



一种基于 Spark 的大数据匿名化系统实现

卞超轶^{1,2}, 朱少敏¹, 周涛¹

(1. 北京启明星辰信息安全技术有限公司, 北京 100193;

2. 北京邮电大学, 北京 100876)

摘要: 分组匿名化框架是一类经典的数据匿名化技术, 它通过构造匿名记录的组, 使得同一组内的不同数据无法被识别区分, 从而达到隐私防护的效果。电力行业大数据分析涉及电力企业核心数据、用户隐私数据, 其数据敏感度更强, 传统的数据匿名化系统已经无法满足电力行业大数据业务应用和安全防护的需要。基于此, 设计并实现了一种基于 Spark 的新型大数据匿名化系统, 提供对 Hadoop 平台上多种数据格式的支持, 并有效提高对大数据的匿名化处理效率。

关键词: 数据匿名化; 隐私; 电力行业; 安全防护; Spark

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018133

Implementation of a big data anonymization system based on Spark

BIAN Chaoyi^{1,2}, ZHU Shaomin¹, ZHOU Tao¹

1. Beijing Venus Information Security Technology Incorporated Company, Beijing 100193, China

2. Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: Group based anonymization is a classical data anonymization framework, which achieves the effect of privacy protection by constructing groups of anonymized data records ensuring that records in the same group cannot be distinguished with each other. The electric power industry big data analysis involves the core data of the power enterprises and the user privacy data, the data sensitivity is stronger, traditional data anonymization systems are unable to meet the needs of big data business applications and safety protection of electric power industry. A new big data anonymization system based on Spark was designed and implemented, which could provide the support for multiple data formats stored on Hadoop and substantially improve the processing efficiency for big data.

Key words: data anonymization, privacy, electric power industry, safety protection, Spark

1 引言

电力行业大数据分析与应用促进了电力系统运行方式、管理体制、发展理念和技术路线等方

面的重大变革, 成为智能电网和全球能源互联网创新发展的重要技术支撑。电力大数据分析涉及电力企业核心数据以及用户隐私数据, 具有实时性更高、数据敏感度更强等特点, 其数据安全与

收稿日期: 2017-12-10; 修回日期: 2018-02-10

隐私保护的需求更加迫切, 安全技术要求更高。例如, 为了分析用户用电行为, 需要将用电数据提供给内外部分析人员, 但是原始数据中包含涉及用户隐私部分, 一旦大规模泄露并被不法分子利用, 存在巨大安全隐患; 在电力企业运行管理方面, 目前广泛利用大数据技术实现电源电网规划、发输变电设备状态评估、配电网供电可靠性等研究与应用, 这些海量异构大数据是电力企业内部关键的运行管理数据, 关系到电力企业的安全稳定运行, 大规模研究与利用对数据处理、使用和存储等方面的信息安全带来了新挑战; 防范大数据环境下电力系统核心数据和用户敏感数据的采集、上传及传播过程中的敏感数据泄露风险, 迫切需要研究适应电力业务大数据分析架构的大数据去隐私化处理理论、模型和算法, 形成大数据多层次全方位敏感数据保护基础环节。

数据匿名化是指以隐私防护为目的而对数据敏感部分进行特殊处理的过程, 典型处理手段包括加密、单向散列、消去、掩盖和模糊泛化等, 以避免数据在公开或共享时发生隐私泄露。分组匿名化 (group-based anonymization) 是一类经典数据匿名化技术框架, 它通过构造匿名数据分组使得组内的各条数据无法互相区分, 从而隐藏数据对应的用户身份。本文设计并实现了一种基于 Spark 实现的大数据匿名化系统, 利用分布式内存计算来提高计算处理效率, 支持 Hadoop 平台多种常用数据格式, 如 HDFS 文件、Hive 表等, 并对系统实现的效果进行了测试验证及性能评估。

2 相关工作

2.1 Hadoop 平台

Hadoop^[1]是由 Apache 负责开发及维护的开源软件框架, 主要目标是提供对大数据分布式存储及计算处理的支持。Hadoop 的核心是分布式文件系统 HDFS 与分布式计算引擎 MapReduce。

