计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190500131

面向缺损数据的(α,k)-匿名模型

张王策　范　菁　王渤茹　倪　旻
云南民族大学电气信息工程学院　昆明 650000

摘　要　在数据集对外发布之前,需要对数据集的准标识符属性进行匿名,以防遭受链接攻击。然而现有的数据匿名算法都是面向完整数据进行,对于数据集中含有缺损数据的元组会进行直接删除操作,降低了数据的可用性。文中提出将缺损数据与完整数据混合匿名的算法,并且结合了(α,k)-匿名算法。实验得出的数据充分证明:改进后的面向缺损数据的(α,k)-匿名模型有效提升了匿名后数据的可用性,实现了数据匿名。

关键词:(α,k)-匿名模型;k-匿名;泛化/隐匿;缺损数据
中图法分类号　TP319.9

(α,k)-anonymized Model for Missing Data

ZHANG Wang-ce,FAN Jing,WANG Bo-ru and NI Min
School of Electrical and Information Technology,Yunan Minzu University,Kunming 650000,China

Abstract　Before a dataset is published,the quasi-identifier attributes of the dataset need to be anonymous in case of a link attack. However,the existing data anonymity algorithms are all oriented to complete data,and the tuples containing defective data in the data set will be deleted directly,which reduces the availability of data. In this paper,the missing data and intact data are mixed into an anonymous algorithm,and the (α,k)-anonymous algorithm is combined. The experiment data fully prove that the improved defective data oriented (α,k)-anonymous model effectively improves the availability of the anonymous data and realizes the data anonymity.

Keywords　(α,k)-anonymous model,k-anonymous,Generalization/concealment,Missing data

1　引言

随着数据分析技术的迅速发展以及越来越多数据集对外共享发布,研究人员可以方便快捷地利用这些数据,但隐私泄漏的风险也随之增加。例如,医院会收集患者的医疗信息并对外公开发布,以便于科研组织的需要。科研机构可以利用医院发布的数据集进行数据分析,以便于统计或者预测相关疾病趋势。但是对外发布的医疗信息可能包含患者的个人隐私。虽然在医疗数据对外发布之前,发布者会删除个体标识符,以防止别人获取到某个患者的敏感信息。但是通过不同数据集之间的链接攻击,仍会导致一些患者的隐私泄漏。文献[1]指出,即便在数据公开之前去掉标识符信息,攻击者依旧能够利用数据中的邮编、婚姻状况、工作等准标识符与其他发布的数据集结合,从而得到用户的隐私信息。正是由于存在这种攻击,越来越多的研究者开始关注数据匿名技术,许多数据匿名算法和数据匿名模型应运而生。

然而,目前绝大多数的数据匿名算法对于带有缺损数据的数据集都力所不及[2-5]。如果某个数据集中的数据存在缺损,算法通常会直接删除其中带有缺损数据的元组,然而在数据收集的过程中,数据缺损的情况是普遍存在的。例如,某机构采集大量人员的家庭收入以及其他信息,如果某些家庭收入较高,通常这些家庭成员会拒绝透露收入信息,那么在收集到的信息中,该家庭成员的家庭收入一项则为空,造成数据缺损。如果直接删除该家庭成员的所有信息,则会造成该家庭成员其他无缺损的信息丢失。但是实际上,我们可以采取更加巧妙的方法来处理这类缺损数据,即将缺损数据和正常数据一起进行匿名处理,采取某种特殊的聚类以及距离计算方法,来保留绝大部分数据。

2　匿名算法模型

定义 1(属性)　假设 $T(A_1,A_2\cdots A_n)$是包含有限个元组的一个数据表,T 的有限属性组是$\{A_1,A_2\cdots A_n\}$。

根据数据表中的属性,我们可以将其分为以下 4 类。

(1)显示标识符 EI(Explicit_Identifer)

通过显示标识符。能够直接确定个体身份的属性,例如工号、姓名、电话号等。通常在数据公开前,将显示标识符加密处理或者直接去除。

(2)准标识符 QI(Quasi_Identifier)

