数据挖掘对数据库安全的影响

陈伟鹤 殷新春 张 锐 谢 立

(南京大学计算机软件新技术国家重点实验室 计算机科学与技术系 南京210093)

Influence of Data Mining on Database Security

CHEN Wei-He YIN Xin-Chung ZHANG Rui XIE Li

(State Key Lab for Novel Software Technology, Nanjing University, Nanjing 210093)

(Dept. of Computer Science and Technology, Nanjing University, Nanjing 210093)

E-mail: chenweihe@dislab. nju. edu. cn

Abstract Main focus of database security research is maintaining data security and providing timely data services. To ensure security, inference attack and aggregate attack are the problems, which should be solved. But these will leave volumes of data to discover some useful knowledge, and some data mining methods used in this field will bring forth some positive or even negative results. In this paper, we survey what data mining has done on database security and present some open problems and possible future research directions.

Keywords Data mining, Database security, Intrusion detection, Audit, Inference attack, Aggregate attack

1. 引言

现代社会信息不断膨胀。在各种数据库、报刊、杂志、网络等媒体上,充斥着各种各样的信息。这些信息中既存在许多过时的、错误的或冗余的信息,也有许多有用的数据,包括一些敏感的数据,或是一些本身不敏感,但可以利用推理、归纳等逻辑手段从这些数据得到数据提供者本不想公开的隐私(private)数据。如何满足既利用好这些数据,同时又能维护数据提供者的隐私(privacy)要求已经成为当前数据库安全和数据挖掘研究中一个极为迫切的问题。

数据库安全研究的内容是如何保证数据库中数据的安全,同时对合法用户提供及时的数据服务,拒绝非法的数据访问请求。

考虑到数据库管理系统作为实现信息存储、共享、维护数据一致性、减少数据冗余的有效手段在信息处理中所起的核心作用,以及其在各种场合的广泛应用,在本文中我们探讨数据挖掘对数据库安全的影响。

一方面,数据挖掘[1,2]是用来对海量数据进行处理,从中抽取和发现感兴趣知识的有效手段。它可以使用户及时、准确地得到所需要的信息。另一方面,为了满足保护信息拥有者的隐私和利益的要求,需要有一些技术手段,它们能够防止用户利用数据挖掘工具对数据库系统提供的数据进行挖掘,得到数据提供者本来不想泄漏的某些敏感信息。因此数据提供者希望控制用户对其所提供数据的使用。显然,这两方面的要求是矛盾的。

数据挖掘对数据库安全的影响可以分为正面影响和反面影响两个方面。它对数据库的正面作用是指数据挖掘可以用来提高数据库的安全性,即数据库管理系统的拥有者在数据库管理系统的审计信息处理和入侵检测中可以利用数据挖掘,从而改进和提高系统的安全性。对数据库的反面作用是指,数据挖掘也完全可以被攻击者用来对数据库系统进行推理攻击和聚集攻击,偷取数据,破坏系统的安全。

2. 数据挖掘对数据库安全的正面影响

2.1 数据挖掘在入侵检测和审计中的应用

入侵检测(intrusion detection)是指监控系统日常运行中发生的各种与系统安全相关的事件,发现任何试图破坏系统功能或是获得对数据库的非法访问权限的行为。入侵检测方法可以分为基于统计规律的(profile-based)和基于已知攻击方法的(signature-based)。为了保证系统的正常运行,监控用户和系统的行为,数据库系统使用时会产生大量的审计数据。如何充分利用这些数据,得到与系统安全相关的信息,往往是数据库系统拥有者所关心的,而这正是数据挖掘的特长。数据挖掘在实现数据库的入侵检测和审计安全威胁方面有广泛的应用。

人们可以利用数据挖掘发现入侵的模式和多种安全威胁,这是信息安全保障(Information Assurance)所强调的。信息安全保障的研究重点在于不仅要有高质量的数据,而且要求能从恶意的攻击中恢复,并且能反击对数据的攻击(包括窃取数据和篡改数据)。而信息安全保障的一个重要前提是有高效、准确的入侵检测技术,使信息系统能够及时发现恶意攻击。只有这样才能对入侵进行限制(intrusion confinement),防止破坏的进一步扩大,并从恶意攻击中恢复[3]。

