计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(12):19−27 [doi: 10.15888/j.cnki.csa.007153]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 互联网标识隐私保护政策及技术研究①

闫宏强, 王　伟, 张　婕

(中国科学院 计算机网络信息中心, 北京 100190)
(中国科学院大学, 北京 100049)
通讯作者: 张　婕, E-mail: zhangjie@cnic.cn

**摘　要**: 互联网标识是互联网中识别和管理物品、信息、机器的关键基础资源, 是互联网运行和发展的重要基础. 在将个人信息与基础资源对应标识的过程中, 也存在个人信息泄露的重大安全隐患. 目前欧盟《通用数据保护条例》(GDPR)、《中华人民共和国网络安全法》等隐私保护法律的出台, 对个人信息安全保护提出了更严格的要求. 本文首先结合隐私政策法律条例, 以典型的互联网标识——域名为切入点, 深入分析互联网标识业务流程中的隐私风险点; 接着结合数据生命周期的不同隐私保护需求, 提出了域名信息在发布、存储、挖掘、使用各环节的隐私保护技术框架; 最后针对每种互联网标识数据生命周期, 提出具体隐私保护技术解决方案.

**关键词**: 互联网标识; 隐私保护; GDPR; 数据生命周期; 个人信息保护

引用格式: 闫宏强,王伟,张婕.互联网标识隐私保护政策及技术研究.计算机系统应用,2019,28(12):19–27. http://www.c-s-a.org.cn/1003-3254/7153.html

## Overview on Privacy Policy and Technology of Internet Identifier

YAN Hong-Qiang, WANG Wei, ZHANG Jie

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
(University of Chinese Academy of Sciences, Beijing 100049, China)

**Abstract**: Internet identifier is the key resource for identifying and managing goods, information and machines in the Internet. It is an important basis for the operation and development of the Internet. In the process of correspondingly identifying personal information and basic resources, there are also major security risks of personal information leakage. At present, the introduction of privacy protection laws such as the GDPR and the Cyber Security Law of the People's Republic of China imposes stricter requirements on the protection of personal information. This study first analyzes the privacy risk of the Internet identifier and discusses the challenges of personal privacy protection; then uses the typical Internet identifier-domain name as an example to deeply analyze the privacy risk in the Internet identifier service process. Combining the different privacy protection needs of the data life cycle, the Internet identifier privacy protection technology framework for the publication, storage, mining and use is proposed. Finally, specific privacy protection technology solutions are proposed for each kinds of Internet identifier data life cycle.

**Key words**: Internet identifer; privacy protection; GDPR; data lifecycle; personal information protection

互联网标识[1]是互联网运行和发展的重要基础, 是在互联网上, 唯一识别目标对象的编码、符号、名字, 是识别和管理物品、信息、机器的关键基础资源. 互联网标识本质上是用真实个人信息与基础资源对应来达到标识目的, 这其中涉及的一个重要问题是个人信息保护问题. 近年来, 有关个人隐私数据泄露事件频发,

① 收稿时间: 2019-04-10; 修改时间: 2019-05-08; 采用时间: 2019-05-20; csa 在线出版时间: 2019-12-10



© 中国科学院软件研究所 http://www.c-s-a.org.cn

引发社会和学术界广泛关注. 据报道, 2017 年 11 月, Google 利用算法成功避开了苹果手机的默认隐私设置, 非法收集大概 540 万名用户的历史浏览数据, 严重侵犯了用户隐私. 2018 年 3 月, 全球用户规模最大的社交应用 Facebook 被曝光有超过 5000 万名用户的个人信息资料遭到泄露, 被第三方公司用于大数据分析, 成为间接影响总统大选的隐形黑手, 在欧美社会产生巨大震动. 同样, 国内的隐私泄露问题也很严峻. 2014 年 3 月, 携程安全支付日历导致用户银行卡信息泄露. 2018 年 8 月, 华住集团旗下连锁酒店 5 亿条用户信息遭到泄露, 泄露的信息包括用户的注册信息、酒店入住信息和开房记录, 这或是国内近五年来规模最大最严重的一次个人信息泄露事件.

国内外的信息泄露事件敲响了互联网个人信息安全警钟. 国际社会和各国政府对隐私保护问题十分重视, 已经建立起较为完善的隐私保护制度. 1980 年, 世界经济与合作发展组织 (Organization for Economic Co-operation and Developmen, OECD) 发布了《隐私保护和个人数据跨境流动指南》(Guidelines on the Protection of Privacy and Transborder Flows of Personal Data)[2], 提出了 8 条隐私保护基本原则, 几十年来已经成为被世界各国广泛接受的隐私保护标准[3]. 欧盟 1995 年颁布《数据保护指令 (95/46/EC)》[4], 指导欧盟成员国隐私保护立法和执法工作. 2016 年, 欧盟实施全面的隐私和数据保护改革, 通过了直接适用于欧盟成员国的《通用数据保护条例》(General Data Protection Regulation, GDPR)[5], 重申并强化数据处理基本原则, 强化了数据主体 (data subjuect) 权利, 建立起严苛的企业问责制度[6]. 美国颁布了《联邦隐私法案》(Privacy Act of 1974)、《金融服务现代化法》(the Gramm-Leach-Bliley Act of 1999)、《联邦儿童在线隐私保护法》(Children's Online Privacy Protection Act of 1998) 等系列法律, 主要以联邦贸易委员会为主开展了一系列隐私保护执法行动[7]. 2007 年美国会计师事务所 (American Institute of Certified Public Accountants, AICPA) 和加拿大特许会计师协会 (Chartered Accountants of Canada, CICA) 发布了一个全球性隐私框架—公认隐私准则 (the Generally Accepted Privacy Principles, GAPP), 旨在帮助特许会计师和注册会计师创建有效的隐私计划, 以管理和预防隐私风险[8]. 2016 年 6 月, 我国颁布《中华人民共和国网络安全法》[9], 加强了对个人信息的保护力度, 完善了个人信息保护基本规则. 截至 2018 年, 世界上 120 多个国家和地区制定了综合性的个人信息保护的相关法律[10].