把准标识符联合起来就能够唯一确定个体身份的属性,比如 Gender,Education,Workclass 等。

基金项目:国家自然科学基金项目(61540063);云南省应用基础研究计划项目(2016FD058,2018FD055);云南民族大学校级教学质量工程建设项目(2018JWC-JG-30)
This work was supported by the National Natural Science Foundation of China (61540063), Yunnan Applied Basic Research Program (2016FD058,2018FD055) and School-level Teaching Quality Engineering Construction Project of Yunnan University for Nationalities (2018JWC-JG-30).
通信作者:范菁(fanjing9476@163.com)

(3)敏感属性 SA(Sensitive_Identifer Attribute)

敏感属性即敏感信息,例如疾病、家庭收入、家庭地址等。

定义 2(k-匿名) 如果数据表 T 中的任意元组 t 都能找到至少 $k-1$ 个元组与 t 在准标识符上无法区分,则该数据表满足 k-匿名。以表 1 为例,任意一个元组都能找到至少 3 个元组在准标识符上具有相同的值,所以该数据表满足 4-匿名。

表 1 4-匿名数据表实例

Table 1 Example of 4-anonymous data table

Num	Age	Workclass	Race	Sex	Occupation
1	[31,46]	Private Eterprises	White	Male	Craft-repair
2	[31,46]	Private Eterprises	White	Male	Adm-clerical
3	[31,46]	Private Eterprises	White	Male	Prof-specialty
4	[31,46]	Private Eterprises	White	Male	Sales
5	[22,50]	Worked	White/Black	Male/Female	Craft-repair
6	[22,50]	Worked	White/Black	Male/Female	Machine-op-inspct
7	[22,50]	Worked	White/Black	Male/Female	Prof-specialty
8	[22,50]	Worked	White/Black	Male/Female	Other-service

定义 3(泛化) 泛化是典型的数据匿名方法,其通常是将元组中某一准标识符属性中值用一个概括的、模糊的范围值代替。例如,某几个元组中的 Workclass 属性存在几个值:Federal-gov, Local-gov, State-gov,则将它们统一用 Government 来替代。这种匿名方法在分类树的结构上尤为适用,如图 1 所示,用分类树中更靠近根节点的节点来代替更靠近叶子节点的节点。当然泛化的程度越大,数据损失就越严重。泛化主要包含局部泛化和全局泛化。全局泛化的算法实现比较简单,但是匿名后造成的信息损失十分巨大。局部泛化的算法相对复杂,但是匿名后的数据集信息损失更低。为了使数据损失降到最低,本文将带有数据缺损的元组与数据完整的元组一起进行局部泛化处理。

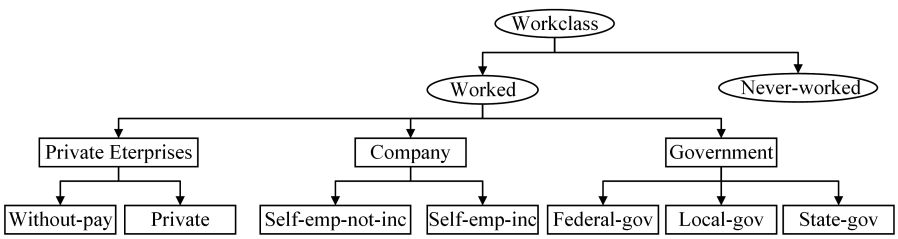

图 1 Workclass 分类树

Fig. 1 Classification tree of Workclass

定义 4(信息损失,$IL(e_l)$) $T(A_1,\cdots,A_n)$ 是一个数据表,$QI(N_1,\cdots,N_x,C_1,\cdots,C_y)$ 是准标识符,$N_i(1\leqslant i\leqslant x)$ 是第 i 个数值型属性,$C_j(1\leqslant j\leqslant y)$ 是第 j 分类属性,则信息损失[18]为:

$$IL(e_l)=\sum_{i=1,\cdots,x}|e_l|\frac{MAX_{N_i}-MIN_{N_i}}{|N_i|}+\sum_{j=1,\cdots,y}|e_l|\frac{H(\wedge(\cup C_j))}{H(T_{C_j})} \tag{1}$$

