, the age of an individual can be collected in the form of a specific number of years, a range of years, or as a descriptive age group. Each data group represents information at different conceptual levels, implicitly offering different levels of privacy. The success of selective divulgence for data mining is based on the premise that data providers are not equally protective of all attribute values and that sensitive data values can be modified with minimal mining accuracy loss, since mining models are often built on aggregate distributions (Agrawal and Aggarwal, 2001).

In addition to providing flexible user control, collecting data at multiple privacy levels/ hierarchies (*P*-levels) preserves the "truthfulness" of the data. Truthfulness refers to how closely values in the collected data set reflect the actual sensitive data. Truthfulness is preserved in two dimensions:

- Enabling divulgence at multiple *P*-levels avoids situations where data providers give false information (or none at all) in the event that they are uncomfortable with the level of detail requested.
- Enabling divulgence at multiple *P*-levels minimizes the need to perturb data. Data perturbation is required in approaches that use data randomization, swapping, hiding and blocking, to the extent that the data miner cannot guarantee the data value is true or false.

Problem Statement: Concept hierarchies have been applied in data mining for data preprocessing, the mining of multi-level association rules, and knowledge representation (Han and Kamber, 2001). The use of concept hierarchies for privacy preserving data collection is yet to be explored. Two major challenges to collecting data at multiple hierarchies are:

1. How can data that exist at multiple granularities be effectively mined?
2. How can privacy breaches due to inference be controlled?

While shedding some light on the first issue, this paper focuses on addressing the problem of inference control. Inference occurs when data which exists at a lower privacy level than the divulged data is extracted from the original data set or the mining process. This paper addresses two main issues raised by the inference problem:

- How can a data provider identify and control a reduction in the *P*-level of data contributed? In other words, are there indicators (during data collection) that can help detect avenues for inference?
- How can a data collector/miner provide guarantees for data collected in multiple *P*-levels? This involves controls and checks to ensure that *P*-levels will not be reduced in data sets that have been collected using concept hierarchies.

1.1 Contributions of this Paper

This paper describes the use of concept hierarchies to achieve user-controlled privacy preserving data collection. It demonstrates how privacy breaches can occur when data collected with the *P*-level model is gathered and subsequently applied to data mining. Furthermore, it provides formal definitions to identify the *P*-level reduction phenomenon and methods which can be used by the data provider and the data collector to control it. We expect that the application of these techniques will increase the control data providers have over privacy preservation, which can subsequently facilitate privacy preserving data collaboration for personalized web services, medical prediction, strategic business planning, and other applications.

1.2 Paper Outline

The organization of this paper is as follows: Section 2 describes related work in the field of privacy preserving data collection. Section 3 provides a background on concept hierarchies, the main technology that is applied to achieving the *P*-level model. We also discuss the unique approach and challenges involved in applying this data collection approach to data mining. In Section 4, we provide a formal definition of the *P*-level model.

Section 5 describes the *P*-level reduction phenomenon and proposes methods for identifying and controlling *P*-level reduction using a motivating example. Section 6 includes a discussion on data quality and usability while Sections 7 and 8 conclude the paper and provide directions for future work respectively.

2. RELATED WORK

Samarati (2001) and Sweeney (2002) introduce the use of concept hierarchies for privacy preserving data mining by providing a *k*-anonymity guarantee prior to the release of data sets. The goal is to meet the need to share person-specific records without divulging the identities of the persons involved. A table is said to achieve *k*-anonymity if each released record has at least *k-1* other records with identical values in the fields that appear in external data. The fields which contain private information, and thus could be linked to external information are termed the Quasi-Identifier, Q.

Samarati (2001) provides a formal representation of how generalization and suppression can be combined to achieve *k*-anonymity with minimal information loss. Generalization involves replacing an attribute value with a semantically consistent but less specific value, thus achieving more privacy

S. No.	Birth	Post Code	Gender	Income
1	196*	T5G1R3	M	$12,000
2	196*	T5G1R3	M	$28,000
3	1975	P4R5Y8	F	$20,000
4	1975	P4R5Y8	F	$20,000
5	1979	F5Y***	M	$20,000
6	1979	F5Y***	M	$36,000
7	1979	F5Y***	M	$22,000

Table 1: Example of *k*-anonymity, where *k* =2 and *Q*= {Birth, Postcode, Gender}

by using a higher conceptual level. While suppression implies that no value is released. Improvements to the k-anonymity algorithm are addressed by its variants (which provide additional privacy guarantees). These variants such as *l*-diversity, *m*-invariance and *t*-closeness can replace applications of the *k*-anonymity algorithm. Domingo-Ferrer and Torra (2008) provide a summary and critique of these methods.

Machananvajjhala *et al* (2005) propose *l*-diversity, an extension of *k*-anonymity which overcomes the privacy breaches inherent in *k*-anonymity when there exists a unique data value and when there is homogeneity in the values of sensitive attributes. For example, having the background knowledge that a respondent was born in 1975, reveals that she earns $20,000 since there is no diversity in the sensitive values of tuples # 3 and # 4 in Table 1.

The main idea introduced with *l*-diversity is that within a set of tuples whose non-sensitive attributes generalize to q*, referred to as a q* block, there should be at least l, where $l \geq 2$, well-represented, different values of sensitive attributes. Well-represented means that if tuples belonging to *l-2* sensitive values are removed, there should exist a 2-diversity table. The authors prove that an adversary would require at least *l-1* pieces of background information to infer a disclosure. Typically, the data collector applies *k*-anonymity and *l*-diversity using concept hierarchies to already acquired data sets. Therefore the data provider has no input in the process. However, the *P*-level model presented in this paper applies concept hierarchies mainly during the data collection phase (allowing for the data provider's input) and later on during data preprocessing to provide inference control. Yang *et al* (2005) present anonymity-preserving data collection, in which data is collected in such a way that the data miner is unable to match data to individual respondents. The focus here is not making the data anonymous (as in the *k*-anonymity approach), but making the data submission procedure anonymous.

The authors provide three protocols for anonymity–preserving data collection. The first is based on the assumption that the data miner and respondents are semi-honest (follow the protocol strictly, but attempt to violate the privacy of honest respondents). The second protocol assumes toleration of a malicious (may not follow the protocol) miner while the third assumes a malicious miner with some colluding malicious respondents. The paper assumes that a piece of data (d_i) originates from a respondent but does not contain information that can be used to associate it with the respondent. The authors combine the use of ElGamal encryption (a form of public key encryption), a re-randomization technique and a joint decryption technique, to achieve a random permutation of the respondents' data which is sent to the data miner.

Anonymity-preserving data collection claims to be useful for the submission of non-identifying personal data without perturbation. According to Sweeney (Sweeney 2002), most individuals can be uniquely identified with a combination of their postal code, gender, and data of birth (Sweeney, 2002). The *P*-level model, on the other hand, allows the data provider to selectively disclose personal data (identifying or otherwise) anonymously by divulging at higher levels of privacy.

Aggrawal *et al* (2004) propose a new set of privacy standards, the Paranoid Platform for Privacy Preferences (P4P), which advocates that the task of preserving privacy should be controlled and retained by the user. The key idea proposed is to release information in a format that is containable by the owner, traceable to the collecting entity (in the event of a breach in privacy), and unusable for purposes beyond the agreed intention. The authors also propose the use of a software agent/agency that makes automated privacy decisions on behalf of the user, based on pre-selected privacy preferences. The agent manages the privacy of service handles such as email address, credit card numbers, etc. by generating temporary handles for third parties, which work for a stipulated time/use and have a restricted source (can be traced to the organization to which it was released).