法律对于个人信息保护提出了严格的要求, 在实践中实施个人信息保护, 还需要系统性的个人数据隐私处理框架及合规体系, 还需要对具体隐私算法进行细化. 互联网标识相关联的个人信息也面临着严峻的安全形势, 主要涉及标识注册信息的隐私保护问题. 国内外学术界围绕隐私保护展开的研究工作主要是针对轨迹隐私保护和隐私计算算法的实现和改进, 缺少针对互联网标识这一应用场景的隐私政策及技术方案研究, 本文致力于填补这个研究空白.

本文结合法律分析了个人信息和隐私保护要求和相关工作的迫切需要, 以最典型的互联网标识——域名为例, 阐述互联网标识业务中涉及的隐私保护场景, 借鉴国内外隐私保护的最新研究理论, 提炼出互联网标识隐私保护技术的最佳方案建议.

本文其余部分的组织结构如下: 第 1 节介绍了互联网标识以及典型领域—域名领域, 第 2 节分析阐述了域名业务领域中涉及隐私泄露风险点, 第 3 节针对第 2 节的风险点、结合个人数据生命周期, 提出了互联网标识隐私保护技术框架, 第 4 节对于数据的每个生命周期中的隐私保护需求, 提出了技术解决方案. 第 5 节, 以随机可逆匿名化算法进行试验验证. 第 6 节总结与展望.

## 1 互联网标识及域名

互联网标识广义上是指用于互联网行为的所有标志性名称, 可以包括图像、文字、数字、声音等几种常见的形式. 狭义上的互联网标识是指机器在网络中的标志和寻址信息, 例如, MAC 地址、IP 地址或者域名可以作为一台机器的互联网标识.

常见的互联网标识有: 域名、自治系统号码、IPv4、IPv6 互联网地址、组播寻址、端口号码、协议号码、统一资源标识符 (URL).

在互联网中, 域名是最常用、最典型的标识, 是互联网上的"门牌号码", 是各种互联网应用的入口. 域名具有网络定位和身份定位双重作用[11], 由一串点分隔的字符组成, 用于在数据传输时标识计算机的电子方位, 在网络应用中起到地址和标识作用. 域名采用分层结构的名称空间, 可以从域名映射到其他标识.

随着互联网高速发展, 互联网用户在迅速增加, 域



© 中国科学院软件研究所 http://www.c-s-a.org.cn

名注册服务市场也在飞速发展. 据中国互联网络信息中心 (CNNIC) 第 43 次《中国互联网络发展状况统计报告》, 截至 2018 年底, 我国域名总数为 3792.8 万个, 其中".cn"域名总数为 2124.3 万个. 面对如此庞大的域名体系, 域名相关个人信息的保护成为行业重要的工作. 近年来, 网络安全形势日益严峻, 相关域名恶意解析和域名纠纷事件频发. 2010 年 1 月 12 日, 国内最大搜索引擎百度长时间无法正常访问, 经查, 原因是黑客篡改了百度域名在域名注册服务商的注册信息, 导致百度域名被指向错误的服务器. 无独有偶, 2005 年天涯社区也被人修改了域名注册信息, 将域名指向另一网站, 一度劫持了天涯社区的访问量.

类似事件都表明, 域名注册信息不仅是域名管理的联系方式, 也是域名持有者对该域名所有权 (使用权) 的法律依据. 对域名注册信息进行有效保护很重要, 如果注册信息不真实、不准确, 一旦域名注册信息被恶意篡改, 域名持有者的隐私安全和域名财产安全将受到损害. 而且, 不法分子往往利用虚假身份信息注册域名实施网络钓鱼、僵尸网络控制、传播违法信息等黑客犯罪行为, 以逃避追查和打击. 国际上在域名服务推广之初, 并未严格要求域名注册信息的实名制, 随着互联网的普及和应用, 不实的域名注册信息比例偏高, 已经成为困扰全球互联网产业健康发展的重要问题. 实施域名实名注册制度是大势所趋.

2003 年, 互联网名称与数字分配机构 (the Internet Corporation for Assigned Names and Numbers, ICANN) 出台了新版的《域名注册信息提醒政策》, 规定姓名、地址、联系方式等完整的注册信息中, 域名持有者必须确保所有信息真实、准确, 如果信息不真实、不准确, 域名会被注销. 2004 年制定实施的《中国互联网域名管理办法》[12]规定, 域名实名制要求用户注册域名时, 填写真实、准确、完整的注册信息, 并且要求全面实施域名实名认证. 以此保护域名注册者的合法权益, 防止域名被恶意盗取和滥用, 维护域名市场环境, 促进网络可信建设.

