大数据治理的数据模式与安全

马朝辉[1]，聂瑞华[1]，谭昊翔[1]，林嘉洺[1]，王欣明[1]，唐华[2]，杨晋吉[1]，赵淦森[1]
1. 华南师范大学计算机学院，广东 广州 510630
2. 华南师范大学软件学院，广东 佛山 528225

摘要
大数据治理的主要目的是使数据的利用价值和利用效率最大化，治理后的数据在利用过程中也不可避免会涉及敏感数据或者隐私数据。从大数据治理出发，基于实际应用案例，讨论大数据治理过程中如何利用数据模式的重组实现数据价值的提升和数据处理效率的提升。同时，也提出了数据安全访问策略的自动生成，保障数据在重组后得到相应的安全防护。

关键词
大数据治理；数据融合；访问控制

中图分类号：TP391　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016033

Research on data schema and security in data governance

MA Chaohui, NIE Ruihua, TAN Haoxiang, LIN Jiaming, WANG Xinming, TANG Hua, YANG Jinji, ZHAO Gansen
1. School of Computer, South China Normal University, Guangzhou 510630, China
2. School of Software, South China Normal University, Foshan 528225, China

Abstract
One of the key objectives of big data governance is to maximize the value and efficiency of data usage. It is less than possible to privacy while processing data that has been subjected to data governance. With case study, the way to improve data value and data processing efficiency by re-construct data schemas was investigated. A mechanism for calculating new access control policies was also presented. The generated access control policies could provide appropriate security protection over reconstructed data.

Key words
data governance, data fusion, access control

1 背景

据IBM公司的分析，人类文明有90%的数据是在过去两年内产生的，到2020年，全世界所产生的数据规模将达到今天的44倍[1]。而我国截至2015年12月，已经拥有6.88亿的互联网用户，13.06亿的手机用户①，每天可产生海量的数据。大数据无论在大型企业，还是政府部门都发挥着相当的作用。在2015年7月1日国务院办公厅印发的《关于运用大数据加强对市场主体服务和监管的若干意见》中提到，充分运用大数据的先进理念、技术和资源是提升国家竞争力的战略选择，是提高政府服务和监管能力的必然要求，有利于政府充分获取和运用信息，更加准确地了解市场主体需求，提高服务和监管的针对性、有效性。此外，大数据为医疗、能源、智慧城市、生物医学、基因组学、交通运输等领域提供了不同的应用视角。如何通过大数据治理来解决上述城市化问题以及更广泛的问题是数字时代的趋势。

数据治理当前已经成为IT业界一门新兴的学科，被广泛研究，但是数据治理这个概念则广泛应用在企业界。数据治理是指“从使用零散数据变为使用同一主数据、从具有很少或没有组织和流程治理到企业范围内的数据治理、从尝试处理主数据混乱状况到主数据井井有条的一个过程，并最终使企业能将数据作为企业的核心资产来管理”[2]。大数据治理这个概念形成于大数据时代，但是对于大数据治理的定义众说纷纭。美国学者桑尼尔·索雷斯[3]将大数据治理定义为：大数据治理是广义信息治理计划的一部分，即制定与大数据有关的数据优化、隐私保护与数据变现的政策。梁芷铭[4]综合不同观点认为：大数据治理是不同的人群或组织机构在大数据时代为了应对大数据带来的种种不安、困难与威胁，运用不同的技术工具对大数据进行管理、整合、分析并挖掘其价值的行为。

大数据治理对国家治理同样重要。大数据技术为提升国家的科学决策、社会监管、公共服务以及应急管理能力都提供了良好的契机，现在国家治理的多元主体已经和信息化、数字化分不开了，但是大量数据藏身于互联网和各种数字媒介，难分真假、难以辨清，国家治理主体容易迷失在其中，因此大数据治理会是国家治理的重要方面。对于国家治理过程中的大数据进行治理，其主要的作用主要体现在以下几个方面。

第一，大数据能有效提升科学决策水平[5]。因为大数据收集了整个国家各个领域方面的信息资源，对这些数据资源进行整合之后相当于一个庞大的信息资源库，面对数据洪流，客观、理性地进行数据分析，强化大数据治理，能更好地帮助国家治理决策科学化，为国家治理提供重要的数据支持和决策依据。

第二，大数据通过增强对现象之间的关联与研究，可以有效减少社会危机发生的不确定性，增强风险预警能力，降低社会危机带来的危害。大数据和社会公共管理的有效对接能够高效实现跨部门、跨领域的管理信息共享，能有效提升公共危机事件的源头治理、事前预警、动态监控和应急处置能力。

第三，数据共享为政府各职能部门的沟通提供了便利，模糊政府各部门之间、政府与公众之间的边界，使得信息孤岛现象大幅度减少。

2 相关工作

2.1 数据融合

数据融合能够成为计算机领域内的研

① http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/

究热点，与实际需求和数据融合技术的巨大潜能息息相关。数据融合最初是由于军事作战需求而提出的，是为了使多种作战设备上多传感器的数据信息能够协调、整合与集成而形成的一种数据横向综合信息处理技术。因而，国内早期研究数据融合的研究者[6]，从技术的观点把数据融合理解为一种技术思路，视为多源信息协调处理技术的总称。随着计算机科学技术的迅猛发展，数据融合概念已经不再局限于多传感器数据融合技术领域，概念的覆盖领域进一步扩充。在计算机领域，随着硬件设备性能和软件服务能力的不断提升，面对多源数据系统的数据融合，数据集成的技术手段不再缺乏。而在如何构建多源数据的集成模型，提供给用户统一的数据视图的问题上，国外数据研究者Lenzerini M提出了自己的一些思考与想法[7]，他针对各种数据源和全局数据模式之间如何建立关联关系，提出了global-as-view和local-as-view两种基础方法论，并对如何在数据整合中处理查询、如何处理数据源不一致性问题等提出了相关的观点和方法。