Spark^[2]是一种分布式内存计算引擎, 相比于

MapReduce, 启用了内存分布数据集, 将计算任务中间结果保存在内存中, 避免了 MapReduce 计算处理中利用 HDFS 中转而带来的大量磁盘 I/O, 大幅提高并行计算效率。在存储方面, Spark 沿用 HDFS, 实现对 Hadoop 平台上的多种数据管理访问直接支持。

2.2 分组匿名化

传统数据匿名化方法仅仅通过去除数据中的身份信息 (例如身份证号、姓名等) 达到信息隐匿效果, 但是仍可利用一些其他信息来定位识别个体。例如, 表 1 中虽然没有包含身份信息, 但是仍可以根据其中的非敏感信息 (年龄、国籍及邮编) 推断识别出具体个人, 从而造成用户用电量这类隐私敏感信息的泄露。这些可以被用于识别个体身份的非敏感信息被称为准标识符 (quasi-identifier)。数据匿名化技术正是通过对数据中的这些准标识符进行处理以达到隐私保护的效果。

表 1 包含敏感信息的数据

序号	非敏感信息			敏感信息
	邮编	年龄/岁	国籍	用电量
1	15072	29	英国	较大
2	15041	27	美国	较大
3	15041	22	日本	一般
4	15072	25	美国	一般
5	12904	51	印度	很小
6	12904	53	英国	较大
7	12907	43	美国	一般
8	12907	41	美国	一般
9	15072	33	美国	很小
10	15072	39	印度	很小
11	15041	38	日本	很小
12	15041	37	美国	很小

分组匿名化是一类经典数据匿名化技术, 其核心思想是通过构造匿名记录组, 使得同一组内的多条数据难以相互区分。 k -匿名化 (k -anonymity)^[3]要求处理后的数据中存在的任意准标识符取值的组合都与至少 k 条记录匹配。匿名化处理的手段有泛化和隐藏: 泛化是指将属性的取值进行模糊替换处理, 比如将“28”替换成



“<30”，将“中国”替换成“亚洲”等；隐藏是指将属性值隐去，例如使用“*”替换。表2给出了对表1数据经过4-匿名化处理后的一种结果。总共12条数据形成了3组，每组内的4条数据拥有完全相同的准标识符，无法互相区分。

表2 4-匿名化示例

序号	非敏感信息			敏感信息
	邮编	年龄/岁	国籍	用电量
1	150**	<30	*	较大
2	150**	<30	*	较大
3	150**	<30	*	一般
4	150**	<30	*	一般
5	1290*	≥40	*	很小
6	1290*	≥40	*	较大
7	1290*	≥40	*	一般
8	1290*	≥40	*	一般
9	150**	3*	*	很小
10	150**	3*	*	很小
11	150**	3*	*	很小
12	150**	3*	*	很小

k -匿名化有缺陷，如表2中最后一组敏感属性取值相同，这样仍然暴露了隐私信息，因此 l -多元化 (l -diversity) [4] 要求每种准标识符组合对应数据至少有 l 种不同敏感属性值。表3展示了对表1数据进行3-多元化处理后的结果，它保证了每组敏感属性至少有3种不同的取值。

表3 3-多元化示例

序号	非敏感信息			敏感信息
	邮编	年龄/岁	国籍	用电量
1	1507*	≤40	*	较大
2	1507*	≤40	*	一般
3	1507*	≤40	*	很小
4	1507*	≤40	*	很小
5	1290*	>40	*	很小
6	1290*	>40	*	较大
7	1290*	>40	*	一般
8	1290*	>40	*	一般
9	1504*	≤40	*	较大
10	1504*	≤40	*	一般
11	1504*	≤40	*	很小
12	1504*	≤40	*	很小

k -匿名化和 l -多元化是广泛应用的分组匿名化框架，已经有相应算法被提出以满足隐私要求[5]，也有适用于小规模数据的开源工具发布[6]，但是针对大数据场景仅有一些利用 MapReduce 来并行匿名化处理过程的工作[7]。为了满足电力大数据分析实际安全需求，本文设计了基于分布式内存计算引擎 Spark 开发，支持对以 HDFS 文件、Hive 表等格式存储于 Hadoop 平台的数据进行高效匿名化处理的系统，满足分组匿名化的隐私要求，实现数据公开及共享时的有效隐私防护。