## 2 域名涉及隐私披露风险点

针对《网络安全法》和欧盟 GDPR 对个人数据 (欧盟称个人信息为个人数据) 提出的更严格的隐私保护要求, 本节以域名业务为例, 分析互联网标识数据在跨域跨境传输过程涉及的隐私披露风险点, 如图 1.

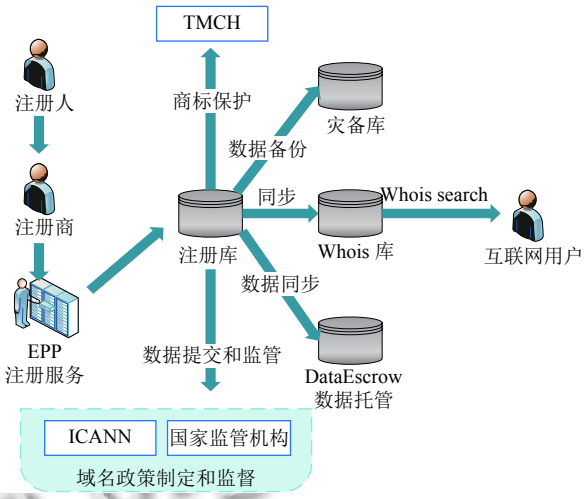



图 1 域名注册业务流程

注册人 (registrant) 选择域名后, 需要向注册商 (registrar) 或其代理商提交域名注册请求, 同时需要提交相关注册信息 (所需域名、注册人姓名、电话、地址等联系信息, 技术联系人信息和管理联系人信息, 以及注册期限). 注册商将检查该域名是否可用, 并按照注册人提供的信息建立一个 WHOIS 记录 (包含注册人、管理联系人和技术联系人的信息、创建日期、更新日期、域名服务器和域名状态), 通过可扩展注册协议 (Extensible Provisioning Protocol, EPP) 注册服务, 构建注册库. 并且向注册局 (registry) 提交数据, 注册局负责顶级域名的注册管理, 维护域名数据库.

为了防止注册局和注册商经营失败或遭受到恶意网络攻击而造成注册数据丢失或该顶级域名无法正常运转, ICANN 要求新通用顶级域名 (New generic Top-Level Domain, NewgTLD) 的申请人选择第三方数据托管服务机构 (data escrow agent) 向其提交注册数据, 进行数据托管.

ICANN 在执行新通用顶级域名计划时, 推出全新商标保护措施, 满足商标持有人的权益保护需求, 同时也避免商标持有人到各个注册局反复注册商标相关域名并提交、校验同样的商标信息. ICANN 推出了全球商标信息交换库 (Trade Mark Clearing House, TMCH), 作为已验证的商标集中存储的商标数据库. 在进行校验的过程中, 注册局或注册商需要向 TMCH 提供包括注册人信息在内的域名注册数据.

ICANN 要求注册管理机构每周向 ICANN 提供一次批量注册数据 WHOIS 的访问权限, 用以随机抽取注



© 中国科学院软件研究所 http://www.c-s-a.org.cn

册数据样本, 供 ICANN 以及其授权的第三方研究机构开展关于域名注册相关调查研究.

## 3 互联网标识隐私保护技术框架

针对第 2 节中讨论的域名注册业务流程, 结合数据隐私保护的生命周期, 具体分析在 GDPR 和《网络安全法》隐私保护新要求下, 业务流程中的相关隐私风险, 同时, 考虑数据在产生、存储、流通、分析挖掘的整个生命周期中, 如何保护用户隐私不被泄露、如何保证数据的可用性. 域名注册数据隐私保护生命周期模型如图 2 所示.

(1) 数据发布

数据发布者即采集数据和发布数据的实体, 包括域名注册局、注册商, 负责采集相关域名注册数据. ICANN 实施 WHOIS 政策, 域名 WHOIS 资料的收集、展示, 以及 ICANN 方都可能产生隐私泄露. 因此域名注册局、域名注册商、注册人在提供 WHOIS 信息、使用 WHOIS 信息上要注重隐私保护. 如何在数据发布时不泄露用户隐私内容, 同时还能保证数据的可用性, 是这一阶段的研究重点.

针对数据的匿名发布技术, 包括 *K*-匿名, *L*-多样性, *T*-接近性匿名等模型, 可以实现对数据发布时的隐私保护.

(2) 数据存储

数据存储方面隐私风险主要指在存储平台中, 数据被不可信的第三方偷窥或篡改的风险. 在域名业务流程中, 在注册局、注册商、ICANN 等域名管理部门以及数据托管商和 TMCH, 内部存储如何保证, 用户存储在系统中的高隐私等级数据不被窃取或篡改, 是数据存储阶段隐私保护的重点. 密码技术方法是解决该方法的关键.