近年来，云计算技术新军突起，成为计算机领域分布式计算的一面旗帜。而伴随着移动互联网时代的到来，信息数据资源激增，也是所谓的“大数据”时代的到来，面对越来越多的信息源和数据源，各种数据使用实体对数据融合的实际需求更加迫切。大数据时代，数据的产生、收集和处理规模空前，在数据集成处理上，Dong X L等从多个维度提出了大数据集成与传统数据集成的区别[8]，这些维度包括了数据源的数量、数据源的动态性、数据源异构和数据源的质量差异。面对大数据，数据融合要充分考虑数据源对象的各种特性，充分考虑大数据融合过程中可能出现的数据问题。为了降低处理大量复杂数据源整合过程中的任务复杂度，Caruccio L等提出了一种基于可视化语言的方法和工具[9]。基于概念层次上的数据融合，该可视化语言能够提供对数据源概念数据模型构建的操作接口或操作方式，这种工具能够生成多个数据源之间的关联模式，自动生成元数据并且提供一种机制，保证阶段性地从各个数据源中加载更新的数据。

《中国大数据技术与产业发展白皮书（2014年）》中对大数据发展趋势的预测总结为“融合、跨界、基础、突破”，可以看出在未来的一个时间阶段内，大数据领域数据融合成为最为显著的发展趋势。数据融合因为实际需求而提出，技术成果要服务于实际应用。互联网将各种异构网络、各种不同的信息系统连在一起，变成一个更庞大的信息资源网络。面对Web数据形式多样、表达自由等特点带来的数据集成信息冗余、准确度差、数据离散等问题，张永新博士对Web数据融合进行了深入探究[10]。数据融合是数据分析挖掘的重要前提，提高集成数据的质量十分关键，张永新针对海量Web信息的数据冲突、多源数据关联、数据融合的可回溯机制等保证数据集成质量的多个方面进行了研究和探讨。此外，为了解决大数据给数据融合带来的新挑战，北京邮电大学穆化鑫尝试使用分布式计算的能力来应对[11]，他提出基于Storm实时计算引擎对物联网的异构数据进行融合处理，其工作主要是构建一种系统架构，将现有的数据融合相关算法与Storm分布式实时计算引擎结合起来，形成一个算法与数据分离、高解耦且可扩展的实时分布式数据融合系统。大数据带来了数据融合的挑战，也催生了解决问题的技术，特别地，数据融合对于大数据与社会治理也提供了强有力的技术支撑。针对电子政务工程建设中政府信息资源利用效率低下的问题，电子科技大学石西庆提出了一种基于“任务”的城市级基础数据融

合服务模型，实现政务基础数据的快速融合服务发布，确保基础数据的时效性和服务能力，进而构建一种电子政务信息共享服务平台[12]。类似地，北京大学化柏林教授对大数据环境下多源信息数据融合的应用进行了深入研究[13]，从国家、社会和企业的不同层次、不同角度的应用研究（如国家政府“单独两孩”政策、城市综合治理和产业优化调整、企业的发展决策等），表征了数据驱动决策的思路贯穿社会多个领域，更体现出数据融合在社会治理中的重要作用。

2.2 数据融合安全

数据融合作为大数据治理的一个重要环节，数据机密性及隐私保护是其主要面临的安全问题。数据融合的生命周期包括收集、融合、检索、处理分析，每个阶段都存在破坏数据的风险。在数据收集阶段，数据融合汇聚了来自多个机构或组织的数据源，每个数据源由不同的安全策略管控，数据很有可能没有按照其安全策略进行收集或者不同机构的安全策略存在冲突[14]。在数据融合阶段，数据被融合集成到一个公共平台，例如data.gov等数据开放平台，孟小峰[15]等指出数据被外包或开放到一个不可信的公共平台，没有索引加密或访问控制等安全保护措施，很可能会引起数据的泄露。在数据检索阶段，融合数据提供检索服务来共享数据，这是最容易发生数据泄露的阶段。因为每个用户都可以从搜索引擎获取数据，如果没有全局安全策略[18]来管控数据，将面临着数据泄露的风险。为了解决这个问题，常见的方法是采用加密文本检索技术[19,20]。在数据处理分析的阶段，同样存在数据泄露的问题，主要原因有：多数据源之间的交叉分析挖掘，很可能发现机密信息或者暴露隐私；数据的处理往往依托大数据平台进行分析，如Hadoop和Spark，平台计算资源是共享的，因而也存在暴露数据的可能。

访问控制是数据融合安全防护的主要机制之一。Carlo等[19]认为多机构合作并共享数据的环境需要提供一种灵活的访问控制来使用资源，因此提出了管理融合数据的访问框架，该框架将系统划分成本地环境以及融合环境，并用属性标记数据资源，通过将本地属性映射到全局属性，以达到统一的访问控制。Huseyin等[20]认为应该为数据集成分析提供细粒度的访问控制，并设计了一种细粒度的访问控制系统GuardMR，该系统使用一种对象约束语言，并自动将策略转换成Java字节码来对MapReduce过程实施访问控制。Gedare和Rahul[21]认为在分布式环境中，访问控制通过一个中心的访问管理器进行决策，但这样会制约系统的性能，因此提出了一种硬件级别的权限缓存，提高系统的决策速度。