The limitation in the P4P approach is associated with any breach in the services of the trusted third party who handles all personal data. Moreover, P4P is a pie-in-the-sky standard for privacy preserving data collection which is yet to be implemented.

Ackerman and Cranor (1999) propose the use of Privacy Critics, semi-autonomous agents, for online privacy protection. Privacy critics offer warnings and suggestions when people make privacy-sensitive decisions, like providing personal data, online. Two important features of Privacy Critics are: they do not take actions on their own but simply provide feedback to users and secondly, a critic-based architecture comprises of multiple critics, each with its own specialty. The authors implemented two prototype critics that provide warnings if a website has been blacklisted on the CyberPrivacy Advocacy Groups' database and if a website requests a combination of information that can be used to uniquely identify an individual. Though, privacy critics do not directly provide privacy protection, they can be used alongside the *P*-level approach for informing decisions on which privacy level to divulge at.

Clifton *et al* (2002), define privacy preserving data mining as getting valid data mining results without learning the underlying data values. They show that secure multiparty computation (SMC), which ensures that parties only learn the mining results, does not guarantee that these mining results do not violate privacy. The authors propose *p*-indistinguishabilty which is preserving the privacy of results in a semi-honest model. Two records ($X_1 \in X; X_2 \in X$) that belong to different individuals are *p*-indistinguishable if for every function $f : X \rightarrow \{0,1\}$, that can be evaluated in polynomial-time,

$$|\Pr\{f(X_{1)} = 1\} - \Pr\{f(X_2) = 1\}| \leq p; \tag{2.1}$$

Where X is the domain of the user provided data.

p-indistinguishabilty provides privacy protection for mining results and thus differs from the *p*-level model which provides privacy protection for data to be collected for the mining system.

In a classification and extended description of privacy preserving algorithms, Verykios *et al* (2004) describe three heuristic-based techniques for association rule hiding. Perturbation-based association rule hiding prevents the mining of a set of sensitive rules by changing a selected set of 1-values to 0-values while maximizing the utility of the released database. A second approach, blocking-based association rule hiding uses data blocking, which is simply replacing the value to be hidden by a question mark. This approach is useful for applications (such as medical applications) where introducing a false value might be dangerous. Blocking based techniques, a third approach, used for classification rules, introduces parsimonious downgrading into classification rule analysis through blocking values for the class label. Decision trees are used for discovering the potential channels for inference in the data set, and then a parametric base set consisting of values $\theta, 0 \leq \theta \leq 1$, is used to replace the values to be hidden. Association rule hiding techniques are useful for preventing inference from data mining results and can be applied as an additional level of privacy protection to data that is collected using concept hierarchies.

3. BACKGROUND

3.1 Introducing Concept Hierarchies

Han and Kamber (2001) define concept hierarchies as: "A sequence of mappings from a set of low-level concepts to higher-level, more general forms". For example, the values of a dimension, *location* can be mapped from *City* to *Province* to *Country* to *ANY*, representing a mapping from a set of low level concepts to a more general, higher level concept. Thus, concept hierarchies organize data in ordered concept levels and help to express data relationships and knowledge in concise high level forms (Han and Fu, 1995). The highest concept is described by the reserved word *ANY*, which

represent all possible attributes in that domain, while the most specific concept corresponds to the specific values of attributes.

Mapping rules (also known as meta-rules) of a concept hierarchy, which indicate desired or meaningful mappings, are often defined by a domain expert. For example, "*2% Foremost milk >> 2% Milk >> Milk*" and "*2% Foremost milk >> Foremost Mi1k >> Foremost*" are two possible hierarchies, but only the former is meaningful or desirable (Han and Fu, 1995). Mappings may organize a set of concepts by a total order such as: *City >>Province >>Country*, or by a partial order, such as: *Day >>Month >>Quarter >>Year* or *Day >>Week >>Year* which would form a lattice.

A concept hierarchy may be defined on a single attribute or across multiple attribute domains. Given a hierarchy H, defined on a set of domains $D_i \ldots D_k$, a concept hierarchy is formally defined as (Han and Fu, 1995):

$$H_p : D_i \times \ldots \times D_k \Rightarrow H_{p-1} \ldots \Rightarrow H_0 \tag{3.1}$$

Where H_p depicts concepts at the primitive level, H_{p-1} represents concepts at the next higher level to that of H_p , and H_0 represents the highest level of concept, ANY.

P-levels, however, are defined in the reverse order of concept hierarchies, i.e., P_0 represents the lowest level of concept and privacy to avoid conflicts in the inherent the use of the words "high" or "low" and the corresponding numbers when describing *P*-levels.

3.2 The Data Collection Interface

The use of concept hierarchies for data presentation, which has been motivated by various reasons including privacy preservation, is not a new concept. However, its use for data collection is novel and fraught with challenges. One obvious challenge is the design of efficient data collection interfaces that are not intrusive to the purpose of data collection (Ackerman and Cranor, 1999). Often, providing users with privacy control implies additional design costs for the data collector and some inconvenience to the data provider. Privacy solutions that will be widely accepted must be easy to use with minimal overhead.

Collecting data using *P*-levels requires a prior knowledge of all possible data values within each domain, to achieve a complete mapping between levels. The data collection interface must guide the provider on possible data values while hiding information that is not required to inform his decision. In an online survey to study data collection using *P*-levels, the use of branching logic and

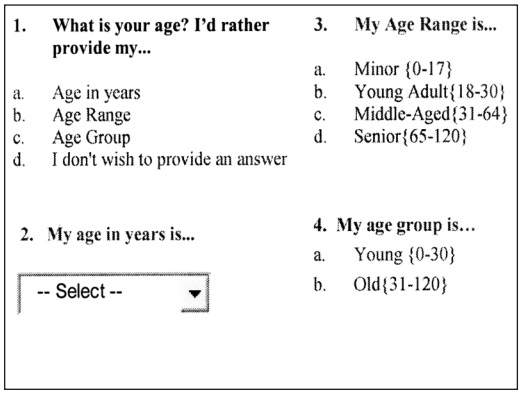

Figure 1:Branching Logic and Skip Patterns-Option 1

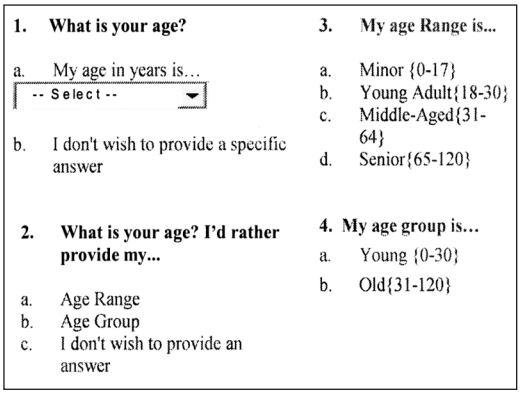

Figure 2:Branching Logic and Skip Patterns-Option 2

skip patterns, has proven quite successful collecting data using *P*-levels. Branching logic and skip patterns, automatically skips irrelevant questions, and vice versa, based on a former response. For example, in Figure 1, a respondent skips Question 2, if his answer to Question 1 is b.