其中,$|e_l|$ 是聚类 e_l 元组的数量,$1\leqslant l\leqslant m$,$|N_i|$ 是第 i 个数值属性的范围,MAX_N 和 MIN_{N_i} 是聚类 e_l 中的最大值和最小值,$H(T_{C_j})$ 是分类树的高度,$H(\wedge(\cup C_j))$ 是具有最小公共祖先的分类子树的高度。

对于图 1 给出的关于 Workclass 的分类树,$H(T_{C_j})$ 为 3,$H(\wedge(Private,Without\text{-}pay))$ 为 1,$H(\wedge(Private,Self\text{-}emp\text{-}not\text{-}inc))$ 为 2。

特别地,如果数据表中某些信息出现缺损,例如表 2,缺损的数据会以"?"的形式出现,那么在计算信息损失时,就要把该缺损数据当作最高泛化层次来计算。例如表 2 中序号为 2 的元组中 Workclass 属性出现缺损,如果把这 4 个元组当作一个分组来进行匿名处理,该分组匿名后的信息损失是:

$$IL(e_l)=4\times\frac{46-31}{74}+4\times\frac{3}{3}+4\times0+4\times1=8.81$$

另外,如果匿名模型直接删除带有缺损数据的元组,那么信息损失将为 $|e_l||QI|$,即删除该元组造成的信息损失为准标识符属性的个数与删除元组数的乘积。例如对于表 2,若直接删除带有缺损数据的元组 2 和元组 3,则造成的信息损失为 $IL=4\times2=8$ 。

表 2 带有缺损数据的数据表实例

Table 2 Instance of data table with missing data

Num	Age	Workclass	Race	Sex	Occupation
1	46	Self-emp-not-inc	White	Male	Craft-repair
2	31	?	White	Male	Adm-clerical
3	38	Private	White	?	Prof-specialty
4	37	Private	White	Male	Sales

定义 5($IL(T)$,数据表 T 信息损失) $T(A_1,\cdots,A_n)$ 是一数据表,则数据表 T 的信息损失为:

$$IL(T)=IL\sum_{e_l\in\varepsilon}(e_l) \tag{2}$$

其中,n 是数据表 T 中元组的数量。可以看出,数据表的信息损失为匿名处理后各个分组信息损失之和。

定义 6(等价类) T' 是一个满足 k-匿名的数据表。如果 T' 中的某几个元组具有相同的准标识符属性的值,那么这几个元组构成一个等价类。

定义 7((α,k)-匿名) 给定一满足 k-匿名的数据表 T'、准标识符 Q 和一个敏感属性值 s,如果任意一个敏感值 s 在每个等价类中的频率都不大于 α,则这个数据表满足 (α,k)-匿名。

该匿名模型主要由 Wong 等提出,以抵御背景知识攻击和同质性攻击。

本节提出了一种面向包含缺损数据的数据匿名方法。其

基本思想是，先以聚类的方式进行最优化分组，之后对各个分组进行泛化。在聚类过程中，通过循环遍历法来寻找使聚类信息损失最少的元组形成聚类，之后在对各个分组泛化时使用基于泛化层次的局部泛化算法，这样每个分组的准标识符便有相同的取值，进而形成等价类，从而可以有效防止隐私信息的泄漏。

2.1 算法描述

算法 1

输入：数据集 S，参数 k，阈值 α

输出：匿名表 S′

开始：

While(|S|≥k)

  随机选取一元组 r，

  将 r 并入聚类 e_l；

  While($|e_l|$≤k)

    根据式(1)查找最优元组 t，

    将 t 并入聚类 e_l；

    S 数据集去除 t；

  End while

    e_l 并入聚类集 ε；

    i=i+1 ；

End while

While(|S|≠0)