数据融合集成了来自多个数据源的数据，每个数据源由不同的安全策略管控，因此上述方法存在以下问题：扩展性受限，上述方法都是对安全策略进行统一管理，随着数据源及数据量的增加，将制约系统的扩展；策略存在冲突，不同机构有自身的安全策略，它们之间很有可能存在冲突的情况。因此，研究数据融合的安全策略融合对其安全防护有重要意义。安全策略融合是将多个访问策略融合，解决安全冲突并生成一个新的策略，该策略能够符合原有的安全要求。现有的研究工作中，Rao[22,23]使用逻辑代数表示安全策略，并提出一种基于代数运算的方法生成融合策略。但由于数理逻辑运算极有可能返回未知的结果，导致系统决策的不确定性，影响系统的可用性。Hu[24]使用基于语义的安全策略，通过本体映射和合并，将

查询语句重写成实体和属性名称，并映射到本地查询。Cruz将本地策略存储在RDF（resource description framework，资源描述框架）中，并在融合过程将本地RDF转变成一个全局RDF。

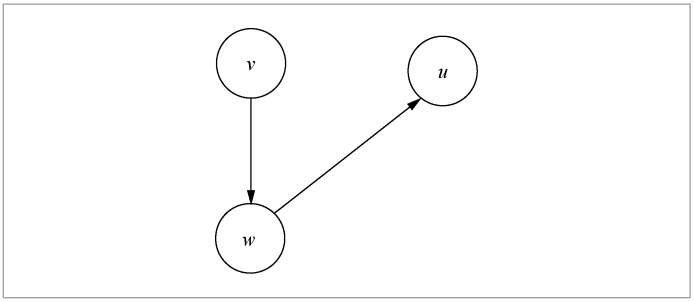


图1 有向图

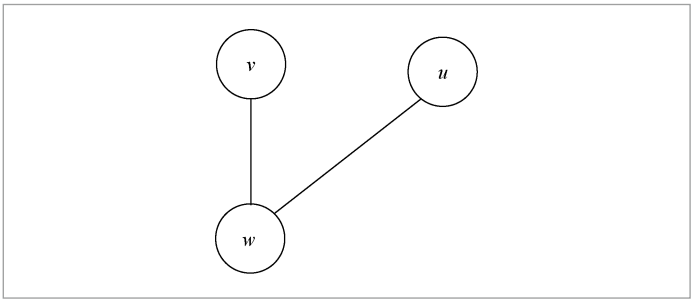


图2 无向图

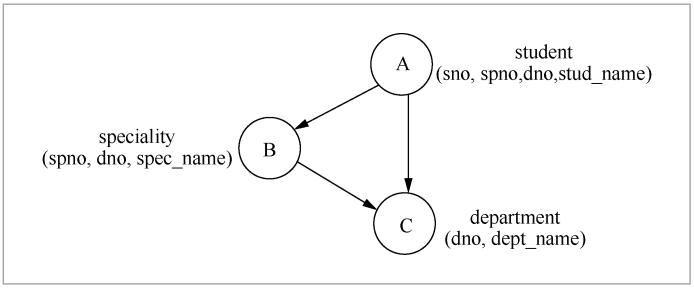


图3 数据库的图表示

3 数据融合中的模式转换

3.1 图模型

图是由一个顶点的有穷非空集合$V(G)$和一个弧的集合$E(G)$组成，通常记作$G=(V,E)$。图中的顶点即数据结构中的数据元素，弧的集合E是定义在顶点集合上的一个关系。用有序对$<v,w>$表示从v到w的一条弧。弧是有方向性的，用带箭头的线段表示，v为弧尾（始点），w为弧头（终点），该图为有向图，如**图1**所示。其中$V(G)=\{v,w,u\}$，$E(G)=\{<v,w>,<w,u>\}$。如果图中从v到w有一条弧，同时从w到v也有一条弧，那么该图称为无向图，如**图2**所示，用无序对$(v,\ w)$表示v和w之间的一条边，其中，$V(G)=\{v,w,u\}$，$E(G)=\{(v,w),(w,u)\}$。

3.2 数据库的图表示

一个学生管理系统的数据库可以采用如**图3**所示的有向图表示其依赖关系。

假设以下条件。

t_i：表示数据库中的一个表。

T：表示数据库中表的集合。

$G=<V,E>$：有向无环图（DAG），表示数据库的关系图。其中，v_i表示图中的一个节点，对应数据库中的一个表i，$V=\{v_1,v_2,\cdots,v_k \mid 1\leqslant k\leqslant n\}$是图中的点集，表示数据库中所有表的集合；$e=<v_i,v_j>$是图中的一条有向边，表示数据库中表$t_i$外键引用表$t_j$，$E=\{<v_i,v_j> \mid 1\leqslant i,j\leqslant n,i\neq j\}$是图中的边集，表示数据库中所有外键引用关系的集合。规定$|V|\geqslant 1$。

数据库DB的表集合$T=\{t_1,t_2,\cdots,t_k \mid 1\leqslant k\leqslant n\}$，则数据库$DB$的图表示为：$G=f(DB)$。令$G=<V,E>$，$\forall t\in DB.T$，则有$v_t\in G.V$和$e_t=<v_t,v_j>\in G.E$，此外没有其他的$V$和$e$属于$G$。

上述建模过程生成了数据库的依赖图，图中节点（数据库的表）的依赖关系由图中的边来表示。因此，当两个节点之间有边相连时，两个节点之间有相应的依赖关系，具体由边的方向决定。

