中国科学: 信息科学 2021 年 第 51 卷 第 7 期: 1100–1115
SCIENTIA SINICA Informationis
论文

基于权限的移动应用程序隐私风险量化

朱敏杰, 叶青青, 孟小峰*, 杨鑫

中国人民大学信息学院, 北京 100872
* 通信作者. E-mail: xfmeng@ruc.edu.cn

收稿日期: 2020–03–02; 修回日期: 2020–06–05; 接受日期: 2020–07–13; 网络出版日期: 2021–06–07

国家自然科学基金项目 (批准号: 61941121, 91846204) 资助

摘要 移动设备的普及带来了移动应用程序市场的蓬勃发展, 各类服务提供商通过移动应用程序的权限大量收集用户数据, 而数据收集过程往往不为用户所知, 因此给用户带来极大的隐私风险. 对移动应用程序进行隐私风险评估, 不仅有助于规范第三方移动应用市场, 而且可帮助用户规避潜在的隐私风险, 而如何评估移动应用程序可能带来的最大隐私风险则是当前面临的重大挑战. 本文通过研究移动应用程序最大化的数据泄露场景, 基于权限请求特征和权限分析原则构建隐私风险最大值量化模型. 该模型基于权限敏感度、权限类别异常度、权限使用率和权限调用者数量 4 个参数, 对移动应用程序的潜在隐私风险进行评估. 在隐私风险量化和恶意应用检测中, 对比当前同类型方法, 该模型在真实数据集上效果均较优, 说明模型的有效性. 实验结果进一步表明, 该模型可用于改善现有第三方移动应用市场的隐私风险预警机制, 进而保护移动用户的隐私.

关键词 隐私保护, 移动应用程序, 隐私风险量化, 权限分析方法

1 引言

随着近些年软硬件技术迅猛发展, 移动设备大量普及. 截至 2020 年 3 月, 中国手机网民规模达 9.04 亿, 网民通过手机接入互联网的比例高达 99.3%[1)]. 移动浪潮进一步密切了人与网络的关系, 但也产生了严重的用户隐私问题. 其中, 移动应用程序 (mobile application) 隐私问题尤为显著. App (application) 类型广泛, 包括视频、社交、购物等, 覆盖了用户方方面面的个人信息. 若这些个人信息被恶意第三方获取, 用户隐私将受到严重侵害. 以安卓系统为例, 许多手机应用在安装时默认获取地理位置、通讯录等授权, 以此来收集用户的个人信息. 如图 1 的移动设备用户个人数据收集流程所示, 当用户使用某 App 时, 用户的个人数据会被开发者和内嵌服务第三方 (即软件工具开发包 SDK, software

1) http://www.cac.gov.cn/2020-04/27/c_1589535470378587.htm.

引用格式: 朱敏杰, 叶青青, 孟小峰, 等. 基于权限的移动应用程序隐私风险量化. 中国科学: 信息科学, 2021, 51: 1100–1115, doi: 10.1360/SSI-2020-0039
Zhu M J, Ye Q Q, Meng X F, et al. Privacy risk quantification of mobile application based on requested permissions (in Chinese). Sci Sin Inform, 2021, 51: 1100–1115, doi: 10.1360/SSI-2020-0039

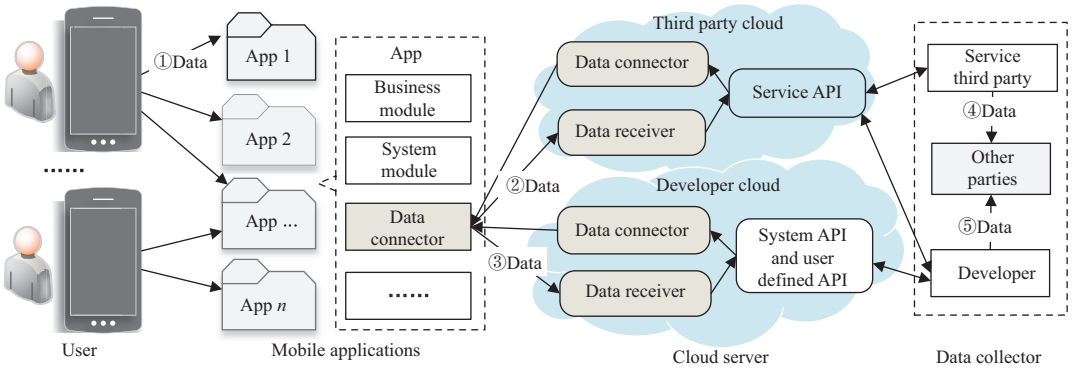

图 1 安卓系统移动设备用户个人数据被收集流程

Figure 1 Individual data collection of Android devices

development kit, 如广告、推送、分类统计类等) 同时收集, 如步骤 ②③ 所示. 此外, 当开发者和服务第三方不可信时, 它们把收集到的用户数据进行分享或交易买卖, 如步骤 ④⑤ 所示, 从而产生巨大的用户隐私风险.