3 系统实现

3.1 匿名化算法

本系统采用 Mondrian 算法[8]来实现数据匿名化。核心思想是先将所有准标识符对应的属性完全模糊化，这种情况下所有数据条目都是相同的，然后逐轮选取一个属于准标识符的属性增加公开信息，对数据进行不断划分，并确保划分出的任意子集包含的数据条目数均不小于 k ，也就是保证 k -匿名化要求被满足。算法每轮迭代中都会选取构成准标识符的一个属性对数据进行划分，循环迭代直至不能再划分，即无论如何继续划分都会造成某个子集不再满足 k -匿名化要求。数据属性有两种类型：数值属性和类别属性。前者的划分是通过选取中间值（或中位数）对数据进行二分，而后者则是根据可取的具体类别数对数据进行划分（二分或多分，由类别属性的取值数决定）。

图1给出了一个对包含年龄、邮编这两个数值属性的数据集使用 Mondrian 算法达到2-匿名化要求的处理过程示例。图1(a)为数据分布情况，以“★”符号表示数据条目。初始数据中的年龄及邮编属性被完全模糊化，例如均以“*”表示。假设首先选择邮编这一属性进行数据划分，则结果如图1(b)所示。这时位于图1(b)左侧的数据子集的邮编属性可由初始的“*”表示变化为“≤27K”，而右侧的数据子集的邮编属性变化为“>27K”。然后对位

于图 1 (b) 左侧的数据子集选择年龄这一属性再次进行划分, 结果如图 1 (c) 所示。相应地, 图 1 (c) 左下方数据的年龄属性由初始的“*”变化为“ ≤ 60 ”, 而左上方数据的年龄属性变化为“ >60 ”。同样对位于图 1 (c) 右侧的数据子集也选择年龄这一属性进行再划分, 结果如图 1 (d) 所示。这样图 1 (d) 右下方数据的年龄属性变化为“ ≤ 62 ”, 图 1 (d) 右上方数据的年龄属性变化为“ >62 ”。此时不能再进行任何划分, 否则会破坏 2-匿名化要求。最后数据被划分为 4 个分组, 每个分组内的数据无法互相区分。例如图 1 (d) 中左上方的两条数据的年龄属性均为“ >60 ”, 邮编属性均为“ $\leq 27K$ ”。

上述的 Mondrian 算法在简单修改后即可支持 l -多元化。具体来说, 在每次数据划分时不检查划分出的子集所包含的数据条目数, 而改为检查划分出的数据子集是否满足 l -多元化要求。如果满足就继续递归划分, 否则就回滚该次划分。

3.2 算法复杂度

Mondrian 匿名化算法应用于仅包含数值属性

的数据集时的工作过程与 kd 树的构造过程相同, 所以具有相同的计算复杂度, 即 $O(n \log n)$, 其中 n 为总数据集大小。而按类别属性进行划分时, 只需要遍历一遍相应数据即可, 与数值属性时按中位数二分的计算复杂度是相同的 (均为 $O(k)$, 这里 k 为要划分的数据集大小); 划分后形成的子集数虽然有差别, 但是对于总体计算复杂度的影响只体现在常数系数上。因此, Mondrian 算法应用于一般数据集的计算复杂度仍为 $O(n \log n)$, 具有良好的可扩展性, 适用于大数据场景。

3.3 实现细节

(1) 数值类型的划分实现方式

Mondrian 算法对数值属性划分时采取的方法是按中位数进行二分。由于在大数据规模下使用快速排序再定位中位数计算, 时间复杂度较高, 且线性时间复杂度的求中位数算法在 Spark 上实现较为复杂, 所以系统实现时通过统计每种取值出现的次数再快速定位出中位数, 从而发挥 Spark 的数据并行处理优势。

