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

Abstract

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), on any attribute. Data collected at a high *P*-level signifies divulgence at a higher conceptual level and thus ensures more privacy. Providing guarantees prior to release, such as satisfying *k*-anonymity (Samarati 2001; Sweeney 2002) , 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.

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 contributors to divulge information anonymously as groups (Isitani *et al.* 2003).

Research into the privacy preferences and behaviors of data contributors 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 informational 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. 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.

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,

Copyright (c) 2007, Australian Computer Society, Inc. This paper appeared at the Australasian Information Security Workshop: Privacy Enhancing Technologies (AISW), Ballarat, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 68. Ljiljana Brankovic, Paul Coddington, John F. Roddick, Chris Steketee, Jim Warren, and Andrew Wendelborn, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

hiding and blocking, to the extent that the data miner cannot guarantee the data value is true or false.

Problem Statement: Data mining involves identifying patterns and trends in large quantities of data. Knowledge that is extracted during the mining process or released to the public as an aggregate could infer information that is at a lower privacy level than the divulged data. This paper addresses two questions:

- How can a data contributor/provider identify and control a reduction in the P-level of data contributed?
- How can a data collector/ miner provide guarantees for data collected in multiple P-levels? This involves identifying and controlling the P-level reduction phenomenon in acquired data sets.

1.1 Contributions of this Paper

This paper describes the use of concept hierarchies to enable 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 contributor and the data collector to control it. We expect that the application of these techniques will increase the control data contributors 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 formally defines the *P*-level model and describes scenarios that could result in *P*-level reduction using a motivating example. Section 3 discusses the methods for identifying and controlling *P*-level reduction while Section 4 describes related work in the field of privacy preserving data collection. Section 5 concludes this paper and provides some directions for future work.

2 Background

This section provides a background on concept hierarchies, the main technology that is applied to achieving the P-model. We present the P-model and describe scenarios for privacy invasion using a motivating example.

2.2 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 C*ity* 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 Milk >> 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):

$$H_p : D_i \text{ x } \ldots . \text{x } D_k \Rightarrow H_{p-1} \Rightarrow \ldots \Rightarrow H_0 \qquad (2.1)$$

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.

2.2 Defining the P-Level Model

A *P*-level, is the degree of generalization, *x*, of a piece of sensitive information *S*, such that, $x = \{0, 1, 2, \ldots h\}$ and *h* is the highest level of generalization possible on *S*. 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 *h,* can be denoted as S^{ph}. 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 $\{\text{Age}^{P1} = \text{middle-aged}\}$ is at a higher *P*-level than $\{\text{Age}^{P0} = 35\}$.

Formally:

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

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:

$$A_j^{Ph} \Rightarrow A_j^{P(h-1)} \Rightarrow \ldots \Rightarrow A_j^{P0} \qquad (2.2)$$

Let $t_i\ [A_j^{px}]$ represent the value of an attribute A_j for a record t_i belonging to the i^{th} data contributor at *P*-level *x*.

Let the *P*-level of value $t_i\ [A_j^{px}]$ be fully represented as x_{ij}.

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

$$t^i = \sum_{k=1}^{k=m} x_{ik} \qquad (2.3)$$

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

$$A^j = \sum_{k=1}^{k=n} x_{jk} \qquad (2.4)$$

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**	

Figure 2a: Data collection at multiple P-levels

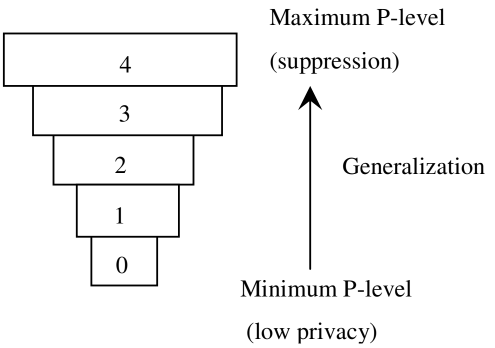

Figure 2b: Achieving Privacy using P-levels

Figure 2a and 2b 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 contributor (t_5). Obtaining information on the sensitivity of data attributes and privacy preferences of data contributors will provide an understanding of the nature of data collected and trust levels of data contributors.