Another useful interface could be designed to request data at the lowest (most specific) level of privacy, with the option of not providing a specific answer. Only if the latter option is selected would the same question could be repeated with possible answers at higher *P*-levels. This approach will simplify the data collection interface for data providers who wish to divulge at the lowest level of privacy, which (for many attributes) is the largest sub-group of data providers. This interface is illustrated in Figure 2. Additional interface scenarios that both protect privacy and encourage honest complete responses could be developed but are left as an area of future work.

3.3 The Effectiveness of *P*-level Collection

In a survey, distributed online between November 2006 and August 2007, with over 800 completes, we study the benefits and shortcomings of the *P*-level data collection approach. We compare *P*-level data collection approach to the regular data collection approach, referred to in this research as the fixed-level approach, in terms of user acceptance and perception and its capacity for data collection. The survey consisted of 14 demographic and employment questions which were anonymously answered. The exact questions used in this survey are available in Appendix A.

Some results, in Table 2, show that the *P*-level approach provided the opportunity to collect data that would have otherwise been withheld or possibly falsely given. For example, only 63% of respondents would have provided their actual age in years. However, an additional 37% (23+11+3) of respondents provided age information at higher privacy levels. Table 2 also shows that marital status and income are the least and most sensitive attributes because 25% and 75% of the respondents provided their specific income at *P*-level *0* for each attribute respectively. Figure 3 shows that about 70% of users perceive that the *P*-level collection approach provides more privacy control and minimizes risk. However, close to 40% of users find it less convenient, and users don't necessarily find it a motivation for truthfulness.

3.4 Data Pre-processing

Collecting data at multiple privacy levels implies that collected data will exist at multiple levels of granularity. A major challenge to mining data collected this way is pre-processing it with minimal loss of data and prediction accuracy. Data preprocessing involves data cleaning (sanitizing missing, inconsistent and noisy data); data integration (merging data from various sources); data

Attribute	*P*-0 (%)	*P*-1 (%)	*P*-2 (%)	*P*-3 (%)
Marital Status	75	20	5	–
Education	68	16	11	5
Age	63	23	11	3
Occupation	53	33	13	–
Race	47	39	15	–
Income	25	53	23	–

Table 2: Capacity for Data Collection

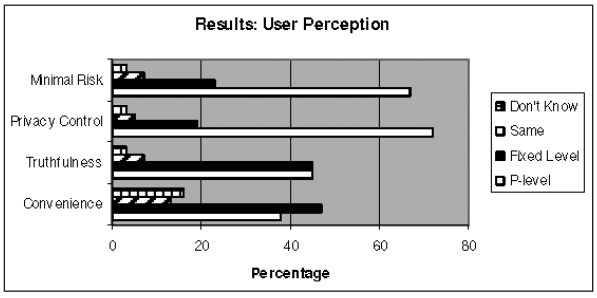

Figure 3: User Perception of the *P*-level Approach

	A	B	C	D	E
1	1	2	1	3	2
2	1	1	2	1	1
3	3	2	1	1	3
4	1	1	3	2	3
5	1	2	1	3	2
6	2	3	1	1	2

	A	B	C	D	E
1	1	?	1	?	?
2	1	1	?	1	1
3	?	?	1	1	?
4	1	1	?	?	?
5	1	?	1	?	?
6	?	?	1	1	?

Table 3: Data at multiple *P*-levels as Missing Data

transformation (aggregating into forms suitable for mining) and data reduction (trimming huge data sets). Data collected at multiple granularities require extensive data cleaning and transformation (normalization and aggregation) techniques before the data set can be meaningfully analyzed. Our approach is to mine data at a single pre-selected privacy level. Data that exists at higher privacy levels are treated as missing data. For example, if the six records in Table 3 will be mined at privacy level 1, data that is provided at levels 2 and 3 can be meaningfully replaced using standard missing data techniques such as multiple imputation (Horton and Kleinman, 2007).

The benefit of this data collection approach is that each missing value can be replaced with more accurate values which are generated from the distribution of the range that has been specified by the data provider, rather than the entire data distribution. For example, an age range response: *0-17* years, can be more accurately substituted with a value generated from the age *0-17* distribution rather than the entire age distribution.

Data that is mined that higher privacy levels than provided can be easily replaced with data on the higher levels by using the concept hierarchies' rules (mappings).

4. THE *P*-LEVEL MODEL

In this section, we discuss the requirements for a data model and formally present the *P*-model.

4.1 Model Requirements

A data model is made up of three components: structure (entities, events and data), operations (rules), and constraints (limitations) (Tsichritzis and Lochovshy, 1982). Our objective for the *P*-level model is to demonstrate that operations are feasible and consistent, and constraints are precise enough to detect violations of the allowable operations on structures within the system. Table 4 shows the *P*-level model structure and provides some definitions.

4.1.1 Constraints

Provision of Limits: Personal data should not be collected if the data provider is not equipped with a means of specifying a desired privacy limit, such as, the ability to divulge on a user-selected privacy level.

$$S_k \subseteq O_i(D_i) \times C_k(U_r) \tag{4.1}$$

Limit by Purpose: Personal data should not be collected at privacy levels below those required to complete the purpose declared for the data collection.

Element	Description	Definition
Data Provider	Entity who owns/ provides personal data	O
Data Collector	Entity who requests for personal data	C
Attribute	Properties of an entity	$A = (A_1, A_2, ...A_j)$
Record	Related attribute values obtained from a data provider	$t_i = (A_{i1}, A_{i2}, ...A_{ij})$
P-level	The degree of generalization of a piece of sensitive information.	px; where $x = \{0...h\}$ h =max. generalization
Personal Data	Data about an identifiable person.	$D \subseteq t_i = (A_{i1}^{px}, A_{i2}^{px}, ...A_{ij}^{px})$
Collection Purpose	The reason for data collection	$U = (u_1, u_2, ..u_r)$
Data Transformation	An alteration to personal data that (ultimately) changes its privacy level	$\exists F_k, F_k := A_{ij}^{px} \rightarrow A_{ij}^{py}$
Collection System	Requirements for collecting personal data	$S \subseteq O(D) \times C(U)$
Minimum Available Privacy level	The lowest privacy level available as a choice for data providers.	$MAP(A_j)$
Minimum Required Privacy Level	The lowest privacy level required for the purpose of collection.	$MRP(A_j)$
Divulged Privacy Level	The privacy level at which a data provider divulges personal data.	$DP(A_{ij}) \in px$
Transformed Privacy Level	The privacy level at which collected datum exists after it has been processed.	$TP(A_{ij})$

Table 4: The *P*-level Model Structure

$$DP(A_{ij}) \geq MRP(A_j) \tag{4.2}$$

Adherence to Limits: While collected personal data can be replaced with data representations with privacy levels below those specified by the data provider, using missing data methods, they must not be replaced with specific, exact data (even if they can be obtained from other sources) at lower privacy levels.

$$\text{If } U_1(A_{ij}^{px}) \rightarrow U_1(A_{ij}^{py}) : (x = DP(A_{ij})) > (y = TP(A_{ij})) \quad U_1(A_{ij}^{py}) \neq U_2(A_{ij}^{py}) \tag{4.3}$$

Operation Limits: A data transformation operation on a collected data set is valid only if it is necessary to achieve the collection purpose.

$$\forall F_k : F_k \in U_r(F) \tag{4.4}$$

Where $U_r(F)$ defines as the set of transformation operations, F required to achieve purpose U_r.