2.2 发现数据库系统中存在的异常数据和模式

数据库管理系统可以用数据挖掘工具发现系统中所存储的数据可能存在的某些异常现象。例如,公司某个雇员到一个城市出差的次数过于频繁。如果这个数据库所具有的数据挖掘工具发现了这样一种数据异常的迹象,它就能对数据库构造出一个查询:即该员工在这个城市是否有往来比较密切的人?如果查询结果显示此员工在该城市确实有关系比较密切的人,那么就发现了该员工的一个异常行为。另外,通过数据挖掘也可以发现偏离正常取值规律的某些字段值,提醒用户,如果这些字段值确实发生数据值错误,系统对其的及早发现就能防止错误的进一步蔓延。

上面所涉及到的都是数据挖掘在数据库安全中的正面应

用，但是正如前面所提到的那样，数据挖掘工具是一把双刃剑，它完全也可以被攻击者用来破坏数据库系统的信息安全。

3. 数据挖掘对数据库安全的威胁和一些解决方法

3.1 数据挖掘在推理和聚集攻击中对数据库安全造成的危害

数据库安全面临的一类主要问题就是推理攻击(inference problem)和聚集攻击(aggregate problem)问题。它们是指用户在某个数据库上提出几个查询，得到相应的返回信息，然后从这些返回结果中推导出一些该用户本来无权访问的信息。这个问题在过去的几十年里进行了广泛的研究，然而数据挖掘的出现使该问题变得比以往任何时候都严重。这是因为在数据挖掘出现之前，任何有推理攻击企图的人必须依靠个人的经验和知识来分析得到的数据，并从中找到感兴趣的东西。但是现在的数据库用户有了功能强大的数据挖掘工具，它们可以用来对得到的数据库查询结果进行高效、简洁、智能化的自动处理，从中发现有价值的信息或模式(包括一些用户根本没有意识到的、潜在的、感兴趣的规律和数据预测)。

推理和聚集攻击一般是由系统的合法用户发起的。攻击的过程如下：攻击者首先确定希望推理出的信息；接着判断进行推理所需的信息，据此构造若干个查询；然后登录进数据库系统，发出查询指令，得到系统返回的结果；分析所得到的系统响应，推断出所需的敏感信息。举一个简单的例子，设某公司有一个存放有公司雇员姓名，工号，月工资的数据库，雇员的姓名和雇员月收入这两个信息分开来看都是公开的，但是两者放在一起时却是一个需要保密的敏感信息(假定公司认为公司雇员的月工资收入是一条需要保密的信息)。因此这个数据库将这些信息分成两张表来存放，一张表是员工姓名和工号，另一张表存放的是员工的工号和月工资(系统不允许用户直接根据员工姓名查询得到员工的工资信息)。一个用户可以通过两个查询得到他所想知道的某个特定员工的月收入信息。第一个查询检索想知道其工资的那个员工的姓名和工号；第二个查询检索工资和工号。然后在这两组信息间做一个匹配处理就可以得到他希望得到的那名员工的月工资信息了。

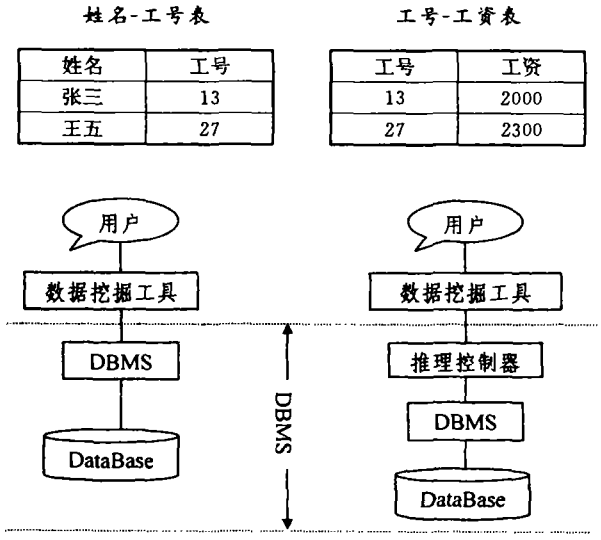

图1 (a)数据挖掘导致推理攻击 (b)推理控制器防止数据挖掘引起的推理攻击

自20世纪70年代开始，由于美国中央统计局对统计数据库的安全问题感兴趣，人们开始在统计数据库上研究推理和聚集攻击问题。到70年代中期，美国国防部开始了对多级安全数据库的研究。自此，原来对统计数据库推理和聚集攻击问题的研究也就扩大到安全数据库领域。这方面的主要研究有Morgenstern[4]，Thuraisingham[5]和Hinke[6]等。