  在 S 中随机选取一元组 t；

  S 数据集去除 t；

  根据式(1)将 t 并入最优聚类 e_l

End while

泛化各个聚类等待发布

输出匿名表 S′

结束

2.2 复杂度分析

设原始数据集 S 中的元组数为 $|S|=n$ 。在生成第一个聚类时，需要循环遍历数据集，依次计算该聚类，若新加入某一个元组的信息损失，则需要计算 $(n-1)+(n-2)+\cdots+(n-k)$ 次，时间复杂度为 $O((n-1)+(n-2)+\cdots(n-k))=O(kn-\frac{(1+k)k}{2})=O(n)$ 。在生成第 2 个聚类时，需要计算信息损失 $(n-k-1)+(n-k-2)+\cdots+(n-k-k)$ 次，时间复杂度为 $O((n-k-1)+(n-k-2)+\cdots+(n-k-k))=O(kn-k^2-\frac{(1+k)k}{2})=O(n)$。以此类推，直到聚类的个数为 n/k，因此聚类的平均时间花销为 $\underbrace{O(n)+O(n)+\cdots+O(n)}_{n/k\text{个}}=\frac{n}{k}O(n)=O(\frac{n^2}{k})=O(n^2)$。之后对聚类后形成的分组进行匿名，时间复杂度为 $O(m)$，m 为需要匿名处理的分组数，一般比较小。因此，总的时间复杂度为 $O(n^2)+O(m)=O(n^2)$ 。

3 实验结果及数据分析

3.1 实验数据及参数

3.1.1 软硬件环境

硬件环境为 2.8GHz Xeon E5-2680 v2 CPU，56GB 内存。软件环境为 Windows7 操作系统，Visual Studio 2017 编译环境，C 语言编写。

3.1.2 实验采用的数据来源

实验采用的 Adult 数据集[1)] 属于 UCI 机器学期数据库，该数据集统计了 1996 年一部分美国人口的部分属性，许多数据匿名研究都用到此数据集。Adult 数据集中共有 15 个属性，这里只取其中的 9 个属性进行实验，分别是：Age，Gender，Race，Marital-status，Education，Native-country，Workclass，Occupation。其中除了 Age 属性为有序属性外，其余属性都为无序属性。数据集结构如表 3 所列。

表 3 Adult 数据集描述

Table 3 Description of Adult data set

No.	Attribute	Type	Distinct values	Height
1	Age	Numeric	74	
2	Workclass	Categorical	8	3
3	Education	Categorical	16	3
4	Marital Status	Categorical	7	2
5	Race	Categorical	5	1
6	Natie Country	Categorical	41	1
7	Sex	Categorical	2	1
8	Occupation	Sensitive	14	

实验从运行时间和信息损失两个角度来分析面向缺损数据的(α,k)-匿名算法的性能。

3.1.3 频率参数 α 的设置原则

实验对于敏感属性的频率参数 α 的设置有一定要求：如果属性有较高的敏感性，则 α 应当较低；另外 α 应该大于或等于敏感属性在原始数据集中的频率，否则难以生成聚类。本文对于 α 不做详细研究，统一设置为 $\alpha=0.5$。

3.2 信息损失量的比较

信息损失量采用式度量。图 2 显示了有 7 个准标识符属性、32 561 个元组时，两种算法的信息损失量随着 k 值的变化情况。可以看到，两种算法的信息损失量都会随着 k 值的变大而变大。这是由于随着 k 值的增加，聚类时每个分组所包含的元组会增加，这样对该分组中准标识符属性进行泛化处理时，泛化的层次就会增高，进而产生更高的信息损失。

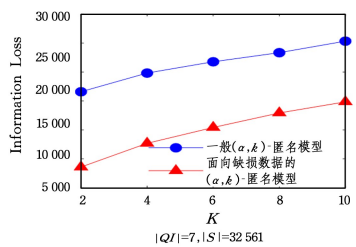

图 2 不同 k 值下信息损失量的比较

Fig. 2 Comparison of information loss under different k values

另一方面，从图 2 能够看出，在 k 值或者准标识符属性个数相同的情况下，面向缺损数据的(α,k)-匿名模型的信息损失量比一般(α,k)-匿名模型的信息损失量小很多。由此可见，相比一般(α,k)-匿名模型，面向缺损数据的(α,k)-匿名模型能够获得更好的隐私信息保护。