数据融合过程在一定程度上是针对图

进行边的消减的过程，以形成一个或者多个独立的节点。每一条边的消除，同时需要把边的两端节点的数据进行融合，减少对应的节点外在依赖，即形成了融合后的数据。当一个节点的所有边都消除后，该节点就成为自包含的数据节点。

算法的主要问题是扩展顺序，即节点间的消边顺序。如**图4**所示，本算法思想是从叶子节点开始往上层节点扩展处理，因为只有叶子节点和孤立节点是当前已经包含完整信息的节点，即数据表。它们不再需要引用其他表的信息，那么它们就是已经包含完整信息的表。所以按照这种顺序扩展后能保证被嵌套扩展的节点已经包含了完整信息，那么扩展后的节点也就会包含完整信息。

核心算法就是从传统关系型数据库的模式图G中的叶子节点集P里取出节点v，取出以该节点为弧尾的边$<u,v>$，对该边的弧头节点u进行扩展，即把v节点的全部信息插进节点u中。当节点u扩展完毕，即没有以该点为弧头的边，就把节点u放入叶子节点集P。当叶子节点v不再被任何节点依赖，即没有以该节点为弧尾，就把该节点v移出节点集P，放入孤立节点集T。如此循环处理叶子节点集，直到叶子节点集P为空集。

本算法输入$G=(V,E)$是有向无环图，其中，V为G的点的集合，E为G的边的集合。规定$|V|\geq 1$。输出是一个二元组序列$S=\{<u,v>|<u,v>\in E\}$，表示扩展顺序。按照顺序S扩展后，模式转换为$G'=(V',E')$。其中，V'为G'的点的集合，E'为G'的边的集合，为空集。为了表述方便，下面将“节点”简称为“点”，“关系边”简称为“边”。

4 安全策略融合

如**图5**所示，在每个数据源上有多个数据集，而这些数据源需要进行整合，融

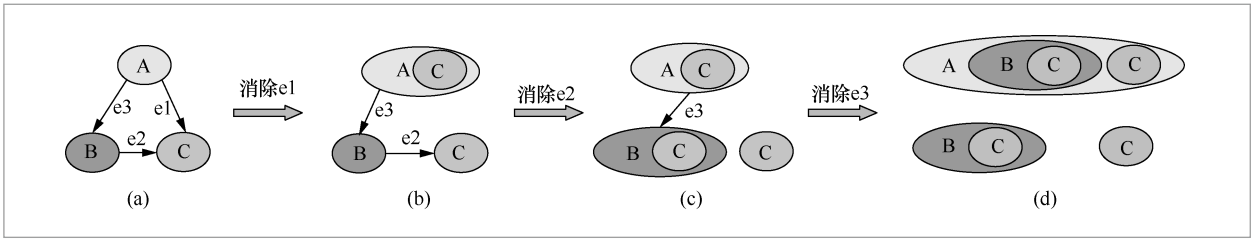


图 4 算法消除边的示意

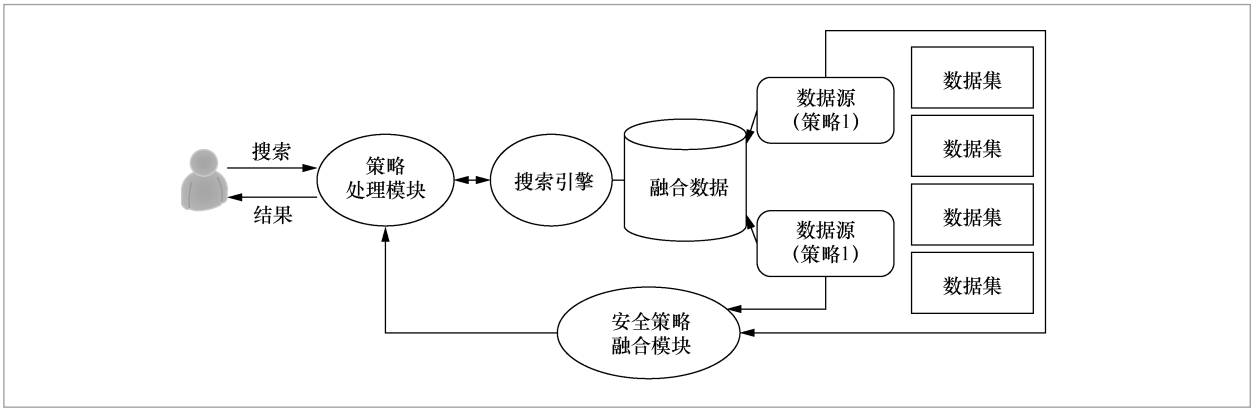


图 5 融合数据搜索系统的架构示意

合在一起形成一个新的数据集。用户在搜索融合的数据集得到查询结果。因此，融合搜索由以下几个关键部分组成，分别是用户、搜索引擎、融合数据、数据源、数据集、记录、安全策略融合模块以及策略处理模块等，其中安全策略融合模块将每个数据源的访问策略进行融合，生成一个融合访问策略，而策略处理模块则是对融合生成数据集根据安全策略进行安全标记，并且过滤不符合安全要求的结果。

假定每个数据源都是基于BLP（Bell-LaPadula）模型下建立访问策略的。因此，根据BLP模式，访问策略P_i定义为$P_i=(f_i,LTC_i,M_i)$，其中，i表示第i个数据源。当不同的数据源合并在一起，就会产生一个新的融合数据集。因为不同的数据源之间存在一些差异，所以融合的访问策略为$P_C=(f_C,LTC_C,M_C)$必须处理融合时的冲突，并且保持与原有数据源中的访问策略一致。而融合过程主要是3部分的融合：Lattice的融合、映射函数的转换以及访问控制矩阵的融合。