下面的一个例子说明数据挖掘工具可以用来破坏数据库系统的安全。在图1(a)中，用户能够构造多个查询，并且能够应用多种数据挖掘工具分析得到的查询结果，从而获得敏感信息。这是一个合法用户恶意利用数据挖掘进行推理/聚集攻击的典型例子。

3.2 相应的解决方法

研究者已提出多种方法用来解决数据挖掘在推理和聚集攻击中可能造成的危害。

推理攻击和聚集攻击这两类问题从本质上说是由计算机所建模的实际应用本身的特点所决定的，即信息泄漏是由应用的语意造成的。用户往往只要了解这种由计算机来处理的应用的实际背景，并具有一些与此应用有关的常识(社会常识或者是对于这个特定应用的特定领域的知识)，就能根据系统所提供的符合数据库系统安全策略的数据，获得数据库系统想要对此用户隐藏的信息。下面是一些解决方法。它们可以分成五种，即：基于推理控制器的方法、基于已有数据挖掘算法的方法、针对数据挖掘所涉及到的数据的方法、研究和开发保护个人隐私数据的数据挖掘算法、基于审计的方法。

3.2.1 *基于推理控制器的方法* 建立一个推理攻击监控器，用来发现用户的攻击企图，防患于未然，阻止攻击的形成。这个推理攻击监控器一般处于数据挖掘工具和数据源或者数据库之间。结构如图1(b)所示[2]。它可以进一步分为基于演绎逻辑的推理控制器和基于归纳逻辑的推理控制器。

Thuraisingham[8]提出了一种基于演绎逻辑来处理推理攻击问题的方法。她提出了安全约束(security constraints)的概念，在数据库管理系统中增加了一个推理引擎(inference engine)模块，它的主要功能就是进行逻辑推理，并依据推理结果判断是否违反了安全约束。工作过程如下：推理引擎对将要流出系统的数据进行推理，得到一些新的数据，如果推理得到的数据违反了系统的安全策略，那么这些本来要流出系统的数据就不提供给用户。

Thuraisingham[2]也探讨了采用基于归纳逻辑的方法来处理推理攻击问题的思路。

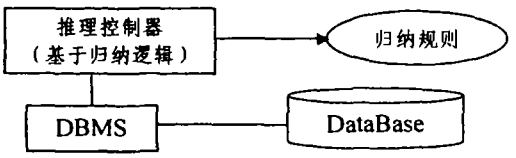

图2 基于归纳逻辑的推理控制器

图2是她给出的一个这种推理控制器的体系结构图。它查询数据库，得到系统响应，通过对系统返回的查询结果进行归纳得到一些规则。所得到的这些规则可能是敏感的或可能导致敏感信息的泄漏。系统采用约束(constraint)的形式定义规则是否敏感或可能导致敏感信息。如果推理控制器发现一条敏感规则，那么它就通知系统安全管理员这些数据的公开将会导致潜在的安全问题，需要对这些数据的安全级别进行重新划分。同样，它也不能彻底解决各种新出现的数据挖掘工具对数据库安全构成的挑战，达到完全防止用户得到无权访问

的信息的目的，因为各种数据挖掘方法不断被研究出来。它的作用只是提示系统安全管理员系统中某些数据可能面临的安全威胁，达到增强系统保护数据安全能力的目的。

3.2.2 *推理和聚集攻击方法的分类* 一是面向特定数据挖掘算法的方法；二是从数据挖掘算法的基本理论入手。前者针对特定的数据库和某一个数据挖掘工具，数据库系统采用相同的数据挖掘工具，再响应用户查询请求把合法的查询信息提供给用户时，系统也同时对这些信息用相同的数据挖掘工具进行挖掘，如果挖掘的结果是能够从这些信息中推理出敏感信息那么就认为存在推理攻击问题，系统可以采取相应的措施来保护敏感数据(如改变数据的安全级别设置等)。

但是这个方案有许多问题。首先它假定用户仅使用一种数据挖掘工具，而且数据库系统的拥有者知道这种工具是什么，并拥有此数据挖掘工具。而实际上数据库管理系统只能对系统内的信息进行保护，一旦信息提供给了用户以后，系统是难以控制用户对这些信息的使用方式的。比如，系统就根本无法对用户所能使用的数据挖掘工具进行限定。而且，构成推理攻击的方式是难以全部考虑到的。具体的可以参考文[7]。