4.2 Defining the *P*-level Model

A *P*-level, is the degree of generalization, x, of a piece of sensitive information A_{ij}, such that, $x = \{0, 1, 2, \ldots h\}$ and h is the highest level of generalization possible on attribute A_j. When data has not been generalized at all, it is said to be at a *P*-level of *0*, representing the lowest level of privacy. Sensitive information which has been generalized to a level x, can be denoted as A_{ij}^{px}. A high *P*-level signifies highly generalized data and therefore, more privacy by virtue of less specificity. For instance, an attribute *Age* with values *{16, 35, 70}* at the primitive level can be

mapped by the rules: *{0-17}* → minor; *{18-30}* → young-adult; *{31-64}* → middle-aged; *{65-120}* → Senior; into a higher concept (privacy) level depicted in the set {minor, young-adult, middle-aged, senior}.

Implying that, an attribute value of {Age^{P1} = middle-aged} is at a higher *P*-level than {Age ^{P0} = *35*}.

Formally:

Let $T = \{t_1, t_2, \ldots, t_n\}$ denote a table with n records corresponding to at most n different data providers.

Let $A = \{A_1, A_2, \ldots, A_m\}$ represent the set of all attributes in T.

All possible data values for an attribute $A_j \in A$, can be expressed as a hierarchy of concepts, ordered from specific concepts (low privacy) to general concepts (high privacy) such that:

$$A_j^{ph} \Rightarrow A_j^{p(h-1)} \Rightarrow \ldots \Rightarrow A_j^{p0} \tag{4.5}$$

Let A_{ij}^{px} represent the value of an attribute A_j for a record t_i belonging to the i^{th} data provider at *P*-level x.

Let the *P*-level of value A_{ij}^{px} be fully represented as x_{ij}

The privacy concern level of the i^{th} data provider can be represented as the sum of *P*-levels selected over all attributes.

$$t^i = \sum_{k=1}^{k=m} x_{ik} \tag{4.6}$$

Likewise, the divulgence level of any attribute, A_j,(a reflection of its sensitivity), in a data set can be expressed as the sum of all *P*-levels chosen by all data providers on that attribute.

$$A^j = \sum_{k=1}^{k=n} x_{kj} \tag{4.7}$$

Table 5 and Figure 4 illustrate a data set collected using the multiple *P*-level data collection approach. In addition to giving each of the data providers control over the divulgence of sensitive information, the data set further provides information to the data collector about the most sensitive attribute (A_5) and the most "private" data provider (t_5).

Obtaining information on the sensitivity of data attributes and privacy preferences of data providers will provide an understanding of the nature of data collected and trust levels of data providers. For instance, attributes relating to weight maybe sensitive for obese respondents while

Attributes /Records	A_1	A_2	A_3	A_4	A_5	t_i
t_1	1	3	2	1	2	**9**
t_2	3	1	2	1	3	**10**
t_3	2	1	2	1	1	**7**
t_4	2	1	0	1	4	**8**
t_5	1	4	1	3	2	**11**
A_j	**9**	**10**	**7**	**7**	**12**	

Table 5: Data collection at multiple *P*-levels

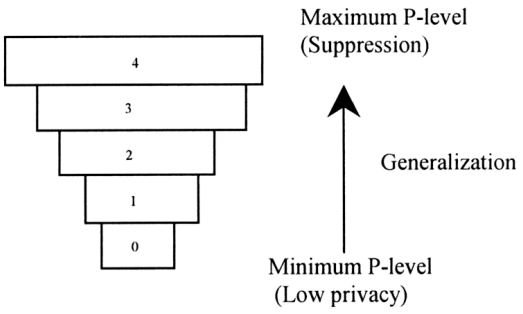

Figure 4: Achieving Privacy using *P*-levels

income attributes maybe sensitive for very low and very high earners. Thus, the *P*-level approach provides different levels of privacy for various attributes depending on the sensitivity of that attribute to each individual data provider.

5. THE *P*-LEVEL REDUCTION PHENOMENON

In this section, we describe scenarios for privacy invasion using a motivating example.

The *P*-level data collection approach provides the data provider with a choice to operate at the privacy level at which they feel comfortable. The data collector must be content with obtaining the highest level of accuracy possible with the data collected and made available under these constraints. Guaranteeing the data provider's privacy requires understanding the privacy breaches inherent in the data mining process which may result in the disclosure of sensitive information at *P*-levels below the disclosed *P*-level.

5.1 Motivating Example

In this section, we present possible privacy breaches using age, education, and address as example attributes that can each be mapped into at least three *P*-levels (P_0, P_1, P_2) using appropriate rules as shown below. Data providers can provide data at any of the specified *P*-levels.

Age(A_1):= *P_0[{0-17}, {18-30}, {31-64}, {65-120}]* $\Rightarrow$ P_1 [{minor}, {young-adult} ;{middle-aged}, {senior}] $\Rightarrow$ P_2 [{young}, {old}] $\Rightarrow$ P_3 [{Any}]

A_1 rules [$P_0 \Rightarrow P_1$]: {0-17} $\rightarrow$ minor; {18-30} $\rightarrow$ young-adult; {31-64} $\rightarrow$ middle-aged ; {65-120} $\rightarrow$ senior

A_1 rules [$P_1 \Rightarrow P_2$]:{minor, young-adult} $\rightarrow$ young; {middle-aged, senior} $\rightarrow$ old

A_1 rules [$P_2 \Rightarrow P_3$]: {young, old} $\rightarrow$ Any

Education(A_2):= P_0 [{G1-G6}, {G7-G9}, {G10-G12}, {College-B.Sc.}, {Professional, Masters Doctorate}];$\Rightarrow$ P_1 [{Elementary}, {High School};{Higher Education}, {Graduate School}] $\Rightarrow$ P_2[{No Degree}, {Degree}] $\Rightarrow$ P_3 [{Any}]

A_2 rules [$P_0 \Rightarrow P_1$]: {G1-G6} $\rightarrow$ Elementary;{G7-G9,G10-G12} $\rightarrow$ High School ; {College - B.Sc} $\rightarrow$ Higher Education; {Professional, Masters, Doctorate} $\rightarrow$ Graduate School

A_2 rules [$P_1 \Rightarrow P_2$]: {Elementary, High School} $\rightarrow$ No Degree; {Higher Education, Graduate School} $\rightarrow$ Degree

A_2 rules [$P_2 \Rightarrow P_3$]: {No Degree, Degree} $\rightarrow$ Any

Address (A_3):= P_0 [{Street, City, Province, country}] $\Rightarrow$ P_1[{City, Province, Country}] $\Rightarrow$ P_2[{ Province, Country }] $\Rightarrow$ P_3 [{Country}] $\Rightarrow$ P_4 [{Any}]

A_3 rules [$P_0 \Rightarrow P_1$]:{Street, City, Province, Country} $\rightarrow$ City, Province, Country

A_3 rules [$P_1 \Rightarrow P_2$] :{ City, Province, Country} $\rightarrow$ Province, Country

A_3 rules [$P_2 \Rightarrow P_3$]: {Province, Country} $\rightarrow$ Country

A_3 rules [$P_3 \Rightarrow P_4$]: {Country} $\rightarrow$ Any

Given two data providers, Alice and Bob, and two colluding data collectors/miners, C_1 and C_2, Alice could divulge information, t_1 to C_1 and choose to divulge information on identical attributes to C_2 at the privacy levels indicated in t_2 as shown in Table 6. Bob, on the other hand, could divulge information, t_3 to C_1 and choose to divulge information identical attributes to C_2 at the privacy levels indicated in t_4.