目前国内外已普遍认识到数据隐私问题的严重性, 并制定了相关的法律法规. 在全球隐私保护浪潮下, 如何分析 App 带给用户的隐私风险, 并进行隐私预警和保护是一个非常迫切且具有重要意义的挑战性问题. 目前隐私研究主要关注隐私保护方法的提出和改进, 针对隐私风险的量化评估研究相对较少, 尤其是移动应用领域. 作为主动式隐私保护框架的重要部分, 隐私主动监测和评估是整个保护框架的基础, 能够引导进行精准的隐私主动管理 [1]. 基于此, 本文重点关注移动应用程序隐私风险的分析量化, 其难点主要包括: (1) 如何识别并分析 App 内的数据收集情况; (2) 如何评估 App 因数据收集而导致的用户隐私风险; (3) 如何简单高效地保护用户隐私.

Demetriou 等 [2] 指出移动应用程序数据泄露评估不应只考虑已发生的事件, 应考虑未来潜在可泄露的最大数据量. 因此本文重点研究分析移动应用程序当前及潜在所能泄露的数据最大范围, 进而量化评估其给用户带来的最大隐私危害程度. 针对难点 (1), 本文整理对比 5 类 App 数据泄露分析方法, 对比后选取权限分析方法来识别潜在可泄露的用户数据. 针对难点 (2), 现有研究对 App 隐私风险量化研究尚有不足: (i) 大部分风险评估研究将安全风险与隐私风险混为一谈 [3,4]; (ii) 基于权限的风险量化方法中存在对权限特征考虑不全的情况 [5~7]; (iii) 风险定义不准确, 认为权限请求数量较少的 App 较为异常 [8]. 针对难点 (3), 基于所提隐私风险量化模型, 建议显式告知用户其待安装或使用中的 App 风险大小, 让用户自主决定是否在其隐私风险接受范围内.

权限是 App 访问数据或资源的请求授权, 本文通过观察权限请求特征确定数据收集情况, 总结得到 6 条权限分析原则, 进而提出基于权限的分类别移动应用程序隐私风险量化模型. 与同类型的隐私风险量化和恶意应用检测方法相比, 本文方法效果更优, 表明了模型的有效性. 本文的主要贡献如下:

(1) 改进基于权限的移动应用程序异常度算法.

(2) 提出分类别移动应用程序隐私风险量化模型, 并在真实数据集上验证了其有效性和可用性.

(3) 验证了目前第三方移动应用商店的隐私风险预警机制的不足, 并提出改善方法.

2 相关工作

2.1 移动应用程序数据泄露分析

移动设备及其安装的 App 包含大量数据和系统资源, 如账户、通讯录、短信等, 这些数据因能被 App 开发者及内嵌服务第三方获取而存在各类风险. 分析 App 数据泄露情况, 主要有以下 5 类方法.

权限分析. 检查 App 下载前和使用中的权限请求来分析其所获取数据. 该方法基于移动设备操作系统提供的权限请求体系, 以结构化的权限请求为分析对象, 简单清晰, 可快速高效地检查移动应用程序的数据收集 [9,10]. 但是该方法只能从较粗粒度上分析 App 可能收集的数据, 无法探测使用过程中实际的数据传输, 常用于研究权限使用特征规律 [11~13].

隐私政策分析. 检查 App 提供的隐私政策中的数据管理规定, 并分析其是否可信 [14,15]. 该方法通过分析数据相关的自然语言实体来识别 App 收集的数据, 因此其依赖于自然语言处理技术, 且无法探测实际的信息传输, 适用于第三方移动应用商店对 App 进行审查. 如 Yu 等 [16] 通过设计 PPchecker 系统来自动化识别隐私政策不完备、不正确、不一致 3 类问题.

静态代码分析. 基于程序代码在更细粒度上分析 App 的数据请求, 分析对象主要为 API 函数, 该方法弥补了权限分析无法探测实际数据传输的缺点, 现有工作通过该方法分析 API 函数与其他数据收集行为间的映射关系 [10,17], 研究应用本身的恶意行为或由数据导致的隐私泄露风险 [18,19].

动态分析. 捕捉并分析 App 使用过程中的网络信息传输, 具体方法包括动态信息流追踪、虚拟个人网络、中间人攻击等 [20~22]. 已有研究中实现系统包括手机污点追踪系统 TaintDroid [23]、基于个人虚拟网络方法设计的 AntMonitor 系统 [24] 和 TaintMan [25] 等.

混合分析. 把以上 4 种分析方法结合起来, 从不同粒度和角度分析 App 数据传输情况, 不同类别的方法进行相互之间的补充和验证 [26]. 例如, 首先利用静态代码分析方法找到异常的代码块, 然后利用动态分析方法来进一步确认异常程度.