图 3 为 k 分别取 2，5，10，准标识符属性个数为 7 且数据集大小变化时，两种算法的信息损失量的比较。可见，在同等条件下，匿名模型所造成的信息损失量均随着数据集的增大而增大，由于数据集的增多会使需要泛化的元组数随之增加，从而造成更多信息损失。另一方面，无论数据集大小如何变

1) http://archive.ics.uci.edu/ml/ datasets/Adult

化，面向缺损数据的(α,k)-匿名模型都比一般(α,k)-匿名模型造成的信息损失小得多。由图4可知：信息损失量会随着准标识符数量$|QI|$的增加而增加，这是由于一旦准标识符属性数量增加，在对元组进行分组时需要考虑的因素就会增多，因此信息损失量会变大。

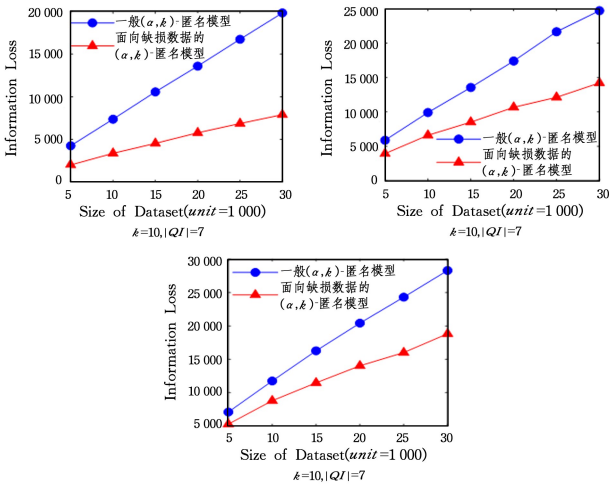

图3 数据集大小变化时信息损失量的比较

Fig. 3 Comparison of information loss as data set size changes

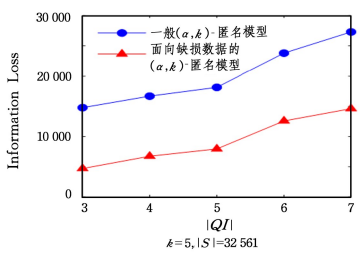

图4 准标识符数量变化时信息损失量比较

Fig. 4 Comparison of information loss as number of quasi-identifier changes

3.3 执行时间的比较

图5显示有7个准标识符属性、32 561个元组时，两种算法随着k值变化的执行时间的对比。可知：两种匿名模型的执行时间都会随着k值的增大而增大，因为这两种算法的聚类过程都是自下向上的。k变大，就意味着聚类次数增多，所以花费的时间就会变多。

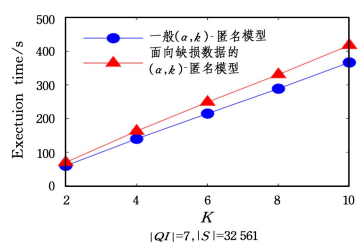

图5 不同k值下运行时间的比较

Fig. 5 Comparison of running time under different k values

图6为k值分别取2,5,10，数据集大小为32 561，准标识符属性个数为7且数据集大小变化时，算法实现两种匿名模型的执行时间的比较。由图6可知，算法的时间花销均会随着数据集的增大而增大，因为随着数据集的增大，聚类次数会增多，所以需要更多时间开销。并且由图6还能看出，算法的时间花销与数据集大小呈二次方关系，这也印证了算法的时间复杂度为$O(n^2)$。由图7可知：准标识符数量越多，算法执行的时间就越长，这是由于在对元组进行分组时需要计算的因素会增多，导致时间花费更多。