(3) 分析挖掘

数据挖掘者试图从获取的数据中挖掘尽可能多的有价值信息, 但这可能会泄露用户的隐私信息. 经简单匿名技术处理的信息, 经过数据关联分析、聚类、分类等挖掘后, 仍可能分析出用户的隐私信息. 如在域名业务场景中, ICANN 或其他域名服务机构, 会委托第三方调研公司, 开展域名相关调查研究. 如何保证数据的可用性、足够研究机构进行调查研究, 又同时防范数据挖掘方法引起的隐私泄露, 是分析发掘阶段的主要隐私风险点.

抑制技术、假名化技术、泛化技术、随机化技术等传统技术可以解决这一阶段的问题.

同时, 基于统计基础的严格和可证明的差分隐私模型, 可以向第三方机构提供查询数据库, 保证隐私数据挖掘和隐私查询.

(4) 数据使用

数据使用者是访问和使用域名数据从数据中挖掘出信息的用户, 通常是企业和个人, 如何确保数据及属性在合适的时间和地点, 给合适的用户访问和利用, 是数据使用阶段面临的主要风险. 角色控制、访问控制等, 是这一阶段的主要解决方案.

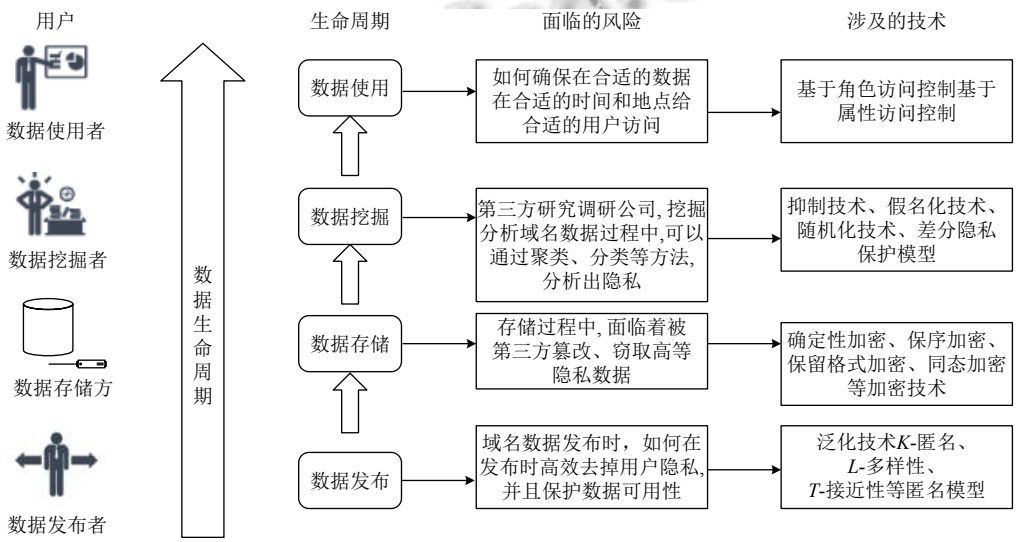



图 2 域名注册数据隐私保护生命周期模型



© 中国科学院软件研究所 http://www.c-s-a.org.cn

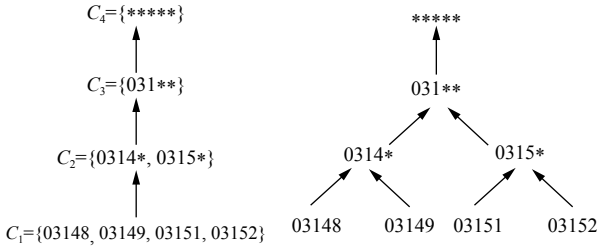
## 4 互联网标识隐私保护技术

### 4.1 数据发布隐私保护技术

注册局、注册商对收集到的域名用户注册数据进行公开发布时, 这些注册数据通常包含注册人的个人信息, 注册局、注册商需要在发布之前对数据进行处理, 防止不必要的用户个人信息泄露. 同时, 考虑用户数据被恶意第三方获取的极端情况, 希望攻击者无法从数据中识别出注册人确切个体数据信息, 匿名技术是实现上述目的的方法之一. 本节重点介绍传统的匿名操作—泛化、典型的匿名隐私保护模型—*K*-匿名模型, 以及 *K*-匿名的扩展改进模型.

(1) 泛化技术

泛化技术[13]是一种能够保护记录级数据的真实性, 同时降低数据集中所选属性粒度的匿名技术, 基本思想是用粗粒度的值代替原始细粒度的属性值, 从而减少属性的唯一值, 增加了推测出数据主体的难度. 泛化技术依据泛化层次树进行泛化, 主要包括域泛化和值泛化两种方式, 如图 3 所示.




(a) DGH 域泛化 (b) VGH 值泛化

图 3 域泛化和值泛化

(2) *K*-匿名模型

传统的匿名方法仅仅在数据表中泛化或者删除有关数据主体身份属性, 但这会遭到链式攻击[14], 攻击者可以通过其他包含数据个体标识符的数据集, 重新建立用户标识符与数据之间的对应关系, 推理得出隐私数据, 从而造成隐私泄露. 为了解决链式攻击问题, 1998 年 Samarati P 和 Sweeney L 提出了 *K*-匿名模型[15], 该模型要求发布的数据中, 指定标识符属性值相同的每一等价类至少包含 *K* 个记录, 使攻击者不能识别出数据主体的具体信息, 从而降低链式攻击所导致的隐私泄露风险. 一般 *K* 值越大, 隐私保护强度越大, 但信息损失也越大.