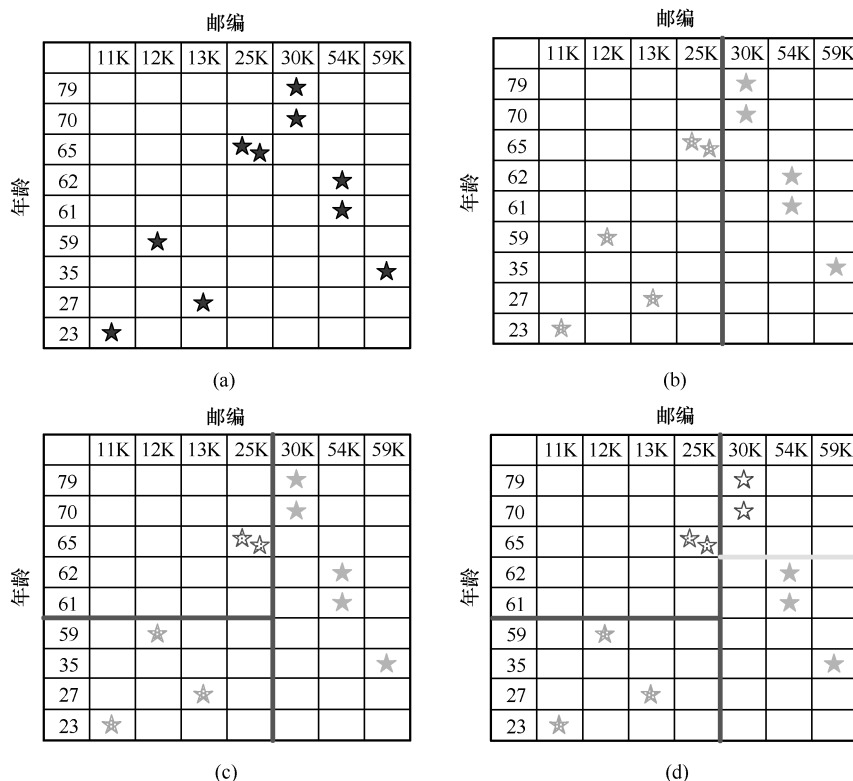


图 1 Mondrian 算法划分数值属性数据示例



(2) 递归算法的实现方法

Mondrian 匿名化算法是一个具有递归思想的算法, 自顶而下地对数据进行划分, 不断对划分出的子数据集使用相同过程进行再次划分, 直至满足停止条件。然而递归实现在 Spark 分布式计算平台上无法使用。本文改进为采用非递归实现, 使用一个队列记录之后要继续进行划分的子数据集, 然后按顺序从队列中取出依次进行。

(3) 输入配置文件

数据匿名化处理需要类别属性的泛化树作为输入。所谓泛化树, 就是类别属性进行泛化处理时的层次关系结构, 如“中国”→“东亚”→“亚洲”→“*”。本系统中采用文本格式配置文件提供这一输入, 文件中的每一行描述一个从叶节点(具体属性取值)到根节点(通常为“*”)的泛化关系, 例如上述的泛化关系可表达为:“中国; 东亚; 亚洲; *”。

4 系统验证

本文系统测试采用机器学习及数据分析挖掘中常用的数据集 Adult 进行。Adult 数据集是从美国 1994 年人口普查数据库抽取而来, 其中包含一个敏感属性——收入状况, 其他属性包括年龄、婚姻状态、职业、性别等。在测试中选取了其中 8 个非敏感属性(构成准标识符), 其中两个是数值属性, 其余 6 个均为类别属性。测试时将数据集以文本文件形式存放在 HDFS 中, 同时将匿名化算法所需要的泛化树等配置信息也置于 HDFS 中供系统初始化时读取。

4.1 效果测试

图 2 显示了测试数据集中准标识符属性的取值样例, 其中第一列与第三列是数值属性, 其余均为类别属性。而图 3 则给出了准标识符属性经过本系统进行 k -匿名化(k 取值为 10)处理后的样例。注意这两个图中的数据条目并非一一对应, 只是随机选取的展示样例。可以看出, 数值属性被泛化成一个区间, 如第一列的“[38,45]”和第