Attributes/Records	Age	Education	Address
Alice: t_1 at C_1	P_1 {Middle-aged}	P_3{Any}	P_2{ Alberta, Canada}
Alice: t_2 at C_2	P_3 {Any}	P_1{Graduate School}	P_0{Tuscany, Calgary, Alberta, Canada}
Bob: t_3 at C_1	P_2 {young-adult}	P_2{No Degree}	P_1{ Calgary, Alberta, Canada}
Bob: t_4 at C_2	P_0 {28}	P_1{High School}	P_3{ Canada}

Table 6: Data Collected at Multiple *P*-levels

5.2 Defining *P*-level Reduction

A *P*-level reduction occurs when any of the data collectors (e.g., C_1,C_2) or an external adversary, C3, with legitimate access, can infer personal information that exists at a privacy level below the divulged privacy level for an attribute *j*, belonging to a data provider, *i*, in a collected data set. Formally,

$$TP(A_{ij}) < DP(A_{ij}) \tag{5.1}$$

All inferences can be generally categorised as occurring through data from one or more of the following hierarchy of sources:

- Data that is external to the data set(s)
- Data within the data set(s) but external to the tuple containing inferred data.
- Other attributes values within the tuple containing inferred data

We demonstrate the three major groups of attacks (derived from these inference sources) on data collected at multiple *P*-levels using the data in Table 6. While this list of attacks is not necessarily an exhaustive one, it includes the key avenues for inference based on these three levels of data.

5.2.1 External Linking Attack

This category of *P*-level reduction describes instances where inferences occur through data sources that are external to the collected data set (s).

Unique Data Value: Assuming the adversary, C_3, has some background knowledge of Alice's address and education. For instance, C_3 knows that Alice lives in {Tuscany, Calgary, Alberta, Canada} and has graduate education. Supposing C_3 has access to data collected by C_2 and is interested in determining the value of an additional sensitive attribute, say Disease. C_3 might be able to single out Alice's record by noting that a single data provider in C_2's data set has specified "Tuscany, Calgary, Alberta, Canada" and "Graduate School". Note that the possibility of inference is increased by divulging a value at *P*-level 0 which happens to be unique to the data set. However, C_3 cannot infer anything at a privacy level below which Alice divulges.

Lack of Diversity: When the sensitive values of released data sets are not adequately diversified, inference could occur. Consider the following scenario:

"The adversary, C_3, knows that Bob has a high school diploma and lives in Canada. C_3 would like to determine Bob's exact age and has access to data collected by C_2.

Let's assume that in C_2's data set, there are three people (including Bob) with the {High School, Canada} value combination. Supposing that the other two people both specify an age of {28} for the age category, C_3 can confidently infer that Bob is 28 years old.

5.2.2 Collector-Linking Attack

This category of *P*-level reduction describes instances where the source of inference is within the collected data set(s). For example, multiple data collectors (or an external adversary) can combine collected data sets to infer information at a lower privacy level than was divulged.

Colluding Collectors: Two malicious data collectors, C_1 and C_2, can collude on the information, t_1 and t_2 respectively, which they have collected on Alice. Assuming that C_1 is interested in obtaining Alice's specific address and is willing to trade it for her age group (in which C_2 is interested). Both data collectors will be unable to identify Alice's record unless each of them have an attribute (e.g., her name) in common with the same exact value by which they can link up both records.

Cooperative Analysis: Suppose two data collectors, C_1 and C_2 would like to combine their data sets to build a stronger classification tool. If neither collector is assured that the other is honest, how can they cooperate without violating Alice's privacy? Both parties can combine their data sets in an open model, such that they both have access to each other's data. This option is not desirable if C_1 and C_2 are competitors and they are constrained by privacy laws not to release their data sets to third parties. Alternatively, they can use the secure multi-party computation model, such that no party learns anything more than its input and the final classification model. Thus in the latter case, the classifier is a black box which all data collectors have access to but do not know how it works.

5.2.3 Attribute-Linking Attack

In this category, we demonstrate scenarios where the data within a compromised tuple might constitute a privacy breach. For example, a data provider might unknowingly give out individual pieces of information which when combined together, would constitute a breach of his own privacy.

Data Linking: When certain attributes are divulged, they could lead to the inference of data at a privacy level below which the data provider willingly revealed. For example, Alice contributed the data t_2 at C_2 with values {any, Graduate school, (Tuscany, Calgary, Alberta, Canada)}, corresponding to {Age, Education, Address}. The value {Graduate school} implies that Alice is not a minor, so an adversary can eliminate that possibility, and with further background knowledge isolate Alice's age group, which is at a higher level than was intended to be divulged.

Analysis Linking: Suppose C_1 builds an association rule classifier that predicts that {Young, Calgary $\Rightarrow$ High School}, which is then applied to collected data. When Bob contributes the data t_3 at C_1 with values {young, No Degree, (Calgary, Alberta, Canada)}, corresponding to {Age, Education, Address}, C_1 can infer that Bob has at most a High school degree, revealing information at a privacy level below the level at which Bob is comfortable.

5.3 Detecting and Controlling *P*-level Reduction

This section presents formal definitions and methods to identify and control the *P*-level reduction privacy breaches demonstrated in the previous section. Each *P*-level reduction attack is discussed in terms of how it can be detected, avoided and controlled, either by the data provider or data collector. Avoidance techniques are preventive in operation and can be applied by the data provider at of point data collection. Control techniques, are a combination of existing data preprocessing techniques such as *k*-anonymity and *l*-diversity (described in Section 2) which the data collector/owner can apply to minimize inference after data has been collected.

5.3.1 Unique Data Value

Detecting *P*-level Reduction: Inference of information through a unique data value is possible

when attribute values are unique at the lowest *P*-levels in any data set. For example, if Alice is the only data provider in C_2's data set that has specified {(Tuscany, Calgary, Alberta, Canada)} as noted above.

Formally, unique data value disclosure occurs if:

$$n([A_j^{pk}] = R) \leq 1 \tag{5.2}$$

Where, $n\ ([A_j^{pk}] = R)$ denotes the number of tuples, that have the attribute A_j set at privacy level $x = k$, for which R is an attribute value. To avoid inference due to a unique data value, the value of $n\ ([A_j^{pk}] = R)$ must be greater than *1*. Inference is greatest when $k = 0$.

The data provider is limited during privacy decision making by not having access to the data provider's entire data set and mining model (incomplete information). In a detailed study, Acquisti (2004) shows that individuals are further limited by bounded rationality (the inability to calculate all parameters relevant to the current choice) and psychological distortions (may still choose to deviate from rationality in favour of immediate gratification) (Acquisti, 2004).

Avoiding *P*-level Reduction: The data provider could avoid disclosure on at the lowest *P*-level on at least one attribute within the set of potentially identifying attributes (attributes that can be used in combination for identification). In a series of experiments based on the 1990 US census data, Sweeney shows that a person can be uniquely identified using a combination of three attributes: postal code, gender and data of birth, referred to as identifying attributes (Sweeney, 2000).

Ensuring that at least one of these attributes is not divulged at *P*-level 0 will significantly decrease the chances of re-identification and inference. We propose that an individual's decisions to avoid *P*-level reduction can be guided by a software agent, such as a privacy critic. Privacy critics (Ackerman and Cranor, 1999) have been implemented to detect and flag potential identifying attributes (Section 2). A software agent can suggest rational options and provide *P*-level reduction warnings.

