数据自治开放模式下的隐私保护

王智慧[1,2]，周旭晨[1,2]，朱云[1,2]
1. 复旦大学计算机科学技术学院，上海 201203；2. 上海市数据科学重点实验室，上海 201203

摘要
数据开放对于提升数据资源的应用价值具有十分重要的意义。但是出于隐私保护的考虑，数据开放应该是有监管的开放，即采取数据自治开放模式。针对数据自治开放可能给隐私保护带来的挑战，提出了面向数据盒的隐私保护系统框架。该系统框架针对数据使用者的数据使用声明进行隐私泄露风险评估，并在评估结果的基础上决定是否授权许可相应的数据使用请求，以支持数据自治开放的实现。

关键词
数据管理；数据自治开放；隐私保护

中图分类号：TP311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018017

Privacy preservation in self-governing openness of data

WANG Zhihui[1,2], ZHOU Xuchen[1,2], ZHU Yun[1,2]
1. School of Computer Science, Fudan University, Shanghai 201203, China
2. Shanghai Key Laboratory of Data Science, Shanghai 201203, China

Abstract
The openness of data is very important for improving the application value of data resource. However, due to the consideration of privacy preservation, data should be open under supervision. That is, data should be in the mode of self-governing openness. For challenges that self-governing openness of data may bring, a possible system framework for the privacy preservation of a data box was presented. The system framework carries out the privacy disclosure risk evaluation for the data usage declaration of a data user, and then determines whether the corresponding data usage request can be authorized or not on the basis of evaluation results, and thus supports the implementation of self-governing openness of data.

Key words
data management, self-governing openness of data, privacy preservation

1 引言

随着大数据时代的来临，数据资源的战略性和商业价值越来越被人们认同和重视，数据资源的开放共享需求也日益紧迫。然而在现实中，数据资源的开放共享却变得越来越困难。造成这种状况的主要原因之一是数据开放共享时的隐私保护问题[1]。

以医疗数据开放为例，医疗数据开放显然能够为医疗数据开发利用提供便利，为临床诊断、药物研发等提供数据支持，推动医疗数据价值的实现。但是，医疗数据常包含较多的隐私信息，例如，患者个人信息、既往病史、就诊记录等。因此，出于对患者的隐私保护的考虑，医疗数据的开放将不可避免地受到制约。

从数据开放模式下的隐私保护需求考虑，一方面要求实现对个体隐私的保护，即保护数据所描述的个体对象的隐私；另一方面也要求对数据自身的稀缺性加以保护，即保护数据开放者提供的原始数据的整体隐私不被非法攫取。因此，从隐私保护的角度来说，数据开放要求必然是有监管的开放，即数据自治开放。

本文考虑在数据自治开放模式下，以数据盒为基本数据单元向数据使用者开放。数据盒是数据自治开放的载体，它封装了被开放的数据以及相应的数据管理系统。数据使用者先向数据开放者申请使用数据盒，在获得数据开放者的授权之后，再通过数据盒提供的数据访问接口使用数据，以此实现数据的开放。在数据使用的过程中，数据盒内的管理系统通过对数据使用者的数据访问进行管控，实现数据开放过程的自治。一个数据盒内封装的数据可能含有隐私信息。为了防止隐私泄露，数据使用者在申请使用数据盒时，需要提供相应的数据使用说明，即说明需要使用哪些数据、以何种方式使用这些数据以及数据使用的预期结果。因此，面向数据盒的隐私保护需要通过分析数据使用者的数据使用说明，判断相应数据使用是否可能导致隐私泄露以及涉及隐私泄露的数据范围和隐私泄露的严重程度。

针对数据盒的隐私保护需求，本文总结了在数据自治开放模式下面临的隐私保护挑战，基于隐私泄露风险评估，设计了面向数据盒的隐私保护系统框架。通过将风险评估与使用控制结合，允许数据开放者在其可接受的隐私泄露风险范围内开放数据，进而为数据自治开放提供有力保障。

2 相关工作

隐私就是任何与特定个人或团体相关、但不愿被不加选择地暴露的信息①[2]。目前许多隐私保护研究基于这样的前提假设：数据开放者持有的电子化数据一般是完整的、未经任何处理的数据，其中涉及隐私的敏感属性事先已知。一种简单的手段是隐藏姓名、身份证号等唯一标识个体身份的显式标识符，但这种简单匿名化的效果是相对脆弱的，难以抵御常见的背景知识攻击和链接攻击等行为[3-5]。背景知识攻击就是利用背景知识直接推知隐私信息的攻击手段。背景知识包括攻击者可能拥有的任何信息，前提是这些信息有助于其进行推理，并获知开放数据中某些记录对应的敏感属性值。而链接攻击则是指攻击者在其背景知识基础上和外部数据集进行链接推理，进而导致隐私泄露的攻击方式。围绕这一问题，目前