Clifton 提出了一种对抗基于数据挖掘的推理攻击的方法[7]，它是从数据挖掘算法的基本理论考虑的。其基本思想是：通过一定的手段使攻击者对其使用的数据挖掘工具的挖掘能力产生怀疑，这样他对这个数据挖掘所得到结果的可靠性也就产生了怀疑。那么怎样才能达到破坏攻击者对所使用数据挖掘工具的目的呢?一种方法是只给攻击者足够少的例子，使他难以用这些数据建立一个好的分类器。那么问题就相应转化为究竟多少的数据样本才是足够少的?Clifton 利用分类理论(Classification Theory)来判定可以泄漏给攻击者的样本数据的上限，但是这方面的研究才刚刚开始，有许多问题需要解决。比如，数据库系统可能泄漏多个样本数据给不同的用户群，这些用户群可能勾结起来，利用所得到的多个样本数据建立一个效果不错的分类器。要想解决这个问题，数据库管理系统必须能够对提供给用户的信息进行追踪控制。而这显然是一件困难的事情。

3.2.3 *数据预处理的方法* 这个方法可以进一步划分为对数据库系统中的数据值进行变换处理的方法，以及对数据的安全级别按保护要求进行敏感级别定义和敏感级别提升的方法。

针对数据库系统中的数据值进行变换处理的方法，Clifton 等[7]讨论了几种防止对数据“过分”挖掘的方法。

(1)对数据增加“噪音”。这是美国统计局所采用的方法。它的基本思想是通过对数据值进行修改，使对这些被修改了的数据所进行的数据挖掘难以得到有用的结果。

(2)消除数据中的附加信息。一些数据由于其产生方式等原因往往具有一些隐含的其它信息，如果某些用户知道这些数据所含额外信息的规律，就可以对其进行利用，得到许多其它信息。例如，美国公民的社会安全号码(social security number)是按照发证机构的序号来编码的(即社会安全号码的前三位数字是发证机构的号码)。且这些发证机构对所发出证件是按顺序进行编码的。了解了上述规律，就可以根据社会安全号码的前三位数字对拥有社会安全号码的人进行按地区的分组(或更进一步可以对这些人按年龄进行分组，因为证件号码是按顺序进行编码的)。其实即使不知道社会安全号码的生成规律，它所具有的一些特征在数据挖掘中也是有用的。比如，通过数据挖掘发现社会安全号码的前三位数字具有较高的出现频率，就可以据此对社会安全号码进行分类。这样，在这些分类的基础上我们可以进一步发现某些现象(具有类似社会安全号码的人的癌症发病率较高)，这就完全可能引起启发用户进行进一步的数据挖掘(比如，这些人又具有那些共同点?)。这就可能导致侵犯个人隐私的问题。再如，以公司电话号码簿为例，如果知道该公司的电话号码是根据电话在公司大楼里的位置而定的，另外再知道存在这样一条规则：即在同一项目组工作的人，他们的电话号码具有一定的规律。那么，我们可以通过发现公司雇员电话号码的规律来将雇员进行分组。从而知道(大约)有几个项目组存在。对于某个需要保密的敏感项目组，可以根据每个组中成员的专业背景和工作性质进一步推断出这些项目组所进行工作的大致内容是关于什么方面的。要解决这个问题，一个办法就是尽量减少数据所隐含的附加信息。只有这样才能防止对其进行数据挖掘，使对其的使用不至于偏离它的本来目的。

(3)故意增加错误数据。在某些应用里，可以在原有的数据中有意识地引入一些错误的数据。在我们很明确地知道这些数据将会被如何使用的前提下，我们可以在系统数据中添加一些起误导作用的数据，它们在正常的系统许可的使用方式下是不会遇到的，而在一些不“合适”的查询里可能会碰上，它们既能误导攻击者，又能作为系统被攻击的标志。比如，在电话簿里添加一些不存在的人名。如果这本电话簿的使用者知道公司里确实有某个人，就可以得到他的电话号码。反之，如果用户不知道公司里的情况，只是凭借这本号簿，那他就完全可能查到的是某个根本不存在的人的所谓的电话号码。如果这个公司是通过接线员来进行人工转接的话，在这个外部人员要求接线员转接到某个根本不存在的分机号(那个根本不存在的公司雇员的分机号)时就表示了一个安全事件。这种方法要求所增加的数据难以被系统外部用户所区分，否则这些人完全可以重新恢复原有的数据集。