4.1 Lattice融合

Hasse图②是一种用于表达有限的偏序关系集合的图，以图形形式表现偏序关系集合的传递关系在偏序集合$<S,\leqslant>$，S的每个元素在Hasse图是一个顶点。而对于两个元素s_1和s_2满足偏序关系，即$s_1\in S$和$s_2\in S$并且$s_1\leqslant s_2$，则在Hasse图里偏序关系表示一段有向线段，从s_2指向s_1。

因为Lattice是一种特殊的偏序关系集合，所以Lattice也可以用Hasse图来表示。因此，Lattice的融合可以转换为两幅Hasse图的合并。合并过程主要分为3个阶段：初始化阶段、冲突处理阶段和化简阶段。初始化阶段是在两幅原始的Hasse图之间添加满足偏序关系的线段。在添加关联线段后，融合Hasse图可能会存在与原有Hasse图的冲突，所以需要对融合Hasse图进行冲突检测和处理，删除一些冲突线段。最后，还需要对融合Hasse图进行化简，删除冗余的线段。

4.1.1 初始化阶段

假设两个Lattice表示为$LTC_1=<S_1,R_1>$和$LTC_2=<S_2,R_2>$。在初始化阶段，需要对两个Lattice之间的节点关系进行考虑。而两个节点之间的关系分为两种：一种是相等关系，另一种是支配关系。

定义1 假设$l_1=<c_1,k_1>$、$l_2=<c_2,k_2>$分别是两个安全等级。当且仅当$c_1=c_2$和$k_1=k_2$时，l_1与l_2是相等关系。

定义2 假设$l_1=<c_1,k_1>$、$l_2=<c_2,k_2>$分别是两个安全等级。当且仅当$c_1\geqslant c_2$和$k_1\supseteq k_2$，则l_1与l_2是支配关系。

如**图6**所示，根据以上两个定义，在Lattice融合的初始化阶段，针对两个Hasse图之间的节点关系，得出以下规则：

- 若两个Hasse图之间的顶点满足相等关系，则在两个顶点之间添加两条互相指向的有向线段；
- 若两个Hasse图之间的顶点满足支配关系，则在两个顶点之间添加一条由支配顶点指向被支配顶点的有向线段。

② https://en.wikipedia.org/wiki/Hasse_diagram

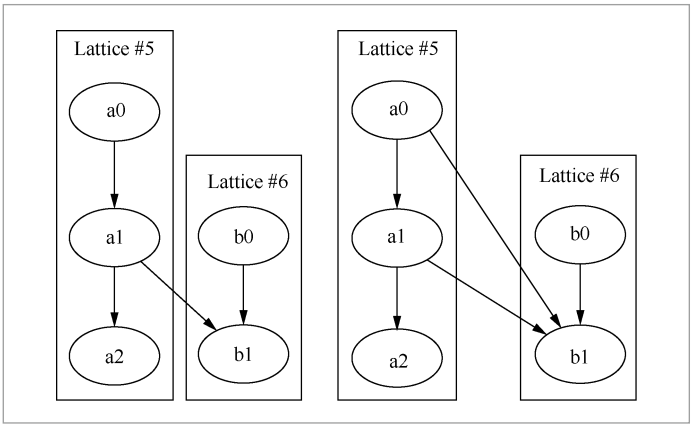


图 6 Lattice 图合并示意

4.1.2 冲突处理阶段

在添加了两个节点的关系线段之后，此时的融合Hasse图可能存在冗余的线段或者冲突线段。因此，接下来要处理的就是那些与原有Lattice的Hasse图冲突的线段。首先，给出Hasse图里的线段定义。

定义3 路径在Hasse图中是一系列的有向线段，连接着一系列的顶点，而连接之间的顶点只出现一次。

定义4 回路在Hasse图中是一条特殊的路径，开始顶点与结束顶点是同一个顶点，且经过多于2个顶点。

在Hasse图里面的两个节点的关系可分为可比关系和不可比关系。

定义5 假设s_1和s_2分别是Hasse图里的两个节点，当且仅当s_1和s_2之间存在路径时，s_1和s_2之间的关系是可比关系。

定义6 假设s_1和s_2分别是Hasse图里的两个节点，当且仅当s_1和s_2之间不存在路径时，s_1和s_2之间的关系是不可比关系。

定义7 当如下两种情况之一出现时，表示一条路径是冲突的：若这条路径是回路；若这条路径起始点和结束点在原有的Hasse图中是不可比关系，但这条路径在合并Hasse图中变得可比。

根据上述定义，对合并过程中出现的两种冲突情况进行讨论，如**图7**所示。

（1）合并Hasse图存在回路

在初始化阶段添加了两个原有Hasse之间节点的全部关联线段后，在生成的合并Hasse图可能会存在一条回路。

（2）在原有Hasse图中，不可比的两个节点在合并的Hasse图中存在路径

在原来的Hasse图中存在两个不可比的节点。但因为初始化节点添加关联线段后，使得这两个节点变得可比。即在某个Lattice里，两个安全等级l_1和l_2是不可比的。但在添加了两个Lattice之间的关联线段后，l_1和l_2之间可能就存在一条路径，使得l_1和l_2变得可比。