Formally, *If*

$$a = \{A_1, A_2, ...A_k\}\ ,\ a \subset A = \{A_1, A_2, ...A_m\}, a \subset PIA,$$
$$\exists A_j \in a \mid A_j^{pk} \mid k > 0 \tag{5.3}$$

Where *PIA* is the global set of potentially identifying attributes and *a* is a subset of *PIAs* that exists in *A*.

Controlling *P*-level Reduction: Prior to public release, the data collector can achieve *k*-anonymity (Sweeney, 2002) or one of its variants (see Section 2) on the set of potentially identifying attributes at *P*-level 0; where *k* is the guaranteed level of anonymity.

5.3.2 Lack of Diversity

Detecting *P*-level Reduction: A *P*-level reduction through lack of diversity can be detected if all records at a certain privacy level have the same value for a sensitive attribute such as the set of potentially identifying attributes.. Formally, a lack of diversity inference occurs if:

$$\exists t : t = \{t_1, t_2, ...t_m\} \subset T = t = \{t_1, t_2, ...t_n\} :$$
$$\forall t_i \subset t : t_i = A_j^{px} = R \tag{5.4}$$

Where *t* denotes the subset of tuples in data set *T*, which have attribute A_j disclosed as the value *R*, on a certain privacy level *x*.

Avoiding *P*-level Reduction: Since the data provider is not aware of the values provided by other respondents, there is not a lot they can do to avoid lack of diversity. However, if the data provider divulges a sensitive value, x, at P-levels greater than 0, it increases the probability that the homogenous block which x belongs to is large." Large blocks of sensitive homogenous values invariably contain enough diversity in the non-sensitive values to introduce uncertainty into any inference efforts.

Controlling *P*-level Reduction: Data collectors should achieve *l*-diversity (Machananvajjhala *et al,* 2005) or one of its variants (see Section 2). Note that l is the number of pieces of background information needed before the adversary would achieve a P-level reduction.

5.3.3 Colluding Collectors

Detecting *P*-level Reduction: An inference can occur when two malicious data collectors, C_1 with data set T_1 and C_2 with data set T_2 can link tuples t_1 at C_1 and t_2 at C_2 using one or more equivalent attributes with identical unique values at privacy level 0 with certainty, as belonging to a data collector Alice. For example, if attributes {sex, date of birth, postal code} have values *{F, 15th August 1973, 12345}* in one data set, colluding collectors can might easily link them to a tuple in a second data set with these exact values, rather than say, *{F, 1973, 123**}*, especially if these values are unique in both data sets. Formally, this can be represented as:

$$\exists j \in \mathbb{N} P([A_j^{p0}] @ C_1) = ([A_j^{p0}] @ C_2) \tag{5.5}$$

Where $\mathbb{N}$ is the set of natural numbers, $[A_j^{p0}]@C_1$ denotes an attribute value at site C_1 , with attribute A_j set at privacy level $x = 0$.

Avoiding *P*-level Reduction: The probability of contributing a unique sensitive value is higher at lower P-levels. Data providers can protect themselves against inference through the linking of a unique sensitive value by disclosing on at least P-level 1. Even if sensitive values are unique at P-level 1, they are no longer identifying.

Controlling *P*-level Reduction: A data collector can set sensitive data at P-level 0 to P-level 1, to prevent data matching in the event of a disclosure of the data set.

5.3.4 Cooperative Analysis

Detecting *P*-level Reduction: According to Clifton *et al* (2002), in the best case of cooperative analysis, if both parties use the secure multi-party computation model, nothing is revealed. In the worst case, when data collectors perform cooperative analysis by directly sharing data, a reduction in P-level is said to occur if a data collector C_1 with attribute value A_{ij}^{px} in T_1 can improve prediction ability (or actually discover) a sensitive value $A_{ij}^{p(x-1)}$ after access to data T_2, collected by a data Collector C_2. Formally, a P-level reduction occurs if:

$$\begin{aligned} &\Pr\{(A_{ij}^{px} \Rightarrow A_{ij}^{p(x-1)}) : T = T_n\} \\ &< \Pr\{(A_{ij}^{px} \Rightarrow A_{ij}^{p(x-1)}) : T = T_n + T_m\} \end{aligned} \tag{5.6}$$

Avoiding *P*-level Reduction: A data provider (or his privacy agent) should divulge only if collecting sites guarantee no direct data sharing, and even then, disclose on at least P-level 1.

Controlling *P*-level Reduction: Honest data collectors who intend to protect the data with which they have been entrusted while performing cooperative analysis with other data collectors, will provide guarantees of non-disclosure for released data sets by using the secure multi-party

computation model (Clifton *et al*, 2002) or a related model. The secure multi-party computation model is discussed in Section 2.

5.3.5 Data Linking

Detecting *P*-level Reduction: Inference through data linking occurs when two or more attributes are naturally associated. A disclosure at a privacy level, x, for one attribute could result in the inference of the associated attribute on at least that privacy level. Formally, *P*-level reduction through data linking occurs if:

$$\begin{aligned} &\exists a : a = \{A_1, A_2, ...A_k\} : A_1 \sim A_2 \sim ... \sim A_k, \\ &\forall A_j \in a, \exists t_i \subset T \;\; if \; x_{ij} = R, \\ &then \{x_{i1}, x_{i2}, ..., x_{ik}\} \text{ can be reduced to } R \end{aligned} \tag{5.7}$$

where $x_{ij} = R$ is the privacy level of an attribute A_j (for a record t_i) which belongs to a set a = { A_1,A_2,…..,A_k } of associated attributes.

Avoiding *P*-level Reduction: Identify the set of associated attributes and disclose both sensitive and non-sensitive attributes at the same *P*-level.

$$\begin{aligned} &\exists a = \{A_1, A_2, ..., A_k\} \subset A = \{A_1, A_2, ..., A_m\} : \\ &A_1 \sim A_2 \sim ... \sim A_k, \forall A_j \in a : \text{ if } \; x_j = R \text{ then}, \\ &\qquad \Rightarrow x_1 = x_2 = ...x_k = R \end{aligned} \tag{5.8}$$

where a, a subset of all attributes A, is the set of associated attributes.

Controlling *P*-level Reduction: For the set of associated attributes, on any record, data collectors should set all attributes values to the highest *P*-level indicated.

5.3.6 Analysis Linking

Detecting *P*-level Reduction: An honest data collector has a significantly higher level of control on *P*-level reduction and this control can be exercised by applying the following measures prior to data mining or public release of collected data:

Here, the likelihood of *P*-level reduction is associated with the accuracy of the data mining model which C_1 is using for prediction. The data mining model refers to the prediction engine (e.g., a set of rules) which is used to predict the outcome (e.g., class) of new instances. Since there are no data mining models with perfect predicative accuracy, a data collector is unable to conclude with certainty the value of an attribute at a reduced *P*-level. The probability of disclosure can be formally defined as:

For Mining Model, M: $AC(M) = Q$,

If M predicts that $A_j^{px} \Rightarrow A_k^{p(x-r)}$

$$\Pr(x_k - r) \leq Q \tag{5.9}$$

Where $AC(M) = Q$ is the accuracy (the correctness of prediction) of the mining model M and r is the reduction in privacy level x_k of attribute A_k.