① www.codata.org

已经有k-anonymity[3,4]、l-diversity[5]、t-closeness[6]、m-invariance[7]、differential privacy[8,9]等多种隐私数据保护模型先后被提出。但是，目前的研究很多都是在数据开放前通过对数据进行添加噪声等相应处理来实现隐私保护的，这样一方面会降低数据在开放过程中的可用性，另一方面也没有体现数据开放过程中对隐私保护的自主可控性。

在大数据环境下，攻击者通过数据挖掘分析技术会拥有更强大的背景知识，而且可以结合多源数据以发动链接攻击。针对这一情况，一些隐私保护研究立足于与数据挖掘相关的隐私保护研究。特别地，一些研究针对关联规则挖掘[10-13]、数据分类[14-18]以及数据聚类[19-21]分别提出了相应的隐私保护挖掘技术。但是，这些研究大都针对特定的数据挖掘任务，缺乏数据自治开放实际所需的普遍适用性。

从上述关于隐私保护的现有技术的分析可以看到，现行方法虽然在一定程度上起到了隐私保护的作用，但是仍存在不足，不能满足数据自治开放模式下隐私保护的本质需求。

3 隐私保护挑战

在数据自治开放模式下，如何实现有效的隐私保护是有待研究的重要问题之一，对数据开放及其流通等有着十分重要的影响。在实现数据自治开放的过程中，面临的隐私保护挑战主要来自以下几个方面。

（1）隐私保护的自主可控性

隐私保护的自主可控性即数据开放者能够根据数据自身的特性或者数据在开放过程中的使用需求，自主地决定并控制哪些数据在数据开放的过程中需要进行隐私保护，并且对这些数据的隐私保护需要达到何种程度。如果不能保障隐私保护的自主可控性，也就无法在数据开放过程中实现对数据的有效监管，从而无法满足数据自治开放的基本要求。

（2）如何平衡数据的可用性与隐私保护之间的关系

数据的可用性反映了在数据自治开放的过程中，数据自身的实际意义得到表达的程度。数据的可用性与具体的应用场景有紧密的关联，其在表现形式上既可以是数据在统计意义上的数据分布信息，也可以是特定数据记录的准确属性值。在数据开放的过程中，如果片面强调数据的可用性，那么隐私泄露的风险必然会增大；如果过分强调隐私保护，那么数据的可用性则会受到影响。

（3）如何兼顾对个体隐私和数据自身稀缺性的有效保护

对个体隐私的有效保护即保护数据所描述的个体对象的隐私不被泄露。这表现为数据使用者不能在未经数据开放者同意的前提下，通过对数据的访问获取特定个体对象的隐私信息。需要注意的是，数据使用者有时虽不能通过单次访问获取个体的隐私信息，但是通过对数据的多次访问，然后利用访问结果之间的相关性可能获取个体的隐私信息。对数据自身的稀缺性进行有效保护是指保护数据开放者提供的原始数据的整体隐私不被非法攫取。这在数据开放的过程中，通常表现为需要防范数据使用者通过“数据拼图”造成对数据整体隐私的侵犯。

数据拼图是指数据使用者能够通过整合数据访问过程中多次获取的数据片段，利用数据片段之间的关联性，非法拼接还原出整个数据的全貌或者其中大量的涉及隐私的敏感数据。数据拼图可以由单个使用者通过拼接在多次数据访问中获取的

数据片段来完成，也可能由多个使用者共同合作，通过共谋来实现对数据片段的拼接。在数据自治开放模式下，一个或多个恶意的攻击者可以通过数据拼图这种攻击手段，未经数据开放者授权就获取被保护的数据对象，从而构成对数据整体隐私的侵犯。

数据拼图会给数据开放带来实质危害，因为攻击者可以借助这种手段实现对原始数据的攫取，这样就意味着数据资源的稀缺性丧失，从而丧失其原有的价值，导致数据拥有者不愿意将数据开放给别人使用。此外，数据拼图的攻击使得数据整体的所有权属也难以得到保护。因为数据使用者可以将通过数据拼图获得的数据再次传播给其他未被授权的数据使用者，造成对原数据权属的二次侵犯。