三列的“[11,14]”; 而类别属性则被泛化成“*”或其他更高层次的泛化表达。经检查确认, 本系统的匿名化处理结果与 Mondrian 算法的开源单机实现版本^[9]相同, 证明了本系统的匿名化算法实现的正确性。类似地, 对于 l -多元化也进行了相应测试, 同样能够得到预期正确结果, 本文不再单独列出。

```
59, Private, 9, Married-civ-spouse, Handlers-cleaners, White, Male, United-States
40, Private, 9, Divorced, Craft-repair, Asian-Pac-Islander, Female, Japan
28, Private, 9, Married-civ-spouse, Transport-moving, White, Male, United-States
39, Private, 9, Separated, Other-service, White, Female, El-Salvador
25, Private, 9, Never-married, Craft-repair, White, Male, United-States
48, State-gov, 10, Divorced, Exec-managerial, White, Female, United-States
62, Private, 5, Married-civ-spouse, Other-service, Black, Male, United-States
44, Local-gov, 9, Divorced, Adm-clerical, White, Female, United-States
28, Federal-gov, 13, Never-married, Prof-specialty, White, Male, United-States
62, Private, 10, Married-civ-spouse, Craft-repair, White, Male, United-States
32, Private, 9, Married-civ-spouse, Craft-repair, White, Male, United-States
59, Federal-gov, 9, Married-civ-spouse, Sales, White, Female, United-States
20, Private, 10, Never-married, Adm-clerical, White, Female, United-States
40, Private, 12, Separated, Tech-support, White, Female, United-States
47, Private, 10, Married-civ-spouse, Other-service, White, Male, United-States
60, Private, 10, Married-civ-spouse, Sales, White, Male, United-States
55, Private, 9, Separated, Protective-serv, White, Male, United-States
18, Private, 8, Never-married, Adm-clerical, White, Male, United-States
43, Private, 13, Never-married, Adm-clerical, White, Male, United-States
31, Self-emp-not-inc, 9, Divorced, Other-service, White, Male, United-States
22, Private, 9, Never-married, Craft-repair, White, Male, United-States
56, Private, 9, Divorced, Tech-support, Black, Female, United-States
41, Private, 9, Separated, Adm-clerical, White, Female, United-States
24, Private, 9, Never-married, Other-service, White, Male, United-States
```

图 2 原始数据集准标识符属性样例

```
[38,45], Local-gov, [11,14], leave, *, *, Male, *
[38,45], Local-gov, [11,14], leave, *, *, Male, *
[38,45], Local-gov, [11,14], leave, *, *, Male, *
[38,45], Local-gov, [11,14], leave, *, *, Male, *
[38,45], Local-gov, [11,14], leave, *, *, Male, *
[38,45], Local-gov, [11,14], leave, *, *, Male, *
[38,45], Local-gov, [11,14], leave, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[41,45], Federal-gov, [11,15], Divorced, *, *, Male, *
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
[47,60], State-gov, [11,16], Divorced, *, *, Female, United-States
```

图 3 经数据匿名化后的准标识符属性样例

4.2 性能评估

为了验证基于 Spark 实现的系统在对数据进行匿名化处理时能否有效提高效率, 对 Adult 数据集进行扩充之后开展了一些简单实验。原始 Adult 数据集包含约 30 000 条数据, 无法体现 Spark 在大数据处理方面的优势, 因此将其复制