该模型可以保证攻击者无法确切指定某个数据主体是否在公开的数据库中; 给定一个数据主体, 攻击者无法确定其是否含有某项敏感属性; 同时也无法将某条数据记录对应到具体数据个体.

但是在实际应用中, 由于攻击者的背景不同, 攻击手段也多种多样, *K*-匿名模型仍会遭到同质攻击 (homogeneity attack) 和背景知识攻击 (background knowledge attack), 不能解决所有隐私泄露问题.

(3) *K*-匿名改进模型

针对 *K*-匿名模型的缺陷, 为了更好地保护数据表中的敏感信息, Machanavajjhala 和 Gehrke 等人提出了 *L*-多样性模型 (*L*-diversity)[16], 该模型要求每一个等价类在每一个敏感属性上至少包含 *L* 个不同值, 使得攻击者最多以 1/*L* 的概率确认每个个体的敏感信息. *T*-接近性 (*T*-closeness)[17]模型在 *L*-多样性模型的基础上, 考虑了敏感属性的分布问题, 要求所有等价类中敏感属性值的分布尽可能接近原始数据集的数据分布. 为了防止概率性推导, 要求任何定价类中的敏感属性的分布于整个数据集中相应属性的分布之间的距离小于阈值 *T*.

### 4.2 数据存储隐私保护技术

数据存储隐私保护是指在数据层层面的个人信息安全. 通信中可以使用 SSL 协议保证数据传输的安全, 因此, 数据层的数据保护主要是针对数据存储和管理的保护, 保证数据的机密性和完整性, 加密技术是解决这一问题的关键.

同态加密[18]是指对密文进行处理得到的结果仍然是加密的结果, 即对密文进行直接处理, 与对明文进行处理后再对处理结果加密, 得到的结果相同. 从抽象代数的角度讲, 保持了同态性. 文献[19,20]利用同态加密技术分别提出了 key-value 隐私存储方式以及多级索引技术, 能够保证数据拥有者和存储平台都不能在用户的节点检索过程中识别出节点.

保留格式加密[21]可以实现明文和密文的格式相同, 有助于增强数据库和数据仓库的安全性, 但是对于数据库敏感数据的保留格式加密, 需要保证密文满足数据库对于格式的约束.

安全多方计算[22]是另外一种数据加密技术, 其核心操作在分布式环境下基于多方参与者提供的数据计算出相应的函数值, 并确保除了参与者的输入及输出信息外, 不会额外暴露参与者的任何信息.

### 4.3 数据分析挖掘隐私保护技术

随着技术的进步, 数据挖掘可以从大量域名注册数据中挖掘出有价值的信息, 但也伴随着隐私泄露的



© 中国科学院软件研究所 http://www.c-s-a.org.cn

风险, 这一课题已经成为研究界的研究热点. 隐私保护数据挖掘, 即在保护隐私的前提下进行数据挖掘. 主要有两个研究方向:

(1) 对原始数据及进行必要的修改, 使得数据接收者不能侵犯他人隐私.

(2) 对数据分析查询、挖掘算法进行研究, 研究如何在挖掘过程中进行隐私保护.

针对第一个研究方向, 方法众多, 主要有抑制技术、假名化技术、随机化等典型代表技术. 针对第二个研究方向, 基于统计基础的严格可证明的差分隐私模型[23], 能够实现隐私查询, 可以确保在数据集中删除或插入一条记录, 对计算结果的影响非常小, 即使攻击者具有所有背景知识, 仍然无法获知某条个人记录.

### 4.4 数据使用隐私保护技术

数据使用者是访问和使用数据从数据中挖掘出信息的用户, 通常是企业和个人, 如何确保数据及属性在合适的时间和地点, 被合适的用户访问和利用, 是数据使用阶段面临的主要风险. 角色控制、访问控制等, 是这一阶段的主要解决方案.

在基于角色的访问控制 (Role-Based Access Control, RBAC)[24]中, 不同角色的访问控制权不同. 通过为用户分配角色, 可实现在对数据的访问权限控制. 因此, 在基于角色的访问控制中, 角色挖掘是前提. 通常, 角色根据职权、责任、工作能力而定.

RBAC 模型中引入了角色 (role) 的概念, 目的是为了隔离动作主体 (user) 和权限, 当一个角色被指定给了一个用户时, 该用户就拥有了该角色所包含的权限. RBAC 基本模型 (RBAC0) 包含了 RBAC 标准最基本的内容, 如图 4 所示.