2.2 Scenarios for *P*-level Reduction

The multiple *P*-level data collection approach provides the data contributor 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 contributor'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.

2.3.1 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 3 *P*-levels (P_0, P_1, P_2) using appropriate rules as shown below. Data contributors can provide data at any of the specified *P*-levels.

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

A_1 rules [P_0 ==>P_1]: {0-17}→minor; {18-30}→ young-adult; {31-64}→ middle-aged ; {65-120} → senior

A_1 rules [P_1 ==> P_2]: {minor, young-adult} → young; {middle-aged, senior}→ old

A_1 rules [P_2 ==> P_3]: {young, old} → Any

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

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

A_2 rules [P_1 ==> P_2]: {Elementary, High School}→No Degree; {Higher Education, Graduate School} →Degree

A_2 rules [P_2 ==> P_3]: {No Degree, Degree}→ Any

Address (A_3):= P_0 [{Street, City, Province, Country}]==>; P_1 [{City, Province, Country }] ==> P_2[{ Province, Country }] ==> P_3 [{Country}] ==> P_4 [{Any}]

A_3 rules [P_0 ==>P_1]: {Street, City, Province, Country} → City, Province, Country

A_3 rules [P_1 ==> P_2]: {City, Province, Country}→ Province, Country

Table 1: Data Collected at Multiple *P*-levels

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}	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}

A_3 rules [P_2 ==> P_3]: {Province, Country}→Country

A_3 rules [P_3 ==> P_4]: {Country}→Any

Given two data contributors, 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 $t_{2.}$ 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 $t_{4.}$

2.3.2 Defining *P*-level Reduction

A *P*-level reduction occurs when any of the data collectors can infer information at a privacy level below that which a data contributor agreed to divulge in the collected data set. A *P*-level reduction also occurs if an external adversary, C_3, is able to infer information (below the P-level of divulgence) belonging to data contributors Alice and Bob by linking data collected from sources C_1 and C_2 . We demonstrate three possible attacks on data collected at multiple *P*-levels using the data in Table 1.

A. External Adversary Linking Attack

This category of *P*-level reduction describes instances where an external adversary (and not the data collectors) has access to the collected data set.

Background Knowledge: Assuming the adversary, C_3, has some background knowledge of Alice's Address. For instance, C_3 knows that Alice lives in {Tuscany, Calgary, Alberta, Canada}, and is interested in determining her education level. Supposing C_3 has access to data collected by C_2, then C_3 might be able to infer that Alice has attained graduate level education by noting Alice is the only data contributor in C_2's data set that has specified Tuscany, Calgary, Alberta, Canada. Note, however, that C_3 cannot infer anything at a privacy level below which Alice divulges.

Lack of Diversity: When 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 3 people (including Bob) with the {High School, Canada} value combination. Supposing that the other 2 people both specify an age of {28} for the age category, C_3 can confidently infer that Bob is 28 years old."

B. Collector-Linking Attack

This category of *P*-level reduction describes instances where multiple data collectors combine collected data sets to infer information at a lower privacy level than was divulged to them.

Colluding Collectors: Two malicious data collectors, C_1 and C_2, can collude on the information t_1 and t_2 respectively 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: Two data collectors, C_1 and $C_{2,}$ respect Alice's privacy preferences and are not malicious. They would, however, like to combine their data sets to build a stronger classification tool. 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. 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.

C. Attribute-Linking Attack

In this category, we demonstrate scenarios where a data contributor 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 contributor 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 ==> 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.

3 Methods for Preserving *P*-levels

This section presents formal definitions and algorithms to identify and control the *P*-level reduction privacy breaches demonstrated in the previous section.

3.1 Detecting P-level Reduction

Background Knowledge: Inference of information in lower privacy levels through background knowledge is possible when attribute values are unique at the lowest *P*-levels in any data set. For example, if Alice is the only data contributor in C_2's data set that has specified {(Tuscany, Calgary, Alberta, Canada)} as noted above.