针对上述数据自治开放模式下的隐私保护挑战，在隐私泄露风险评估的基础上，本文提出了一种面向数据盒的隐私保护系统框架来应对这些挑战。

4 面向数据盒的隐私保护系统框架

数据盒作为实现数据自治开放的载体，将数据封装在盒内，其本质是在获得数据开放者许可的前提下，将数据使用权开放给使用者，而不是直接将数据移交给使用者。由于数据盒内封装的数据可能涉及隐私信息，因此在数据使用者使用数据盒内的数据之前，评估其相应的数据使用请求，判断是否会导致隐私泄露，对于数据自治开放中的隐私保护至关重要。本节在隐私泄露风险评估的基础上，给出了一个具有层次化结构的面向数据盒的隐私保护系统框架。该系统框架如**图**1所示。

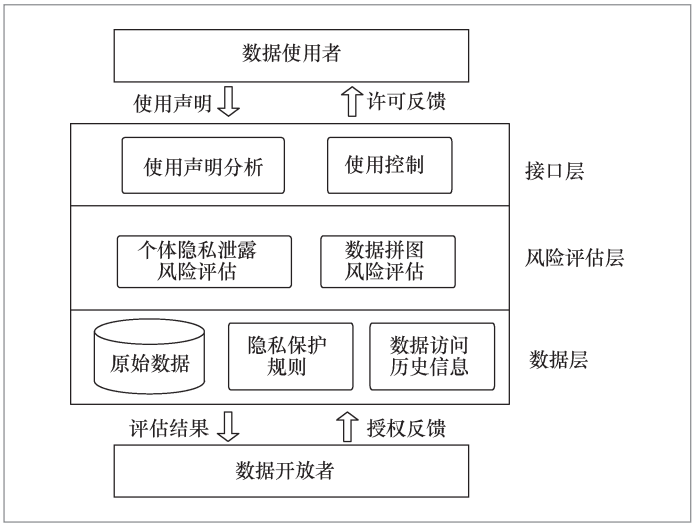


图1 面向数据盒的隐私保护系统框架

在系统框架中，最下层的数据层存储数据开放者提供的原始数据、隐私保护规则以及当前所有数据访问的历史信息记录集合，为评估隐私泄露的风险提供数据基础。隐私保护规则包括个体隐私保护规则和数据拼图防范规则，可以由数据开放者预定义，以便依据风险评估的结果确定是否违背数据开放者的隐私保护需求。

数据使用者可以采用系统预定义的数据使用声明规范描述语言或接口，表达自己对数据开放者提供数据的使用请求，并将其提交给系统框架的接口层。接口层包括使用声明分析模块和使用控制模块，其中使用声明分析模块首先分析数据使用者的数据使用声明，在此基础上抽取数据使用者可能的数据使用行为、数据使用范围以及数据使用的预期结果，并传递给风险评估层做相应的风险评估。在获取风险评估层的反馈结果后，如果当前数据使用请求带来的风险超出了数据开放者所能承受的限度，使用控制模块将拒绝数据使用者的使用请求。

系统框架中的风险评估层将根据接口层传递下来的数据使用者可能的数据使用

行为及数据使用范围，对数据使用者的数据使用请求进行风险评估。在进行风险评估时，将针对当前数据使用请求涉及的数据使用行为及数据使用范围，结合数据层中存储的数据访问历史信息，对当前数据使用请求可能导致的个体隐私泄露风险和数据拼图风险分别进行评估，给出相应的个体隐私泄露风险系数或等级以及数据拼图的风险系数或等级。同时，可以根据数据开放者预设的个体隐私保护规则和数据拼图防范规则，自动检查相应的风险系数或等级是否超出数据开放者的承受范围，也可以将风险评估的结果直接展现给数据开放者，由数据开放者自行决定是否授权许可当前的数据使用请求，并可进一步通过接口层的使用控制将授权许可反馈给数据使用者。

在风险评估的过程中，具体的隐私信息认定、隐私泄露的风险系数或等级的设定等可以根据应用场景由数据开放者自主确定。**表1**记录了某公司的员工收入情况，每条数据记录均与某个特定员工相对应。假设**表1**中的收入属性值为涉及隐私的敏感数据，其中单条数据记录中的收入属性值为个体隐私，数据集内所有记录的收入属性值集合构成了整体隐私。个体隐私泄露的风险系数设定为数据使用者依据数据盒反馈的数据访问结果，可推测出单条数据记录中的收入属性值的可能性。数据拼图风险系数设定为数据使用者根据数据盒反馈的数据访问结果，能获知的收入属性值在数据集内所有记录的收入属性值集合中所占比例。