针对数据的安全级别按保护要求进行敏感级别定义和敏感级别提升的方法，Steven Dawson 等[10]提出了一种对数据库中的数据按照其中信息敏感度进行分级的新方法，它不仅满足数据库系统对数据安全的直接要求(explicit classification)，而且考虑了防止推理攻击和聚集攻击的要求。文[10]中提出了一个形式化的框架对数据库中客体的安全级别所必须满足的约束进行描述。这些约束既有描述了单个数据对象必须满足的由系统安全策略所显式表达的安全要求(explicit requirement)，又有针对数据对象集合的关联约束(association constraints)和针对不同数据对象的安全级别之间所存在的逻辑联系的推理约束(interference constraints)，这部分约束定义了每个(或一组)数据对象必须满足的信息敏感度级别的下界。另外，考虑到信息的最大化共享和主体原有的知识，这个框架中还有一部分约束是用来描述可以赋给数据对象的信息敏感级别的上限的。这个方法的优点在于根据这组安全约束确定数据对象的信息敏感度级别时能保证数据的最大可见性。所谓数据的最大可见性是指数据在符合数据库系统的安全约束集合的基础上，能实现信息的最大共享。在文[10]中提出了极小(minimality)的概念描述这种既满足数据库系统的安全保护要求又不会把数据对象的信息敏感度级别定得过高的性质。在文[10]中给出了一个算法来实现这种对数据敏感度级别进行划分的方法，它具有低阶的多项式复杂度。

3.2.4 *研究和开发保护个人隐私数据的数据挖掘算法*

目前，在数据挖掘方面一个较为热门的研究方向就是维护隐私的数据挖掘技术。Agrawal 等[9]提出一种不影响个人隐私的数据挖掘方法。数据挖掘的实质是在聚集的数据上进行一定的运算，得到准确反映这些数据规律的数据模型，Agrawal 等考虑是否可以在不访问每个单独数据值(即不涉及隐私数据)的基础上就能得到一个准确的模型?在文[9]中，研究目标建立一个基于决策树的分类器，用来建立分类器的数据集中的每一个数据都被加入了噪声，而且这个经过噪声处理过的数据集中数据的分布也不同于未加噪声前的数据分布规律。这样尽管每个数据原来的值难以推断出来，但是却可以通过一个重构过程准确估计原有数据集所具有的数据分布规律，就能以此为基础建立分类器，它的精度可以与建立在未加噪声的数据集上的分类器相媲美。

3.2.5 *基于审计的方法* 通过审计发现合法用户对信息的滥用，系统进而可以采取一定的补救和惩罚措施。问题是审计系统该存储哪些东西?怎样发现用户在对数据库系统进行数据挖掘?流出数据库的数据构成了哪些推理攻击问题?对于审计数据又该怎样利用?

显然数据挖掘对数据库安全构成的威胁是由数据库所具有的信息共享特点造成的，如果数据库系统不向外提供信息，就不会发生推理攻击的问题。但是，这完全违背了数据库管理系统的根本目的。上述方法各有优缺点，在实际应用中可以将多种方法结合起来，保护数据库中数据的安全。

在数据仓库(Data Warehousing)里，推理和聚集攻击依旧是一个问题。比如，用户可以访问数据仓库中存储的平均工资，然后推断出数据源中的某些个人的工资。对数据仓库中的推理和聚集问题的研究还很少，在这方面有许多工作要做。

4. 需要解决的问题

(1)研究特定数据挖掘算法的技术细节。如果知道某个算法是如何定义它的兴趣度量(interest measures)的，那么可以通过使数据库系统向用户提供的数据不包括用户"感兴趣"的规则，或是通过提供一些经过处理的数据，使从这些数据中只能挖掘出一些伪规则，真正的规则却被这些数据隐藏了起来[7]。因此知道这些特定算法是如何生成规则的、兴趣度是怎样定义的、对低兴趣度项目集是如何剪枝的，我们就能采取有效的措施来对抗并削弱数据挖掘工具破坏数据库系统安全的能力。最理想的是能够找到各种数据挖掘算法的共同特性，这样我们就能提出一些通用的方法来避免由这些挖掘算法造成的安全威胁。