Formally, background knowledge disclosure occurs if:

$$n\,(\,[A_j^{pk}\,] = R) \;\leq\; 1 \qquad\qquad (3.1)$$

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 background knowledge, the value of $n\,([A_j^{pk}\,] \;= R)$ must be greater than 1. Inference is greatest when $k = 0.$

Lack of Diversity: A *P*-level reduction through a lack of diversity can be detected if all records at a certain privacy level have the same value for the sensitive attribute.

Formally, a lack of diversity inference occurs if:

$$\exists\, t: t= \{\, t_1, t_{2,\,........}\, t_i \,\} \;\subset\; T = \{t_1, t_{2,\,........}\, t_n\}:$$

$$\forall\; t\;,\; t\,[A_j^{px}\,] \;=\; R \qquad\qquad (3.2)$$

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.

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

$$\exists a{:}a = \{\, A_1,A_{2,\,........,}\, A_k \,\}\; : A_1,\sim A_2\; \sim_{,\,......}\sim_{..}\, A_k\,,$$

$$\forall A_j\;\;\varepsilon\;\; a\,,\;\exists\, \mathrm{t_i} \;\subset\; T$$

$$\text{if}\;\; x_{ij} = R\,,\;\; \text{then}\,\{\, x_{i1}\;, x_{i2,}\;\; \ldots,\, x_{ik}\,\}$$

$$\text{can be reduced to } R \qquad\qquad (3.3)$$

where $x_{ij} = R$ is the privacy level of an attribute A_j (of a record t_i) which belongs to a set $a = \{\, A_1,A_{2,.....,}\, A_k\,\}$ of correlated attributes.

Analysis Linking: Here, the likelihood of *P*-level reduction is associated with the accuracy of the data mining model which C_1 is using for prediction. Here, 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 $A_j^{px}\;\; ==>\; A_k^{px\,-r}$

$$Pr(x_k\; - r\,)\;\leq\; Q \qquad\qquad (3.4)$$

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 .

Colluding Collectors: An inference can occur with two malicious data collectors, C_1 and C_2 if both data sets T_1 at C_1 and T_2 at C_2 contain at least one unique attribute value, R at privacy level 0 on the same attribute with which t_1 at C_1 and t_2 at C_1 can be linked with certainty, as belonging to a data collector Alice. Formally, this can be represented as:

$$n\,(\,[A_j^{p0}\,]\; at\; C_1\;\; =R\,)\;\; =$$

$$n\,(\,[A_j^{p0}\,]\; at\; C_2\;\; =R\,)\;\;\; = 1 \qquad\qquad (3.5)$$

where, $n\,(\,[A_j^{p0}\,]\; at\; C_1\;\;\; =R\,)$ denotes the number of tuples at site C_1 , that have the attribute A_j set at privacy level $x = 0,$ for which R is an attribute value.

Cooperative Analysis: 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 $\mathrm{T_1}$ can improve prediction ability (or actually discover) a sensitive value $A_{ij}^{p(x\text{-}1)}$ after access to data T_2, collected by a data Collector C_2. Formally, a *P*-level reduction occurs if:

$$Pr\{(A_{ij}^{px}\;\; ==> A_{ij}^{p(x\text{-}1)})\,:\, T = \;T_n\,\}$$

$$<\;\; Pr\,\{(A_{ij}^{px}\;\; ==> A_{ij}^{p(x\text{-}1)}\;\;):$$

$$T = \;T_n\; + T_m\,\} \qquad\qquad (3.6)$$

3.2 Controlling P-level Reduction

This section describes formal methods to control *P*-level reduction during analysis, using the six *P*-level reduction scenarios described in the previous sections. First, we describe measures to protect a data contributor providing data to a malicious data collector at multiple *P*-levels.

The second sub-section applies to honest data collectors who intend to protect the data with which they have been entrusted, perform cooperative analysis with other data collectors, and provide guarantees of non-disclosure for released data sets.