表1 示例数据

编号	年龄/岁	收入/元
1	20	3 500
2	22	3 800
3	45	5 000
4	57	5 500
5	28	4 000
6	60	7 000
7	37	4 500
8	48	6 400
9	62	4 300
10	32	6 500

考虑数据使用者曾提出数据使用请求Q_1：“统计年龄21~25岁之间的员工数”。该请求获得了授权，数据使用者已获知相应的结果（即数据使用者已获知年龄为21~25岁的员工只有1人）；同时Q_1作为数据访问历史信息被记录下来。现在数据使用者进一步提出新的数据使用请求Q_2：“计算年龄为21~25岁的员工平均年龄及平均收入”。如果Q_2获得授权，那么数据使用者将可以推测出“年龄22岁的员工收入为3 800元”。因此，依据前述风险系数的设定，对Q_2进行评估将会得出其所对应的个体隐私泄露风险系数为100%，数据拼图风险系数为10%。按照风险评估的结果，如果数据开放者认为风险系数过高，则可以拒绝对Q_2进行授权。

面向数据盒的隐私保护系统框架可以较好地适用于数据自治开放的应用场景，即在数据开放者自主可控的情况下，将数据甚至部分隐私数据作为开放的资源。在基于隐私泄露风险评估的系统框架中，数据开放者可以依据系统提供的风险评估结果给出授权与否的直接反馈，或者通过设定合适的个体隐私保护规则和数据拼图防范规则，决定是否许可当前的数据使用请求，可以在数据使用过程中兼顾个体隐私保护与整体隐私保护，体现了对隐私保护的自主可控。同时在实际使用中，数据开放者针对不同的数据使用者，可以依据隐私泄露风险的评估结果，给予不同的授权反馈，从而也可以实现对数据可用性与隐私

保护之间的动态平衡。此外，该系统框架允许数据使用者在数据开放者能够接受的隐私泄露风险范围内，直接使用部分原始数据，因此与传统的通过引入噪声数据来实现隐私保护的方式相比，该方式在一定程度上可以提高数据可用性。

该系统框架还存在着一些技术细节及局限性需要进一步研究。一是数据使用者在数据使用前提出的数据使用请求与其在数据使用时的真实行为之间的一致性问题，即所谓的言行一致问题。数据使用者可以先提出低风险的数据使用请求，通过风险评估获得使用授权，但在实际使用数据的过程中却不遵照事先的数据使用请求而非法获取隐私信息。因此，还需要在数据的实际使用过程中对数据使用者的数据访问行为进行必要的管控，以确保与其事先的数据使用声明一致。二是由于数据使用者可能会通过多次精心设计的不同数据访问，再利用数据之间内在的相关性，逐步推导出数据集的全貌或者其中的个体隐私信息，因此需要维护大量的数据访问历史信息，防止可能的隐私泄露。但是，这样做的后果是，随着数据访问历史信息的逐渐增加，风险评估系统的效率会逐步降低。因此，有必要研究通过索引技术等提高系统的实现效率。三是在某些场景下，数据可能会发生更新的情况或者以数据流的形式出现，鉴于数据的变化可能会产生隐私泄露以及可能会影响到先前的风险评估结果，因此，需要进一步研究由于数据更新或数据流的方式带来的隐私泄露以及相应的风险评估问题。

5 结束语

数据自治开放可以为数据的开放共享提供便利，但是在实际过程中可能产生个体隐私信息泄露以及数据的整体隐私受到危害（即数据自身可能被非法攫取）等问题。这些问题已经严重阻碍了数据拥有者开放其自身数据。本文分析了数据自治开放模式下的隐私保护挑战，并基于隐私泄露风险评估，针对这些挑战提出了面向数据盒的隐私保护系统框架。通过将风险评估与使用控制结合，允许在数据开放者可接受的隐私泄露风险范围内使用数据，给数据自治开放提供有力保障，推动数据资源的开放利用。

参考文献:

[1] BOYD D, CRAWFORD K. Critical questions for big data[J]. Information, Communication & Society, 2012, 15(5): 662-679.

[2] 方滨兴, 贾焰, 李爱平, 等. 大数据隐私保护技术综述[J]. 大数据, 2016, 2(1): 1-18.
FANG B X, JIA Y, LI A P, et al. Privacy preservation in big data: a survey[J]. Big Data Research, 2016, 2(1): 1-18.