针对这两种情形，给出以下两条规则来处理冲突的线段。

- 规则1：删除在冲突路径中出现次数最多的关联线段。
- 规则2：若规则1不适用，则删除在冲突路径中涉及的安全级别最高的关联线段。

4.1.3 化简阶段

经过冲突处理阶段后，合并Hasse图应该不存在任何具有冲突的路径，但此时的图可能会比较冗余，因此需要对

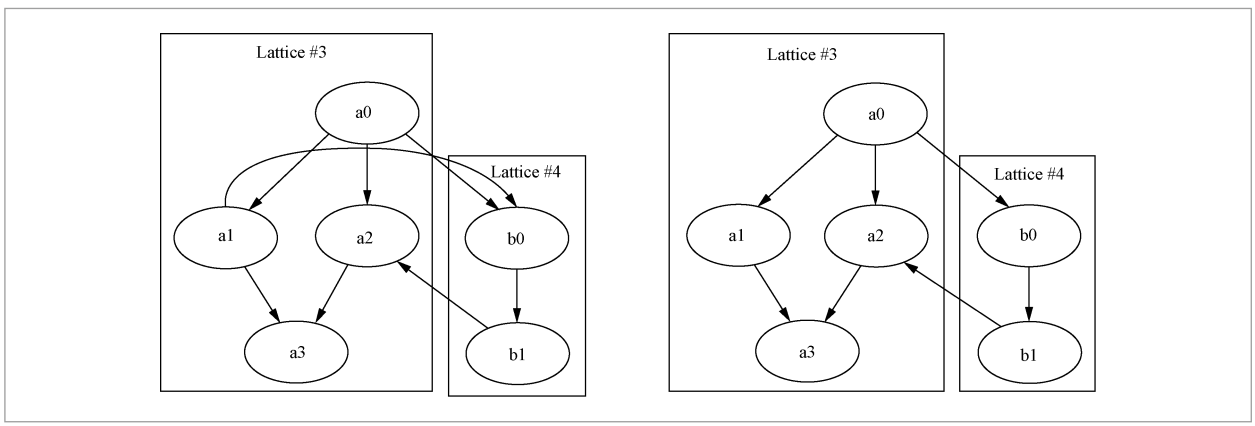


图 7 Lattice 图冲突解决示意

Hasse图进行最后一个步骤，化简操作，如**图8**所示。

定义8 假设在Hasse图中有两个节点s_1和s_2。当且仅当两条路径互相直接指向对方，即$s_1 \rightarrow s_2$和$s_1 \rightarrow s_2$，则这两条路径是平等关系。

定义9 假设在Hasse图中有两个节点s_1和s_2。当且仅当一条路径是s_1直接指向s_2，如$s_1 \rightarrow s_2$，而另一条路径是由s_1到s_2，并且中间经过若干个节点，如$s_1 \rightarrow \cdots \rightarrow s_2$，则这两条路径是覆盖关系。

定义10 冗余线段就是指那些满足平等关系或覆盖关系的关联线段。

因此，若冲突处理后的Hasse图存在冗余线段，按照以下两条规则对冗余线段进行删除，并化简Hasse图，得到最终简化的Hasse图。

- 若两条路径是平等关系，则对路径涉及的两个节点进行合并，生成新的节点。
- 若两条路径是覆盖关系，则删除那条从起始点直接指向结束点的关联线段。

4.2 映射函数转换

在安全策略融合后，需要将原始的Hasse图上的安全等级映射到新生成的Lattice图的安全等级。在Hasse图中，每个安全级别对应的是图中的节点。因此，安全级别的映射转换就等同于在原有Hasse图上的节点映射到融合Hasse图的节点。

本文定义了两个映射函数的转换函数。f_i^G表示从原始$Lattice_i$映射转换为融合$Lattice$映射，其中，i表示原始的格LTC_i。f_i^G表示从融合$Lattice$映射转换为原始$Lattice_i$映射。f_i^G函数是将原始的安全等级转换为全局的、融合的安全等级。而f_G^i则相反，即将全局的、融合的安全等级转换为原始的安全等级。

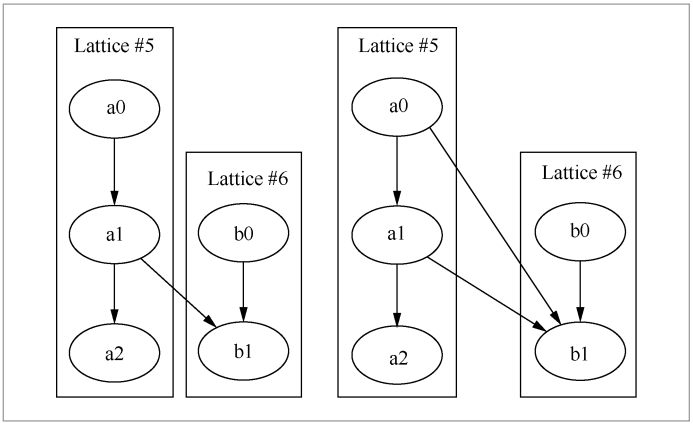


图 8 Lattice 图化简示意

4.3 访问控制矩阵融合

访问控制矩阵标识了主体对客体是否拥有访问权限，若主体拥有客体的访问权限，则将访问矩阵对应的元素设置为真。在合并两个访问控制矩阵形成新的访问控制矩阵时，融合数据集的访问属性与进行合并的数据集访问属性相关。为了保护数据的机密性，当合并前两个数据集在访问矩阵中均能访问时，合并后的数据集才可以访问。

当两个访问控制矩阵进行合并时，新的访问控制矩阵的主体是两个数据源的主体的并集，客体是两个数据源的并集与新融合的数据集。若主体对两个融合数据集具有访问权限，则主体对两个数据集都具有访问权限，那么主体对新数据集拥有访问权限，新矩阵中对应的元素设置为真，否则设置为假。