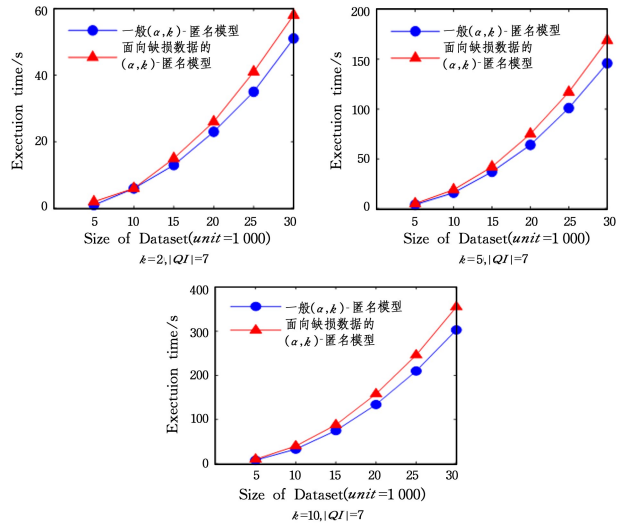

图6 数据集大小变化时运行时间的比较

Fig. 6 Comparison of running time as data set size changes

另一方面，由从图5、图6能够看出，在同等情况下，向缺损数据的(α,k)-匿名模型和一般(α,k)-匿名模型时间花销相差不大，所以向缺损数据的(α,k)-匿名模型以与一般(α,k)-匿名模型近似的时间代价获得更好的隐私保护。

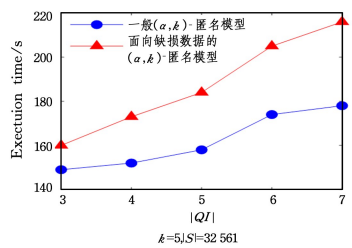

图7 准标识符数量变化时运行时间的比较

Fig. 7 Comparison of running time as number of quasi-identifier changes

结束语 本文面对数据集中的缺失数据，采用了一个基于聚类的(α,k)-匿名模型。该模型可以对带有缺损数据的数据集进行匿名处理，确保匿名后的数据集满足(α,k)-匿名模型，保证隐私安全。实验结果证明，面向缺损数据的(α,k)-匿名模型以与一般(α,k)-匿名模型近似的时间代价获得了更好的隐私保护。

本文算法和模型中的聚类分组中包含的元组数基本是固定的，下一步将考虑在聚类过程中采用变化元组数目。

参 考 文 献

[1] SWEENEY L. k-Anonymity: A model for protecting privacy [J]. Int'l Journal on Uncertain, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[2] SAMARATI P. Protecting respondents' identities in microdata release[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010-1027.

[3] LI T C, LI N H. Towards optimal k-anonymization[J]. Data and Knowledge Engineering, 2008, 65(1): 22 -39.

[4] 韩建民，于娟，虞慧群，等. 面向敏感值的个性化隐私保护[J]. 电子学报，2010，38(7)：1723-1728.

[5] MACHANAVAJJHALA A, GEHRKE J, KIFER D. L-diversi-

ty: privacy beyondk-anonymity[C]// Proceedings of the 22nd International Conference on Data Engineering. Atlanta, GA, USA: IEEE Press, 2006: 24-36.

[6] TRUTA T M, VINAY B. Privacy protection: p-sensitive kanonymityproperty[C]// Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW). Washington, DC, USA: IEEE Computer Society, 2006: 94.

[7] WONG C R, LI J, FU A, et al. (α,k)-anonymity: an enhancedk-anonymity model for privacy preserving data publishing[C]// Proceedings of the 12th ACM SIGKDD Conference. Philadelphia, PA: ACM Press, 2006: 754-759.

[8] LI N H, LI T C, VENKATASUBRAMANIAN S. t-Closeness: privacy beyond k-anonymity and l-diversity[C]// Proceedingsof the 23rd International Conference on Data Engineering(ICDE). Istanbul, Turkey: IEEE Press, 2007: 106-115.

[9] XIAO X K, TAO Y F. Personalized privacy preservation[C]// Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. Chicago, USA: ACM Press, 2006: 229-240.

[10] YANG X C, LIU X Y, WANG B, et al. K-Anonymization approaches for supporting multiple constraints[J]. Journal of Software, 2006, 17(5): 1222-1231.

[11] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Incognito: Efficient full-domain K-anonymity[C]// Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). ACM Press, 2005: 49-60.