Avoiding *P*-level Reduction:
In multi-level data mining, prediction often occurs at adjacent concept levels (Han and Fu, 1995). For example, if age is collected at *P*-level 2, alongside other attributes, prediction will occur on at least level 1 and on at most *P*-level 3 (the adjacent levels to levels to *P*-level 2) for the age attribute.

Controlling *P*-level Reduction:
Apply rule hiding techniques (see Section 2) to rules that reveal sensitive data (Verykios *et al*, 2004).

6. QUALITY AND USABILITY OF COLLECTED DATA

In this section, we provide a discussion on the quality and usability of data collected using the *P*-level approach and the effect of the proposed techniques on the data.

6.1 Quality of Collected Data

Bartini *et al* (2005) present the various dimensions of data quality such as accuracy, currency, consistency and completeness. Accuracy is a measure of measure of how close (semantically and syntactically) a value v' in a data set is to the real world value, v). Currency (and other time-related data quality measures such as timeliness and volatility) focuses on how data changes with time. Consistency measures discrepancies in data that violate sematic rules defined over a data set, while completeness of a dataset measures the "the extent to which data are of sufficient breadth, depth and scope for the task at hand" (Bartini *et al*, 2005).

In this work, we focus on completeness because the major issue introduced by data collection using the *P*–level approach is that of incomplete data. Completeness in a data set can be measured with respect to the presence of and meanings of null values. A null value can mean one of the following: the value is not existing, the value existing but not known or it is not known if the value is existing. Attribute completeness measures the ratio of specified values for an attribute with respect to the total number of values that should have been specified (Pilar and Lachlan, 2005). This is formally represented as:

$$\frac{\text{Number of Specified Values}}{\text{Total Number of Values}} \tag{6.1}$$

We compute the attribute completeness at *P*-level 1 for the data collected in our survey. We select *P*-level 1 because most data mining applications analyze data at this (or higher) level of specificity, below which the data groupings are too many to extract useful patterns. We also note that the privacy-concerned data provider will simply not provide any data if they can only provide the most detailed data. Thus, the value of the higher granularity data is undoubtedly less than fine detailed data but it is far better than no information at all." We present our results in Table 7:

Attribute	***P-level 1 Value***	***Completeness***
Age	*Age Range*	*86%*
Marital Status	*Married/Not*	*95%*
Occupation	*Occupational Group*	*86%*
Education	*Class of Education*	*86%*
Country	*Continent*	*97%*
Income	*Income group*	*78%*

Table 7: Quality of Collected Data

6.2 Usability and Usefulness of Techniques

When collected data is analyzed at particular *P*-level, data that exits at higher levels can as well be regarded as "null values". These null values (except ANY) inherently contain a lot of information to the extent that a semantically accurate approximation them can be generated using data imputation methods. For example, For example, if the task for which data has been collected is data mining, and an attribute, say, Age is to be mined at *P*-level 1 (Age Range). Data at the *P*-level 0 can be accurately mapped while values at higher levels can be replaced with approximate values using missing data methods. Imputed values can significantly increase the quality of data. For example, imputing data from *P*-levels 2 for the Age attribute in Table 7 increases completeness to 96%.

7. CONCLUSION

Data providers are getting increasingly concerned about preserving informational privacy (Ackerman *et al*, 1999). By implementing privacy during data mining runtime, existing solutions place the task of preserving privacy solely in the hands of the data miner who may deliberately or accidentally betray the trust of the data provider (Aggarwal *et al*, 2004). Recent research suggests that user-controlled privacy preservation approaches hold the solution to this trust problem (Aggarwal *et al*, 2004). This paper presents the use of concept hierarchies, represented as *P*-levels, at the data collection stage of data mining to enable user control at the datum level. We formally define the *P*-level model and describe the *P*-level reduction privacy breach, which can occur when data collected at multiple *P*-levels is aggregated, linked to external data or used for data mining purposes. We use an example to demonstrate occurrences of *P*-level reduction and formally identity these situations. Finally, we propose methods for controlling *P*-level reduction which can be used by both the data provider and the data collector.

8. FUTURE WORK

This research is preliminary, thus there are many avenues for future work. The wide acceptance of the *P*-level collection approach will be dependent on its ability to provide data accuracy at least matching those attainable using the fixed-level approach. It should be noted that while this collection approach will have the same number of missing specific values as the fixed level approach, there will be more information about these missing values since they could be available at higher granularities. However, what needs to be ascertained is that since the replacement missing data is informed, the level of accuracy obtained from mining this data will be higher or at least comparable to data collected using the fixed level approach.

Interfaces that allow data collection at multiple privacy levels will be required for various types of data. Issues such as improved usability and effectiveness are challenges that must be met before this collection approach can be widely applied.

ACKNOWLEDGEMENT

We thank the paper reviewers for their detail and insightful comments which has in no small measure improved the clarity and quality of this paper.

REFERENCES

ACKERMAN, M. and CRANOR, L. (1999): Privacy critics: UI components to safeguard users' privacy: In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'99)*, Short Papers 2: 258-259.

ACKERMAN, M., CRANOR, L. and REAGLE, J. (1999): Privacy in E-commerce: Examining user scenarios and privacy preferences: In *Proceedings of the ACM Conference on Electronic Commerce (EC'99)*, Denver, Colorado, USA, 1-8.

ACQUISTI, A. (2004): Privacy in electronic commerce and the economics of immediate gratification: In *Proceedings of the ACM Electronic Commerce Conference (EC 04)*. New York, 21-29.

AGGARWAL, G., BAWA, M., GANESAN, P., GARCIA-MOLINA, H., KENTHAPADI, K., MISHRA, N., MOTWANI, R., SRIVASTAVA, U., THOMAS, D., WIDOM, J. and XU, Y. (2004): Vision paper: Enabling privacy for the paranoids: In *Proceedings of the 30th VLDB Conference*, Toronto, Canada, 708-719.

AGRAWAL, D. and AGGARWAL, C. (2001): On the design and quantification of privacy preserving data mining algorithms: In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Santa Barbara, California, United States, 247-255.

BATINI, C., SCANNAPIECO, M. and MISSIER, P. (2005): Data quality at a glance. Datenbankspektrum, Vol. 14, Available at: http://www.datenbank-spektrum.de/pdf/dbs-14-6.pdf

CLIFTON, C., KANTARCIOGLU, M., and VAIDYA, J. (2002): Defining privacy for data mining, In *National Science Foundation Workshop on Next Generation Data Mining*: KARGUPTA, H., JOSHI, A. and SIVAKUMAR, K. Eds., Baltimore, MD, November,126-133.

DOMINGO-FERRER, J. and TORRA, V. (2008) A critique of k-Anonymity and some of its enhancements: In *Proceedings of the Third International Conference on Availability, Reliability and Security (ARES 08)*, March, 990-993.

FULE, P. and RODDICK, J. (2004): Detecting privacy and ethical sensitivity in data mining results: In *Proceedings of the 27th conference on Australasian Computer Science*, 26:159-166.

HAN, J. and FU, Y. (1995): Discovery of multiple-level association rules from large databases: In *Proceedings of the 21st International Conference on Very Large Data Bases*, September, 420-431.

HAN, J. and KAMBER, M. (2001): *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Fransico, USA, 56-58.

ISHITANI, L., ALMEIDA,V. and MEIRA, W. (2003): Masks: Bringing anonymity and personalization together. *IEEE Security & Privacy*, 1(3): 18-23.

JUTLA, D. and BODORIK, P. (2005): Sociotechnical architecture for online privacy: *IEEE Security and Privacy*, 3(2): 29- 39.