100 倍、1 000 倍, 并对复制数据条目的数值属性进行随机修改, 从而获得供性能评估测试用的数据。具体来说, 复制 100 倍、1 000 倍后的数据条数分别达到约 3×10^6 和 30×10^6 , 相应的数据文件大小分别约为 380 MB 及 3.8 GB。使用单机版的匿名化程序和基于 Spark 实现的本系统分别进行相同的处理并记录各自消耗的时间, 对于两种大小的数据集均重复进行了 10 次实验。其中实验用的 Spark 集群由 3 台 Dell R720 服务器(CPU: Intel® Xeon® E5-2603 v2@ 1.80 GHz, 8 核; 内存: 64 GB DDR3@ 1 600 MHz) 组成, 而单机版则在其中一台服务器上运行。图 4 给出了实验结果, 直方块表示 10 次实验的平均值, 误差条表示标准差。结果显示基于 Spark 实现的匿名化系统比单机版程序对 380 MB 和 3.8 GB 数据集进行匿名化处理所用的时间分别少约 21% 和 47%, 表明基于 Spark 实现的系统能够有效提高数据匿名化处理的效率, 而且在数据规模增大时优势更为明显。

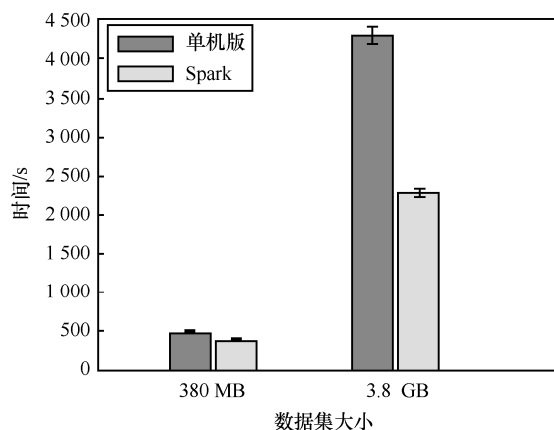


图4 数据匿名化处理时间对比

5 结束语

本文设计了一个基于分布式内存计算引擎 Spark 开发实现的大数据匿名化系统, 支持对 Hadoop 平台的多种存储格式的数据进行高效匿名化处理, 使用 Mondrian 算法达到 k -匿名化、 l -

多元化等分组匿名化框架, 能够在数据公开时对其中包含的隐私信息进行防护隐藏使其不会被恶意攻击者获取。系统性能相比于单机处理有明显提升。在未来的研究工作中, 将参考 Spark MLlib 中决策树算法的实现来进一步提高系统的处理效率, 并对数据经本系统匿名化处理后的价值损失进行理论分析与评估。

参考文献:

- [1] Apache Software Foundation. Apache Hadoop[EB]. 2011.
- [2] Apache Software Foundation. Apache Spark[EB]. 2014.
- [3] SWEENEY L. K -anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.
- [4] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. L -diversity: privacy beyond k -anonymity[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 3.
- [5] FUNG B, WANG K, CHEN R, et al. Privacy-preserving data publishing: a survey of recent developments[J]. ACM Computing Surveys (Csur), 2010, 42(4): 14.
- [6] PRASSER F, KOHLMAYER F. Putting statistical disclosure control into practice: the ARX data anonymization tool[M]. Berlin: Springer International Publishing, 2015.
- [7] JADHAV R H, INGLE R B. A survey on data anonymization approaches using MapReduce framework on cloud[J]. Digital Image Processing, 2015, 7(2): 48-49.
- [8] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Mondrian multidimensional k -anonymity[C]//International Conference on Data Engineering, April 3-7, 2006, Atlanta, GA, USA. Piscataway: IEEE Press, 2006: 25.
- [9] QIYUAN G. GitHub-qiyuangong/basic_mondrian[EB]. 2015.

[作者简介]



卜超轶 (1987-), 男, 博士, 北京启明星辰信息安全技术有限公司高级研究员, 启明星辰博士后工作站—北京邮电大学博士后流动站联合培养博士后, 主要研究方向为大数据自身安全、大数据安全分析等。

朱少敏 (1983-), 男, 博士, 北京启明星辰信息安全技术有限公司前线技术专家团成员, 主要研究方向为电力系统信息安全、二次系统安全防护、多媒体信息处理等。

周涛 (1979-), 男, 博士, 北京启明星辰信息安全技术有限公司教授级高级工程师, 主要研究方向为大数据安全分析、事件关联分析、入侵检测等。