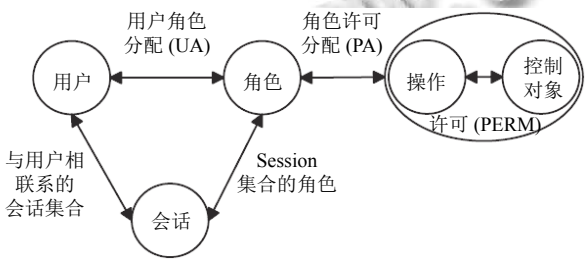



图 4 RBAC 模型核心

基于属性的访问控制 (Attribute-Based Access Control, ABAC)[25]通过将各类属性, 包括用户属性、资源属性、环境属性等组合起来用于用户访问权限的设定. RBAC 模型以用户为中心, 而没有将额外的资源信息, 如用户和资源之间的关系、资源随时间的动态变化、用户对资源的请求动作 (如删除、编辑等) 以及环境的上下文信息进行综合考虑. 而 ABAC 模型通过对全方位属性的考虑, 可以实现更加细粒度的访问控制. ABAC 框架示意图如图 5 所示.

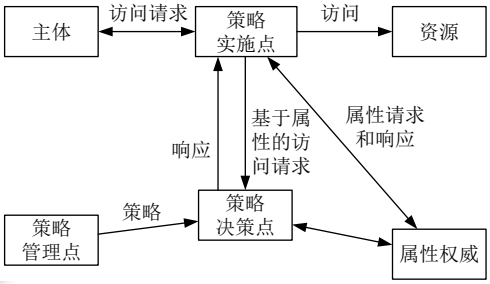



图 5 ABAC 框架示意图

## 5 实验分析

本节以数据发布阶段的隐私保护方案做为实验验证, 针对“.cn”域名注册数据中数值文本数据, 提出具体的方案流程, 如图 6 所示, 并对于可用性和隐私性进行对比分析.

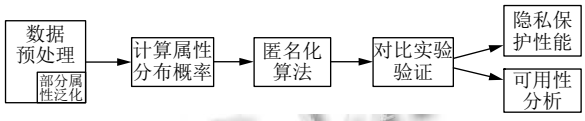



图 6 数值文本隐私保护方案流程

首先对数据需要进行预处理, 按需求对部分属性进行泛化、删除缺省数据、非法值.

然后计算属性的概率分布, 部分属性统计表 1 所示, 用于匿名化算法的实现, 以及后续对比实验.

表 1 原始表属性概率分布统计

| $Q_4$ 注册商取值 | 中国万网 | 35 互联 | 商务中国 | 时代互联 | 新网 |
|---|---|---|---|---|---|
| $Dist_4$ | 0.3 | 0.2 | 0.1 | 0.2 | 0.2 |
| $Q_5$ 年龄取值 (岁) | 20–30 | 30–40 | 40–50 | 50–60 | 60–70 |
| $Dist_5$ | 0.1 | 0.4 | 0.3 | 0.1 | 0.1 |
| $Q_6$ 地址取值 | 辽宁 | 河南 | 山东 | 四川 | 福建 |
| $Dist_6$ | 0.3 | 0.1 | 0.3 | 0.1 | 0.2 |

采用匿名化算法对数据表进行匿名处理, 可以采用 $K$-匿名、$L$-多样性算法以及其改进算法.

最后对于匿名化算法的隐私保护性能和可用性进行分析. 本文结合“.cn”域名注册数据, 实现了一种随机



© 中国科学院软件研究所 http://www.c-s-a.org.cn

可逆匿名化算法, 依据原始属性值概率分布, 随机替换需要匿名化的对象值, 具体算法如下:

算法. 随机可逆匿名算法

1. 输入: 原始数据集 $D$, 准标识符属性集合 $Q$, 准标识符属性被选概率 $p_i$,$i\in[1,n]$, $n$ 为准标识符属性个数
2. 输出: 随机匿名后的数据集 $D'$.
3. begin
4. $k=|D|$
5. $Dist_i=\emptyset$
6. for $i$=1 to $n$ do
7. begin
8. $Dist_i=Q_i$ 的概率分布
9. end
10. for $j$=1 to $k$ do
11. begin
12. 对于记录 $j$, 以概率 $p_u$ 随机从准标识符属性集合中选取一个属性 $Q_u$
13. 根据概率分布 $Dist_u$, 随机生成一个新的值
14. 用新值替换原属性值
记录替换关联, 以备可逆还原
13. end
14. end

数据原始表和匿名后的数据表如表 2、表 3 所示.

表 2 预处理后的数据原始表

| ID | 准标识符 QI | | | | | | 敏感属性 S | |
|---|---|---|---|---|---|---|---|---|
| | 注册日期 | 国籍 | 邮编 | 注册商 | 年龄 | 地址 | 手机号 | 邮箱 |
| 1357144 | [2017/11/01–2017/11/30] | 156 | 114000 | 中国万网 | [40–50] | 辽宁鞍山 | 156****2911 | N**@163.com |
| 1973789 | [2017/06/01–2017/06/30] | 156 | 455000 | 35 互联 | [30–40] | 河南安阳 | 180****2629 | se***@126.com |
| 3017568 | [2017/10/01–2017/10/31] | 156 | 118300 | 新网 | [50–60] | 辽宁丹东 | 197****2629 | l***@yahoo.com |
| 2975416 | [2017/05/01–2017/05/31] | 156 | 264200 | 商务中国 | [20–0] | 山东威海 | 180****4975 | a*@hotmail.com |
| 4497736 | [2017/11/01–2017/11/30] | 156 | 276800 | 时代互联 | [30–40] | 山东日照 | 187****2628 | m***@cnic.cn |
| 4195213 | [2017/05/01–2017/05/31] | 156 | 266000 | 中国万网 | [40–50] | 山东青岛 | 189****7263 | me***@163.com |
| 2725274 | [2017/11/01–2017/11/30] | 156 | 614100 | 新网 | [60–70] | 四川乐山 | 152****9455 | Ir***l@163.com |
| 5801711 | [2017/11/01–2017/11/30] | 156 | 350200 | 时代互联 | [30–40] | 福建福州 | 178****6424 | l***@yahoo.com |
| 5631342 | [2017/05/01–2017/05/31] | 156 | 122100 | 35 互联 | [40–50] | 辽宁锦州 | 185****4635 | n*@126.com |
| 4364300 | [2017/10/01–2017/10/31] | 156 | 366300 | 中国万网 | [30–40] | 福建龙岩 | 158****6794 | r***@yahoo.com |