3.2.1 Avoiding *P*-level Reduction at the Collection Stage

The data contributor is limited during privacy decision making by not having access to the data contributor's entire data set and mining model (incomplete information). In a detailed study, Acquisti 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). Therefore, we propose that an individual's decisions to avoid *P*-level reduction be guided by a software agent, such as a privacy critic (Ackerman and Cranor 1999). A software agent can suggest rational options and provide *P*-level reduction warnings based on a pre-defined set of actions such as described below:

Background Knowledge: Avoid disclosure on a low *P*-level on the set of potentially identifying attributes (attributes that can be used in combination for identification). Privacy critics (Ackerman and Cranor 1999) have been implemented to detect and flag potential identifying attributes. Formally,

$$\ni \; a = \{ A_1, A_2, \ldots\ldots\ldots, A_k \} \subset A = \{A_1, A_2, \ldots\ldots\ldots, A_m\} \; :$$

$$a \subset PID, \quad \forall A_j \; \mathcal{E} \; a \mid x_j > 0 \qquad (3.7)$$

Where *PID* is the global set of potentially identifying attributes and *a* is a subset of *PIDs* that exists in *A*

Lack of Diversity: Divulge insensitive attributes at *P*-levels at which you are willing to tolerate a disclosure of sensitive attributes.

$$\exists a : a = \{ A_1, A_2, \ldots\ldots\ldots, A_k \} \subset A = \{A_1, A_2, \ldots\ldots\ldots, A_m\}$$

$$\exists b : b = A - a \; ,$$

$$If \; \exists A_k : A_k \; \mathcal{E} \; b \quad and \quad x_k = R$$

$$If \; \exists A_j : A_j \; \mathcal{E} \; a \quad and \quad x_j = Q$$

Then the allowed P- level reduction is,

$$\mathrm{P} = Q - R \qquad (3.8)$$

where *a* & *b* is the set of sensitive and insensitive attributes respectively

Data Linking: Identify the set of correlated attributes and disclose both sensitive and non-sensitive attributes at the same *P*-level.

$$\ni \; a = \{ A_1, A_2, \ldots\ldots\ldots, A_k \} \subset A = \{A_1, A_2, \ldots\ldots\ldots, A_m\} \; :$$

$$A_1, \sim A_2 \sim, \ldots\ldots \sim_{..} A_k, \; \forall A_j \; \mathcal{E} \; a \; : \; \text{if} \; x_j = R$$

$$then, ==> x_1 = x_2 = \ldots\ldots = x_k = R \qquad (3.9)$$

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

Analysis Linking: 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, predication 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. A protection strategy will be to disclose at one *P*-level higher than the data contributor's P-level of comfort with sensitive attributes.

Colluding Collectors: The probability of contributing a unique sensitive value is higher at lower P-levels. Data contributors can protect themselves against inference through the linking of a unique sensitive value by disclosing on at least *P*-level 1.

Cooperative Analysis: Divulge only if collecting sites guarantee no direct data sharing, and even then, disclose on at least *P*-level 1.

3.2.2 Avoiding *P*-level Reduction at the Preprocessing and Data Mining Stage

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:

1. *Background Knowledge:* Achieve *k*-anonymity (Sweeney 2002) on the set of potentially identifying attributes; where *k* is the guaranteed level of anonymity.
2. *Lack of Diversity:* Achieve *l*-diversity (Machananvajjhala *et al.* 2005), where *l* is the number of pieces of background information needed before the adversary would achieve a *P*-level reduction.
3. *Data Linking:* For the set of correlated attributes, on any record, set all attributes values to the highest *P*-level indicated.
4. *Analysis Linking:* Apply rule hiding techniques (Verykios *et al.* 2004) to rules that reveal sensitive data
5. *Colluding Collectors:* Set all *P*-level 0 data to *P*-level 1 to prevent data linking in the event of a disclosure of the data set.
6. *Cooperative Analysis:* Cooperate using the secure multi-party computation model (Clifton *et al.* 2002).

4. 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. 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 by using a lower conceptual level. While suppression implies that no value is released.

Machananvajjhala *et al.*(2005) propose *l*-diversity, an extension of k-anonymity which overcomes the privacy breeches inherent in *k*-anonymity when the adversary has background knowledge and when there is homogeneity in the values of sensitive attributes. 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.