[12] XU J, WANG W, PEI J, et al. Utility-Based anonymization using local recoding[C]// Proc. of the 12th ACMSIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD). ACM Press, 2006: 785-790.

[13] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Mondrian multidimensional K-anonymity[C]// Proc. of the 22nd Int'l Conf. on Data Engineering (ICDE). IEEE, 2006: 25.

[14] XIAO XK, TAO Y. Anatomy: Simple and effective privacy preservation[C]// Proc. of the 32nd Int'l Conf. on Very Large Data Bases(VLDB). VLDB Endowment, 2006: 139-150.

[15] TAO YF, CHEN H K, XIAO X, et al. ANGEL: Enhancing the utility of generalization for privacy preserving publication[J]. IEEE Trans. on Knowledge and Data Engineering (TKDE), 2009, 21(7): 1073-1087.

[16] WONG R C W, LI J Y, FU A W C, et al. (α,k)-Anonymity: An enhanced k-anonymity model for privacy preserving data publishing[C]// Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge discovery and Data Mining (SIGKDD). ACM Press, 2006: 754-759.

[17] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. l-Diversity: Privacy beyond k-anonymity[J]. ACM Trans. on Knowledge Discovery Data (TKDD), 2007, 1: 3.

[18] 任向民. 基于K-匿名的隐私保护方法研究[D]. 哈尔滨：哈尔滨工程大学，2012.

ZHANG Wang-ce, born in 1994, master, is a student member of CCF. His main interests include network security and machine learning.

FAN Jing, postgraduate, professor, is a member of China Computer Federation. Her main research interests include network security, intelligent sensor net and intelligent control.

(上接第 390 页)

[3] CANARD S, JAMBERT A. On extended sanitizable signature schemes[C]// Cryptographers' Track at the RSA Conference. Berlin: Springer, 2010: 179-194.

[4] KLONOWSKI M, LAUKS A. Extended sanitizable signatures [C]// Proc of Information Security and Cryptology-ICISC. Berlin: Springer, 2006: 343-355.

[5] BRZUSKA C, FISCHLIN M, LEHMANN A, et al. Unlinkability of sanitizable signatures[C]// Proc. of Public-Key Cryptography-PKC. Berlin: Springer, 2010: 444-461.

[6] LAI W F, ZHANG T, CHOW S M, et al. Efficient Sanitizable Signature Without Random Oracles[C]// Proc. of ESORICS. Springer, 2016: 363-380.

[7] FLEISCHHCAKER N, KRUPP J, MALAVOLTA G, et al. Efficient unlinkablesanitizable signatures from signatures with re-randomizable keys[C]// Proc. of Public-Key Cryptography-PKC. Berlin: Springe, 2016: 301-330.

[8] POINTCHEVAL D, SANDERS O. Short randomizable signatures[C]// Cryptographers' Track at the RSA Conference. Springer, Cham, 2016: 111-126.

[9] LV J Q, WANG X M. Verifiable ring signature[C]// Proc. of 9th International Conference on Distributed Multimedia System. Miami, USA, 2003: 663-665.

[10] 王化群，郭显久，于红，等. 几种可转换环签名方案的安全性分析和改进[J]. 电子与信息学报，2009，35(15)：135-137.

[11] 李晓琳，梁向前，刘奎，等. 可验证环签名方案的分析与改进[J]. 计算机应用，2012，32(12)：3466-3469.

[12] BONEH D, LYNN B, SHACHAM H. Short signatures from weil pairing[C]// Proc of Advances in Cryptology-ASIACRYPTY. Berlin: Springer, 2001: 512-532.

[13] BRZUSKA C, FISCHLIN M, LEHMANN A, et al. Sanitizable Signatures: How to partially delegate control for authenticated data. [C]// Proc. of Special Interest Group on Biometrics and Electronic Signatures. Bonn: GI, 2009: 117-128.

[14] LV X, XU F, PING P, et al. Schnorr ring signature scheme with designated verifiability[C]// 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES). IEEE, 2015: 163-166.

ZHANG Jun-he, born in 1991, postgraduate. His main research interests include mimic defense, and digital signature.