表 3 匿名化后的数据表

| ID | 准标识符 QI | | | | | | 敏感属性 S | |
|---|---|---|---|---|---|---|---|---|
| | 注册日期 | 国籍 | 邮编 | 注册商 | 年龄 | 地址 | 手机号 | 邮箱 |
| 1357144 | [2017/11/01–2017/11/30] | 156 | 114000 | 中国万网 | [40–50] | 辽宁鞍山 | 156****2911 | N**@163.com |
| 1973789 | [2017/06/01–2017/06/30] | 156 | 455000 | 新网 | [30–40] | 河南安阳 | 180****2629 | se***@126.com |
| 3017568 | [2017/10/01–2017/10/31] | 156 | 614110 | 35 互联 | [50–60] | 辽宁丹东 | 197****2629 | l***@yahoo.com |
| 2975416 | [2017/05/01–2017/05/31] | 156 | 264200 | 商务中国 | [30–40] | 山东威海 | 180****4975 | a*@hotmail.com |
| 4497736 | [2017/11/01–2017/11/30] | 156 | 276800 | 时代互联 | [20–30] | 山东日照 | 187****2628 | m***@cnic.cn |
| 4195213 | [2017/05/01–2017/05/31] | 156 | 266000 | 新网 | [40–50] | 山东青岛 | 189****7263 | me***@163.com |
| 2725274 | [2017/11/01–2017/11/30] | 156 | 118300 | 中国万网 | [60–70] | 福建龙岩 | 152****9455 | Ir***@163.com |
| 5801711 | [2017/05/01–2017/05/31] | 156 | 350200 | 中国万网 | [30–40] | 福建福州 | 178****6424 | l***@yahoo.com |
| 5631342 | [2017/11/01–2017/11/30] | 156 | 366300 | 35 互联 | [40–50] | 辽宁锦州 | 185****4635 | n*@126.com |
| 4364300 | [2017/10/01–2017/10/31] | 156 | 122100 | 时代互联 | [30–40] | 四川乐山 | 158****6794 | r***@yahoo.com |

采用隐私保护评价的重要指标——数据查询准确率作为评价指标, 以相对误差进行横向对比. 采用 Aprioi 方法进行关联恢复, 验证匿名数据可用性.

查询方式为模糊查询:

$$\sum_{m}\left(*p(query_s)\prod_{i=1}^{n}\frac{*p(query_i\cap V(Q_i))}{*p(V(Q_i))}\right)$$

其中, $*p(a)$ 代表 $a$ 在准标识符敏感属性分组中出现的次数, $V(Q_i)$ 代表该准标识符属性可能的取值.

相对误差准确率计算方式为:

$$Error=\frac{\left|R(q,DB)-R\left(q,DB'\right)\right|}{R(q,DB)}$$

实验结果图 7 表明, 采用随机可逆匿名化算法的数据查询准确率相对误差远低于其他算法, 证明数据匿名性较好. 图 8 表明, 大部分准敏感关联规则得到了保留, 而其他 3 种方法的关联大部分被丢失, 说明采用随机可逆匿名化算法的匿名后的数据可用性较好.



© 中国科学院软件研究所 http://www.c-s-a.org.cn

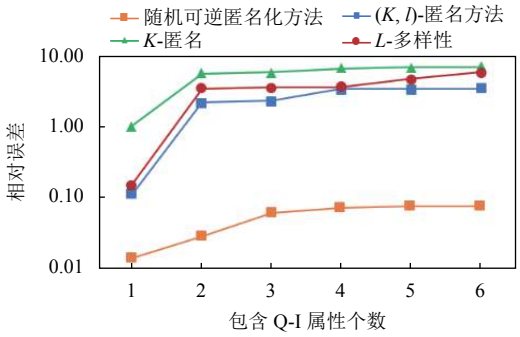



图 7 数据查询准确率

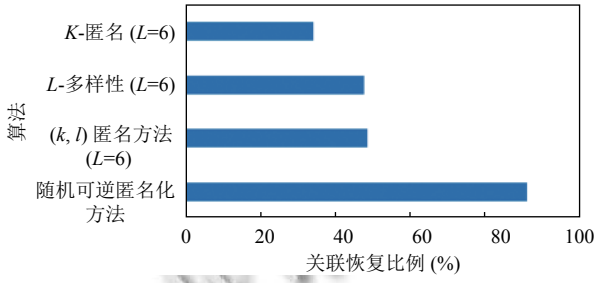



图 8 关联恢复比例

## 6 总结与展望

欧盟 GDPR 和国内网安法对个人信息和隐私安全保护提出了更严格的要求. 针对互联网标识中涉及的隐私数据泄露的问题, 本文以最典型的互联网标识——域名为例进行深入讨论. 首先分析域名业务流程中涉及隐私泄露的风险点, 针对隐私泄露风险点, 结合个人数据生命周期, 提出了互联网标识隐私保护框架. 本文以域名业务场景为例, 但是问题的分析思路、隐私保护框架以及技术解决方案具有普适性, 仍适用于其他互联网标识的隐私保护分析.