Yao *et al. (2006)* highlight uncertainty and indistinguishability as two independent aspects of privacy. Uncertainty refers the inability of an intruder to select from a group of values the private value of an individual, while indistinguishability is the inability of an intruder to discern the differences among a group of people. *k*-anonymity provides indistinguishability in anonymized tables. The key contribution of this paper is defining indistinguishability for general situations (such as multiple database views) based on the symmetry among the possible private values associated with individuals. Symmetric indistinguishability (SIND) is accomplished when a released data set remains unchanged after two symmetrically indistinguishable (on the quasi-identifier) tuples exchange their private values while keeping other tuples unchanged. *k*-SIND is achieved if each SIND set has a cardinality of at least *k.* SIND is a strict metric because it requires symmetry in all possible private values. RSIND is a more flexible metric which based on a subset of private values (including the current private value). The authors also provide practical algorithms (and their computational complexities) for checking whether a set of database views provide enough indistinguishability.

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 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.

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 breech 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 decision 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 success of selective divulgence for data mining is based on the premise that data contributors 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).

5. Conclusion

Data contributors 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 contributor and the data collector.

5.1 Future Work

This research is preliminary, thus there are many avenues for future work. We intend to experimentally validate the effectiveness of the *P*-level model approach to data collection by comparing the mining accuracies of data collected at multiple *P*-levels with data collected at fixed levels. Interfaces that allow data collection at multiple privacy levels will be required for various types of data.

Issues such as usability and effectiveness are challenges that must be met before this collection approach can be widely applied. There will be a need to provide data preprocessing techniques for data collected at multiple *P*-levels to achieve optimal data utility to ensure cleaner data sets and higher mining accuracies. Finally, we will experimentally validate the effectiveness of the *P*-level reduction techniques proposed in 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), S*hort Papers (v.2.), 1999, Pg. 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, Pg. 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, Pg. 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, Pg. 708-719.

D. Agrawal and C. Aggarwal, (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, Pg. 247-255.

C. Clifton, M. Kantarcioglu, and J. Vaidya, (2002): "Defining Privacy for Data Mining", In *National Science Foundation Workshop on Next Generation Data Mining :* H. Kargupta, A. Joshi, and K. Sivakumar, Eds., Baltimore, MD, November 2002, Pg. 126-133.

P. Fule and J. Roddick, (2004): "Detecting Privacy and Ethical Sensitivity in Data Mining Results": In *Proceedings of the 27th conference on Australasian Computer Science*, Volume 26, 2004. Pg. 159-166.

J. Han and Y. Fu, (1995): "Discovery of Multiple-Level Association Rules from Large Databases": In *Proceedings of the 21st International Conference on Very Large Data Bases,* September, 1995, Pg. 420-431.

J.Han and M.Kamber, (2001): "*Data Mining: Concepts and Techniques*": Morgan Kaufmann Publishers, San Fransico, USA, Pg. 56-58.

L. Ishitani, V. Almeida and W. Meira, (2003): "Masks: Bringing Anonymity and Personalization Together": *IEEE Security & Privacy*, Volume 1, Number 3, May/June 2003, Pg.18-23.

D. Jutla and P. Bodorik, (2005): "Sociotechnical Architecture for Online Privacy": *IEEE Security and Privacy*, Volume 3, Issue 2, March-April 2005, Pg. 29- 39.

Machananvajjhala, J.Gerke and D. Kifer, (2005): "*l*-Diversity: Privacy beyond *k*-Anonymity": Technical Report, Cornell University.

P. Samarati, (2002): "Protecting Respondents Identities in Microdata Release", *Knowledge and Data Engineering, IEEE Transactions,* Volume 13, Issue 6, November-December 2001, Pg. 1010 – 1027.

L. Sweeney, (2002): "Achieving *K*-Anonymity Privacy Protection using Generalization and Suppression, "*International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*": Volume 10, Issue 5, Pg. 571–588.

V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Sagin and Y. Theodoris, (2004): "State-of-the-art in Privacy-Preserving Data Mining": *SIGMOD Record*, Volume 33, Number1, March 2004, Pg.50 – 57.

Z.Yang, S. Zhong and R.Wright (2005): "Anonymity-Preserving Data Collection": In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining,* Chicago, Illinois, USA, Pg. 334-343.

C. Yao, L. Wang, S.Wang and S. Jajodia (2006): "Indistinguishability: The Other Aspect of Privacy": In *Proceedings of the Third VLB Workshop on Secure Data Management (SDM)* 2006, Pg. 1-17.