HORTON, N. and KLEINMAN, K. (2007): Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models, *The American Statistician*, 61(1): 79-90.

MACHANANVAJJHALA, A., GERKE, J. and KIFER, D. (2005): *l*-Diversity: Privacy beyond *k*-Anonymity: Technical Report, Cornell University.

PILAR, A. and LACHLAN, M. (2005): Quality measurement and assessment models including data provenance to grade data sources. In PETRATOS, P. and MICHALOPOULOS, D. (Eds.), *Computer Science and Information System,* 101-116. Athens:ATINER

SAMARATI, P. (2002): Protecting respondents identities in microdata release, *Knowledge and Data Engineering, IEEE Transactions*, 13(6):1010 – 1027.

SWEENEY, L. (2002): Achieving *K*-Anonymity privacy protection using generalization and suppression, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*: 10(5): 571–588.

SWEENEY, L. (2000): Uniqueness of simple demographics in the US population, LIDAP-WP4. Carnegie Mello University for International Data Privacy, Pittsburgh, PA.

TSICHRITZIS, D. and LOCHOVSHY, F. (1982): *Data Models*, Prentice-Hall Inc., 5–14.

VERYKIOS, V., BERTINO, E., FOVINO, I., PROVENZA, L., SAGIN, Y. and THEODORIS, Y. (2004): State-of-the-art in privacy-preserving data mining: *SIGMOD Record*, 33(1): 50–57.

YANG, Z., ZHONG, S. and WRIGHT, R. (2005): Anonymity-preserving data collection: In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, Illinois, USA, 334-343

APPENDIX A: APPROACHES TO DATA COLLECTION SURVEY

QUESTIONNAIRE 1: Fixed-Level; Approach

1. What is your age?
2. What is your sex?
3. What is your marital status?
4. What is your race?
5. In what country were you born?
6. How are you related to the person who owns/rents the house you live in house?
7. What is the highest level of school you have completed or the highest degree you have received?
8. How are you employed?
9. What kind of work do you do, that is, what is your occupation?
10. How many hours per week do you usually work (main & other jobs)?
11. How much do you earn per annum? (An estimate)
12. How much did you approximately receive as interest on investments in 2005 (e.g. bonds, stocks, real-estate):
13. How much did you approximately lose on investments in 2005 (e.g. bonds, stocks, real-estate)

QUESTIONNAIRE 2: P-level Approach (Implemented using branching logic)

1. What is your age? I'd rather provide my …

a. Age in years
b. Age range: minor {0-17}; young-adult {18-30}; middle-aged{31-64}; senior{65-120}
c. Age group: young {0-30}; old {31-120}
d. I don't wish to provide an answer

2. What is your sex? I wish to provide

a. A specific sex: Male, Female
b. No answer

3. What is your Marital Status? I'd rather provide

a. My specific marital status: Married, spouse present; married spouse present; never married, divorced, widowed
b. A generic Yes/No answer: Married {Spouse absent, spouse present}; Not Married {Never married, divorced, widowed}
c. I don't wish to provide an answer

4. What is your race? I wish to provide

a. A specific Race: Black, White, Asian, Hispanic, Native, Other, Mixed, etc
b. A generic Yes/No answer: Mixed; Not Mixed
c. I don't wish to provide an answer

5. In what country were you born? I wish to provide

a. Specific Country
b. A Continent
c. No answer

6. **How are you related to the person who owns/rents the house you live in house? I wish to provide**
   a. A specific Relationship {Self, spouse, child, parent, partner, friend, etc}
   b. A generic relationship: family or friend
   c. I don't wish to provide an answer

7. **What is the highest level of school you have completed or the highest degree you have received? I'd rather provide a …**
   a. Specific Level of Education: G1-G6; G7-G12; Specific Technical Qualification; College Dipl.; University Degree; Professional Qualif., Masters; Doctorate
   b. Classification of Education: Elementary {G1-G6}; High School {G7-G12}; Technical Qualifications; Higher Education {College diploma, university degree}; Graduate School {Professional Qual., Masters, Doctorate}
   c. Yes/No answer: No Degree/Diploma; Obtained Degree/Diploma
   d. I don't wish to provide an answer

8. **How are you employed? I'd rather provide my …**
   a. Type of Organization: Private Organisation, Self-Emp-not-Inc, Self-Emp-Inc, Federal-gov, Local-gov, State-gov, Volunteer, Not-working.
   b. Type of Employment: Volunteer; Not working; Self employed, Employee
   c. Earning Status: Earning; Not earning
   d. I don't wish to provide an answer

9. **What kind of work do you do, that is, what is your occupation? I wish to provide**
   a. A Specific Occupation
   b. An Occupational Group: Health; Education; Business; Technical; Arts; Social Services, Entertainment, Agriculture; Legal, etc
   c. I don't wish to provide an answer

10. **How many hours per week do you usually work (main & other jobs)? I wish to provide**
    a. Number of hours
    b. A range of hours: Less than 35 hours or more than 35 hours
    c. No Answer

11. **How much do you earn per annum? (An estimate) I wish to provide**
    a. Estimated amount in dollars
    b. Select range of amount: less than 50k ; greater than 50k
    c. No Answer

12. **How much did you receive as interest on investment in 2005 (e.g. bonds, stocks, real-estate): I wish to provide**
    a. Estimated amount in dollars
    b. Select a range of amounts: Less than $5k; Btw $5k and $10k; Btwn $10k and $20k dollars; Greater than $20k
    c. No answer

13. Enter Amount of (investment) money lost in 2005: I wish to provide

a. Estimated amount in dollars
b. A range of amounts: Less than $5k; Btw $5k and $10k; Btwn $10k and $20k dollars; Greater than $20k
c. No answer

REVIEW OF METHODS

1. Which one of the data collection approaches did you find more convenient to use?

a. Method 1
b. Method 2
c. I don't know
d. They are pretty much the same

2. Which one of the data collection approaches did you feel gave you more control on how much information you were willing to disclose?

a. Method 1
b. Method 2
c. I don't know
d. They are pretty much the same

3. Which one of the data collection approaches you would prefer to use if you did not quite trust the data collector?

a. Method 1
b. Method 2
c. I don't know
d. They are pretty much the same

4. Which one of the data collection approaches do you feel encourages you to provide more accurate/truthful information?

a. Method 1
b. Method 2
c. I don't know
d. They are pretty much the same

BIOGRAPHICAL NOTES

Pele Williams

Pele Williams is a PhD candidate at the Computer Science Department, University of Calgary, Canada. She holds an MSc. in Computer Science from the University of Manitoba, Canada. Her research interests include privacy-preserving data mining, web workload characterization and network security. She is currently investigating methods to increase user-control in privacy-preserving technologies for private data. The P*-level model is proposed as part of her thesis research work, which is being supervised by Dr Ken Barker.*

Ken Barker

Ken Barker is a Professor of Computer Science at the University of Calgary with particular expertise in the area of database management systems. He holds a Ph.D. in Computing Science from the University of Alberta (1990) and has 25 years of experience working with industrial computer systems and in research. Dr Barker has published extensively in areas as diverse as distributed systems, software engineering, transaction systems, simulations and security. His current interests include developing data repository systems that provide privacy protection for data suppliers while allowing collectors to utilize the data within the guidelines explicitly agreed to by the provider at the time it was acquired. The research objectives include developing a privacy preserving database system and privacy preserving data mining strategies.