5 案例分析

刑事共犯的追踪主要是要融合相关情报部门整理的多个情报源的数据，根据给定人员的信息，通过融合的情报数据对关联任务进行发现和追踪。情报部门的每个情报源刻画的是一个社会侧面的活动，如

出租屋信息刻画的是社会人员租赁房屋和居住的信息，铁路出行刻画的是市民利用铁路作为交通工具的乘坐信息。融合后的数据可以同时反映出不同侧面的活动，提供了更加完整的信息。刑事共犯的数据融合将相关人员的证件号码、电话号码等信息作为关联的依据。

通过这些信息，融合后的数据可以提供同行同住、频繁邻近空间交往、疑似同伴等侦查过程需要的分析挖掘能力，如图9所示。若依靠传统手段，如市民A做了坏事，市民B是A的亲戚，A做不做坏事，B都跟A是亲戚，没有意义。融合后的数据要找的是A做了坏事，当时跟A在一起的有什么人，比如他们在相近时间住在相邻的酒店、他们经常在某些地方先后出现等。这种关联不是很明显，但是它是很有价值的，因为就算他们不是同行，他也有可能是见证人，有可能见证了事件的发生。所以需要融合数据来分析怎么把不相关的事情关联起来，这就需要从数据处理的角度分析，在事件网络上做信息的协同挖掘，找到他们有可能关联的行为。

6 结束语

本文从大数据治理中的数据模式转换和安全防护的角度，讨论了大数据中割裂数据的融合问题，通过发现结构化数据的数据模式和识别数据中的实体以及实体之间的关联关系，依据关联关系重组数据的存储和组织形式，消除数据的外部依赖，以减少在大数据分析挖掘过程中对数据的重复查找和组合的工作。同时，针对数据的重组过程提出了基于Bell-LaPadula模型的数据保护机制。该机制在数据按照相应需求进行重组的同时，对数据访问控制的安全策略进行了相应调整。调整后的新安全策略能够使数据的私密性得到保障，提供不低于原有安全策略的数据访问保护。

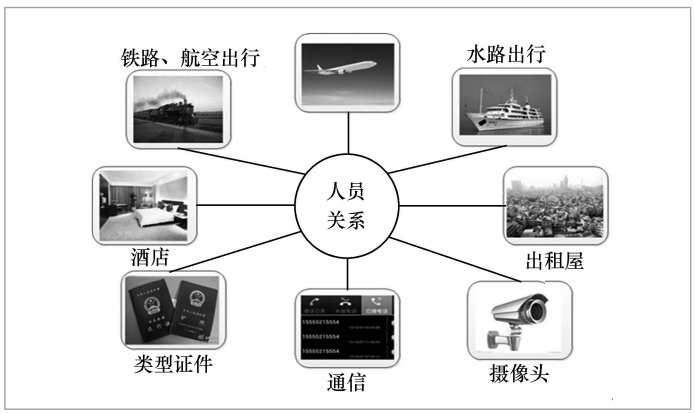


图 9 刑事共犯数据融合示例

参考文献:

[1] 马双荣.该如何面对大数据来袭[N].解放军报，2014-04-17.
MA S R. How to face the incoming data[N]. Jiefangjun Bao, 2014-04-17.

[2] 张一鸣. 数据治理过程浅析[J].中国信息界，2012(9)：15-17.
ZHANG Y M. Analysis of the data governance process[J]. Information China, 2012(9): 15-17.

[3] 桑尼尔·索雷斯. 大数据治理[M]. 匡斌，译. 北京：清华大学出版社，2014.
SUNIL S. Big data governance[M]. Translated by KUANG B. Beijing: Tsinghua University Press, 2014.

[4] 梁芷铭. 大数据治理：国家治理能力现代化的应有之义[J].吉首大学学报（社会科学版），2015，36(2)：34-41.
LIANG Z M. Mega data governance: an essential approach to the modernization of state governance[J]. Journal of Jishou University(Social Science Edition), 2015, 36(2): 34-41.

[5] 张兰廷. 大数据的社会价值与战略选择[D].北京：中共中央党校，2014.
ZHANG L T. Social value and strategic choice of big data [D]. Beijing: Party

School of the Central Committee of C.P.C, 2014.

[6] 谢红卫, 汪浩, 苏建志. 数据融合技术[J]. 系统工程与电子技术, 1992(12): 40-49.
XIE H W, WANG H, SU J Z. Data fusion technology [J]. Systems Engineering and Electronics, 1992(12): 40-49.

[7] LENZERINI M. Data integration: a theoretical perspective[C]//The 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, June 2-6, 2002, Madison, WI, USA. New York: ACM Press, 2002: 233-246.

[8] DONG X L, SRIVASTAVA D. Big data integration[C]// 2013 IEEE 29th International Conference on Data Engineering (ICDE), April 8-11, 2013, Brisbane, Australia. New Jersey: IEEE Press, 2013: 1245-1248.

[9] CARUCCIO L, DEUFEMIA V, MOSCARIELLO M, et al. Data integration by conceptual diagrams[C]// Database and Expert Systems Applications, Sep 1-5, 2014, Munich, Germany. Berlin: Springer International Publishing, 2014: 310-317.

[10] 张永新. 面向Web数据集成的数据融合问题研究[D]. 济南: 山东大学, 2012.
ZHANG Y X. Research on data fusion for web data interation[D]. Jinan: Shandong University, 2012.