(2)研究不同数据挖掘算法，找到它们的特点(尤其是每个算法的适用领域和弱点)，就能判断数据库系统向外提供的数据集是否容易受到数据挖掘攻击。了解了算法的计算复杂性等特点，我们就能建立数据库系统，使其向外提供的数据或者难以为数据挖掘算法所处理(即对这些数据的挖掘或者难以得到理想的结果，或者挖掘得到的结果不会影响数据库中敏感数据的安全)。

(3)利用数据挖掘工具发现数据库中潜在的推理攻击可能。一个数据库只要它同时包含了公开数据和只向部分具有特定安全级别用户提供的敏感信息，就有必要采用此方法。因为，完全有可能存在一些规则，它们以数据库向外提供的公开数据作为逻辑推理的前件，经过推理运算就能得到一些敏感信息(后件)。而且，这可能会导致更多敏感信息泄漏。这是因为，如果一个普通用户知道这些公开数据中某些数据所表示实体的若干敏感信息，他就能够在这些公开数据上做进一步的数据挖掘。

(4)研究保护数据提供者隐私的数据挖掘算法。它研究的目标是提供给数据挖掘算法的样本数据即使是已经被加了噪音、甚至连原始数据集的分布规律都改变了，且不能重新恢复原始数据值的情况下，依旧能进行所需的数据挖掘。

(5)从数据而言，数据库系统必须做到只向外公开系统安全策略允许提供给用户的数据。要达到这个目的，我们必须准确知道用户能从数据库提供给其的数据集里得到哪些额外的信息?在某些情况下，数据库管理系统提供给用户的数据不必完全精确，甚至可以故意引入一些起误导作用的错误数据，具体的原因在前面一节已经讲过了。问题的关键是这些提供给用户的数据的准确度应该是多少?如何保证数据的准确度?解决了这两个问题我们就能够在数据库中有选择地引入一些错误数据，达到限制数据挖掘工具挖掘效果的目的。

结论 本文论述了数据挖掘与数据库安全的关系。显然，数据挖掘作为一种对海量数据进行处理，从中发现有价值规律、模式的手段，它既可以被数据库系统拥有者用来提高系统的安全性能，也可以被攻击者用来偷取数据库里存储的敏感信息，关键在于如何使用数据挖掘和数据库系统所提供的数据集，以及由谁来使用数据挖掘。最理想的情况是，既能够充分共享数据又能够避免由于数据共享可能导致的敏感信息泄漏。要实现这个目标，可以采用诸如：对用户可以访问的数据进行限制；对数据增加噪声；对将要提供给用户的数据，系统先进行数据挖掘，避免其中可能存在的关联和聚集关系；或是研究不破坏数据库系统中敏感信息安全性的新型数据挖掘算法等方法。

参 考 文 献

1 Han Jiawei, Kamber M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, Inc. 2001

2 Thuraisingham T. Data Mining Technologies, Techniques, Tools, and Trends. CRC Press, 1999

3 Liu P, Jajodia S, McCollum C D. Intrusion Confinement by Isolation in Information Systems. in Research Advances in Database and Information System Security, Vijayalakshmi Atluri, John Hale, eds. Kluwer Academic Publishers, 2000

4 Morgenstern M. Security and Interference in Multilevel Database and Knowledge Base Systems. In: Proc. of the 1987 ACM SIGMOD Conference, San Francisco, CA, June 1987

5 Thuraisingham B. Security Checking in Relational Database Systems Augmented by an Interference Engine. Computers and Security, Vol. 6, 1987

6 Hinke T. Interference and Aggregation Detection in Database Management Systems. In: Proc. of the 1988 Conf. on Security and Privacy, Oakland, CA, April 1988

7 Clifton C, Marks D. Security and Privacy Implications of Data Mining. ACM SIGMOD Workshop on Data Mining and Knowledge Discovery, Montreal, Canada, June 2 , 1996

8 Thuraisingham B. Design and Implementation of a Database Interference Controller. Data and Knowledge Engineering Journal, North Holland, 1993(8)

9 Agrawal R, Srikant R. Privacy-Preserving Data Mining. In: Proc. Of the ACM SIGMOD Conf. on Management of Data, Dallas, May 2000

10 Dawson S, et al. Maximizing Sharing of Protected Information. Journal of Computer and System Sciences, 2001