[3] SWEENEY L. k-Anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[4] SWEENEY L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571-588.

[5] MACHANAVAJJHALA A, GEHRKE J, KIFER D. l-Diversity: privacy beyond k-anonymity[C]//The 22nd International Conference on Data Engineering, April 3-8, 2006, Atlanta, USA. New York: ACM Press, 2007, 1(1): 3.

[6] LI N H, LI T C, VENKATASUBRAMANIAN S. t-Closeness: privacy beyond k-anonymity and l-diversity[C]// The 23rd International

Conference on Data Engineering, April 17-20, 2007, Washington, USA. Piscataway: IEEE Press, 2007: 106-115.

[7] XIAO X K, TAO Y F. M-invariance: towards privacy preserving re-publication of dynamic datasets[C]// The ACM SIGMOD International Conference on Management of Data, June 11-14, 2007, Beijing, China. New York: ACM Press, 2007: 689-700.

[8] DWORK C. Differential privacy[C]// The 33rd International Colloquium on Automata, Languages, and Programming, July 10-14, 2006, Venice, Italy. Heidelberg: Springer, 2006: 1-12.

[9] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4): 211-407.

[10] RIZVI S J, HARITSA J R. Maintaining data privacy in association rule mining[C]// The 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China. [S.l.]: VLDB Endowment Inc., 2002, 34(6): 682-693.

[11] EVFIMIEVSKI A V, SRIKANT R, AGRAWAL R, et al. Privacy preserving mining of association rules[C]// The 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, July 23-25, 2002, Edmonton, Canada. New York: ACM Press, 2002: 217-228.

[12] KANTARCIOGLU M, NIX R, VAIDYA J. An efficient approximate protocol for privacy-preserving association rule mining[C]// The 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, April 27-30, 2009, Bangkok, Thailand. Heidelberg: Springer, 2009: 515-524.

[13] BONOMI L. Mining frequent patterns with differential privacy[C]//VLDB Endowment, August 26-30, 2013, Riva del Garda, Trento. [S.l.]: VLDB Endowment Inc., 2013, 6(12): 1422-1427.

[14] AGRAWA R L, SRIKANT R. Privacy-preserving data mining[C]// The ACM SIGMOD Conference on Management of Data, May 16-18, 2000, Dallas, USA. New York: ACM Press, 2000: 439-450.

[15] YU H, JIANG X Q, VAIDYA J. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data[C]//ACM Symposium on Applied Computing, Data Mining Track, April 23-27, 2006, Dijon, France. New York: ACM Press, 2006, 39(9): 603-610.

[16] VAIDYA J, CLIFTON C, KANTARCIOGLU M, et al. Privacy-preserving decision trees over vertically partitioned data[J]. ACM Transactions on Knowledge Discovery from Data, 2008, 2(3): 1-27.

[17] MANGASARIAN O L, WILD EW, FUNG G. Privacy-preserving classification of vertically partitioned data via random kernels[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 2(3): 1-16.

[18] YANG B,SATO I, NAKAGAWA H. Bayesian differential privacy on correlated data[C]//SIGMOD Conference, May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 747-762.

[19] JAGANNATHAN G, WRIGHT R N. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data[C]// The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 21-24, 2005, Chicago, USA. New York: ACM Press, 2005: 593-599.

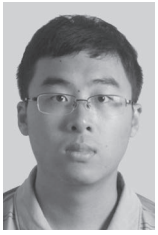
[20] CUI Y J, WONG W K, CHEUNG D W. Privacy-preserving clustering with high accuracy and low time complexity[C]//The 14th International Conference on Database Systems for Advanced Applications, April 21-23, 2009, Brisbane, Australia.

Heidelberg: Springer, 2009: 456-470.

[21] DISHABI M R E, AZGOMI M A. Differential privacy preserving clustering in distributed datasets using Haar wavelet transform[J]. Intelligent Data Analysis, 2015, 19(6): 1323-1353.

作者简介

王智慧(1975-),男,博士,复旦大学计算机科学技术学院讲师,主要研究方向为数据管理、数据挖掘、数据安全与隐私保护。



周旭晨(1993-),男,复旦大学计算机科学技术学院硕士生,主要研究方向为隐私保护、差分隐私。

朱云(1986-),女,复旦大学计算机科学技术学院硕士生,主要研究方向为数据管理、隐私保护。

收稿日期: 2018-02-05

基金项目: 上海市科技发展基金资助项目(No. 16JC1400801)

Foundation Item: Shanghai Science and Technology Development Fund (No.16JC1400801)