[11] 穆化鑫. 基于Storm引擎的物联网异构数据融合系统的设计与实现[D]. 北京: 北京邮电大学, 2015.
MU H X. Design and implementation of IoT data fusion system based on Storm[D]. Beijing: Beijing University of Posts and Telecommunications, 2015.

[12] 石西庆. 基于数据融合技术的电子政务信息共享服务平台模型[D]. 成都: 电子科技大学, 2013.
SHI X Q. A model of e-government information sharing service platform based on data fusion technology[D]. Chengdu: University of Electronic Science and Technology of China, 2013.

[13] 化柏林, 李广建. 大数据环境下多源信息融合的理论与应用探讨[J].国书情报工作, 2015(16): 5-10.
HUA B L, LI G J. Discussion on theory and application of multi-source information fusion in big data environment[J]. Library and Information Service, 2015(16): 5-10.

[14] PAN L, XU Q. Visualization analysis of multidomain access control policy integration based on treemaps and semantic substrates [J]. Intelligent Information Management, 2012, 4(5): 188-193.

[15] 孟小峰, 张啸剑. 大数据隐私管理[J]. 计算机研究与发展, 2015(2): 265-281.
MENG X F, ZHANG X J. Big data privacy management[J]. Journal of Computer Research and Development, 2015(2): 265-281.

[16] SELLAMI M, GAMMOUDI M M, HACID M S. Secure data integration: a formal concept analysis based approach[J]. Database and Expert Systems Applications, 2014(8645): 326-333.

[17] SUN W, WANG B, CAO N, et al. Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking[C]//The 8th ACM SIGSAC Symposium on Information, Computer and Communications Security, May 8-10, 2013, Hangzhou, China. New York: ACM Press, 2013: 71-82.

[18] CAO N, WANG C, LI M, et al. Privacy-preserving multi-keyword ranked search over encrypted cloud data[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(1): 222-233.

[19] RUBIO-MEDRANO C E, ZHAO Z, DOUPé A, et al. Federated access management for collaborative network environments: framework and case study[C]//The 20th ACM Symposium on Access Control Models and Technologies, June 1-3, 2015, Vienna, Austria. New York: ACM Press, 2015: 125-134.

[20] ULUSOY H, COLOMBO P, FERRARI E, et al. GuardMR: fine-grained security policy enforcement for MapReduce systems[C]// The 10th ACM Symposium on Information, Computer and Communications Security, Apr 14-17, 2015, Singapore. New York: ACM Press, 2015: 285-296.

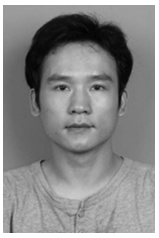
[21] BLOOM G, SIMHA R. Hardware-enhanced distributed access enforcement for role-based access control[C]//The 19th ACM Symposium on Access Control Models and Technologies, June 25-27, 2014, London, ON, Canada. New York: ACM Press, 2014: 5-16.

[22] RAO P, LIN D, BERTINO E, et al. An algebra for fine-grained integration of XACML policies [C]// The 14th ACM Symposium on Access Control Models and Technologies, June 3-5, 2009, Stresa, Italy. New York: ACM Press, 2009: 63-72.

[23] RAO P, LIN D, BERTINO E, et al. Fine-grained integration of access control policies [J]. Computers & Security, 2011, 30(2-3): 91-107.

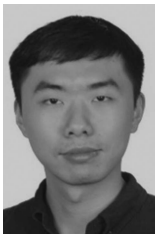
[24] HU Y J, YANG J J. A semantic privacy-preserving model for data sharing and integration [C]//The International Conference on Web Intelligence, Mining and Semantics, May 25-27, 2011, Sogndal, Norway. New York: ACM Press, 2011: 1-12.

作者简介

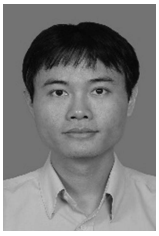


马朝辉(1974-),男,华南师范大学计算机学院博士生,广东外语外贸大学思科信息学院讲师,主要研究方向为网络安全、云计算和大数据等。

聂瑞华(1963-),男,华南师范大学计算机学院教授,中国计算机学会高性能计算专业委员会委员,广东高等教育学会信息网络专业委员会副理事长,华南师范大学“教育部互联网应用创新开放平台示范基地”负责人,主要研究方向为计算机网络及应用、云计算与大数据等。



谭昊翔(1990-),男,华南师范大学计算机学院硕士生,主要研究方向为信息安全和大数据等。



王欣明(1980-),男,博士,华南师范大学计算机学院讲师,IEEE会员,主要研究方向为软件工程、程序分析和大数据等。

唐华(1973-),男,华南师范大学软件学院院长助理、副教授,广东省科技咨询专家库专家,中国计算机学会计算机应用专家委员会委员,主要研究方向为计算机网络、信息安全、云计算和大数据等。

林嘉洛(1992-),男,华南师范大学计算机学院硕士生,主要研究方向为大数据和数据挖掘等。

杨晋吉(1968-),男,华南师范大学计算机学院教授,主要研究方向为逻辑、信息安全。

赵淦森(1977-),男,博士,华南师范大学计算机学院教授、副院长,广东省服务计算工程中心副主任,中国电子学会云计算专家委员会专家委员,粤港信息化专委会委员,中国信息系统专委会委员,广东省计算机学会常务理事,主要研究方向为信息安全、云计算和大数据等。

收稿日期:2016-02-28

通信作者:赵淦森,gzhao@scnu.edu.cn

* 本文为2015中国大数据技术大会(BDTC)演讲约稿