隐私保护是目前信息安全领域的研究热点之一, 也取得了丰富的研究成果. 近 2 年来, 随着各国加强立法, 注重隐私保护, 其技术研究也出现了新的深度. 但是从实际应用角度来看, 还有很多内容需要深入研究, 本文从立法和技术以及行业流程的不同角度回答了互联网标识隐私保护所面临的一些挑战的解决方案, 希望能够给后续的研究提供一些参考.

## 参考文献


1 孙昌璐. 工业互联网标识管理与解析技术. 信息通信, 2017, (9): 161–162. [doi: 10.3969/j.issn.1673-1131.2017.09.078]

2 Gassmann HP. OECD guidelines governing the protection of privacy and transborder flows of personal data. Computer Networks (1976), 1981, 5(2): 127–141. [doi: 10.1016/0376-5075(81)90068-4]

3 Landau S. Control use of data to protect privacy. Science, 2015, 347(6221): 504–506. [doi: 10.1126/science.aaa4961]

4 EU Directive. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Communities, L281, 1995. 31–50.

5 Regulation GDP. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Official Journal of the European Union (OJ), L119, 2016. 294.

6 Voigt P, Von Dem Bussche A. The EU general data protection regulation (GDPR). A practical guide. Cham: Springer, 2017.

7 O'Connell, Emmet J. Privacy in America: The traditions, changing views, and response. Senior Projects Spring 2018. 348. https://digitalcommons.bard.edu/senproj_s2018/348.

8 Prosch M. Protecting personal information using Generally Accepted Privacy Principles (GAPP) and continuous control monitoring to enhance corporate governance. International Journal of Disclosure and Governance, 2008, 5(2): 153–166. [doi: 10.1057/jdg.2008.7]

9 中华人民共和国网络安全法. 新疆农垦科技, 2017, 40(1): 80−82.

10 Banisar D. National Comprehensive Data Protection/Privacy Laws and Bills 2018. Privacy Laws and Bills, 2019.

11 杨学莲. 域名注册管理相关法律问题研究[硕士学位论文]. 济南: 山东大学, 2010.

12 中国互联网络域名管理办法. 信息技术与标准化, 2005, (1–2): 4–6.

13 Samarati P. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010–1027. [doi: 10.1109/69.971193]

14 Sweeney L. Computational disclosure control: A primer on data privacy protection [Ph. D. thesis]. Cambridge, Massachusetts: Massachusetts Institute of Technology, 2001.

15 Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557–570. [doi: 10.1142/S0218488502001648]

16 Machanavajjhala A, Gehrke J, Kifer D, *et al*. L-diversity:




© 中国科学院软件研究所 http://www.c-s-a.org.cn

Privacy beyond k-anonymity. Proceedings of the 22nd International Conference on Data Engineering. Atlanta, GA, USA. 2006. 24.

17 Li NH, Li TC, Venkatasubramanian S, Venkatasubramanian S. *T*-closeness: Privacy beyond *K*-anonymity and *L*-diversity. Proceedings of 2007 IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey. 2007. 106−115.

18 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战. 计算机研究与发展, 2013, 50(1): 146–169. [doi: 10.7544/issn1000-1239.2013.20121130]

19 Chen X, Huang QM. The data protection of MapReduce using homomorphic encryption. Proceedings of 2013 IEEE 4th International Conference on Software Engineering and Service Science. Beijing, China. 2013. 419–421.

20 Brakerski Z, Vaikuntanathan V. Efficient fully homomorphic encryption from (Standard) LWE. Proceedings of 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science. Palm Springs, CA, USA. 2011. 97–106.

21 Black J, Rogaway P. Ciphers with arbitrary finite domains. Proceedings of Cryptographers' Track at the RSA Conference. San Jose, CA, USA. 2002. 114–130.

22 Goldreich O. Secure multi-party computation. Manuscript. Preliminary Version, 1998, 78.

23 Dwork C. Differential privacy. In: Bugliesi M, Preneel B, Sassone V, *et al*, eds. Automata, Languages and Programming. Berlin, Heidelberg: Springer, 2006. 1–12.

24 Sandhu RS, Coyne EJ, Feinstein HL, *et al*. Role-based access control models. Computer, 1996, 29(2): 38–47. [doi: 10.1109/2.485845]

25 Hu VC, Kuhn DR, Ferraiolo DF, *et al*. Attribute-based access control. Computer, 2015, 48(2): 85–88. [doi: 10.1109/MC.2015.33]

www.c-s-a.org.cn



© 中国科学院软件研究所 http://www.c-s-a.org.cn