以上 5 种方法通过检验数据存在或流动的不同形式来识别数据泄露, 各具优缺点. 隐私政策分析依赖于自然语言技术, 结构性较弱且粒度也较粗; 权限分析依赖于结构性最强且易解析的的权限请求, 但识别出的数据粒度一般; 静态代码分析识别粒度较细, 但函数命名的不规则使得数据请求识别难度较高; 动态分析方法则要求人工介入来捕捉 App 与外界的网络信息传输, 数据识别粒度最为全面, 但其数据捕捉与实时的人工操作相关, 分析范围有限, 且加密的网络数据报文使其无法大规模应用. 本文分析研究 App 当前及潜在可能泄露的最大数据量, 与用户具体操作中实时泄露的特定数据无关, 因此动态及混合分析方法不适用. 考虑到本文针对大规模 App 进行分析, 不规则的静态代码分析不可行. 比较数据识别规则和粒度, 权限分析皆优于隐私政策分析, 因此本文采用权限分析方法.

2.2 移动应用程序隐私风险评估

信息安全风险评估指识别风险源, 并基于其发生可能性和发生后产生危害对其风险程度进行评估 [27]. 同理, 隐私风险评估指识别用户数据泄露源, 基于其发生可能性及发生后对用户隐私产生危害评估其风险程度. 隐私风险评估不同于隐私泄露概念, 前者评估潜在的隐私泄露风险, 该事件尚未发生, 因此需要对风险上下界进行估计, 而后者指数据泄露事件已经发生, 并导致隐私泄露问题.

针对 App 隐私风险, 因权限与用户数据间的映射关系, 基于权限分析的评估方法是目前的主流方向, 主要考虑 3 个因素: 权限选取、参数定义和评估模型. Sarma 等 [5] 最早提出的 CRCP (category-based rare critical permission) 风险信号方法基于权限使用率和触发机制来刻画 App 风险, 设定权限使用率的最小阈值. Peng 等 [3] 随后也利用权限的使用率特征提出概率生成模型计算 App 产生概率,

并基于该概率与其风险值的反比关系得到风险值. 该模型把权限当作独立的伯努利变量, 并引入权限敏感度分级概念, 即本文 4.1 小节定义的 “权限敏感度” 参数. 但该模型得到的风险值并不限于隐私风险, 且模型的单调性会因使用率大于 50% 的权限而受到破坏. DroidRisk 方法 [4] 彻底打破了囿于权限使用率评估 App 风险值的限制, 使用权限的恶意应用发生概率和敏感度计算恶意应用风险. Liccardi 等 [6] 提出的 SensitivityScore 方法则简单把各敏感权限的敏感度视为 1, 把敏感权限个数视为敏感度分值. Mylonas 等 [28] 则提出使用 EBIOS 方法 [29] 分析权限敏感度. Quattrone 等 [8] 基于权限特征使用孤立森林 (isolation forest, iForest) 算法 [30] 对 App 权限异常程度进行量化, 以权限异常程度表示其隐私风险. 但该方法未考虑实际的数量分布趋势, 认为请求权限数量较少的移动应用程序具有较高的异常值. Hamed 等 [7] 则综合考虑权限及权限对的使用率及敏感度, 但是该方法并未考虑 App 功能类别不同导致的权限差异. 此外, 部分现有工作考虑权限调用差异, 进一步对可调用权限的 App 内嵌服务第三方所带来的隐私风险进行研究和分析 [2, 31~33].

3 数据泄露及权限分析

基于权限分析方法, 本节着重分析权限与隐私数据间的映射关系, 并研究基于权限的数据泄露特征. 最后, 基于权限请求特征和相关文献, 归纳得到 6 条权限分析原则.

3.1 权限到隐私数据映射关系

隐私范畴难以界定, 个人因主观隐私意识不同, 其对信息的隐私性认知也不同.《信息安全技术个人信息安全规范》[2)] 中把自然人的隐私信息描述为 “一旦泄露、非法提供或滥用可能危害人身和财产安全, 极易导致个人名誉、身心健康受到损害或歧视性待遇等的个人信息”. 具体来讲, 互联网环境下的个人隐私数据主要分为信息隐私、位置隐私、设备隐私、通信隐私、网络行为隐私、社会行为隐私, 详见补充材料表 S1 个人隐私范畴分类表. 分析移动应用程序泄露的个人隐私数据, 其涵盖除社会行为隐私外的另 5 大类隐私, 如属于信息隐私范畴的个人账号、属于位置隐私的 GPS 数据、属于设备隐私的任务进程、属于通信隐私的短信息、属于网络行为隐私的操作日志.

权限是 App 请求系统资源或数据的请求授权, 本文以个人隐私数据范畴内的权限为研究对象, 并将其定义为敏感权限. 安卓 6.0 以上版本[3)] 共有 144 个系统权限 (Manifest.permission, 2019), 本文选取 39 个隐私数据范畴内的敏感权限展开研究. 这些权限分别对应 18 类个人数据, 包括日历、图片、联系人、位置、音频、通话、短信、存储、网络、Wi-Fi、蓝牙、账户、日志、同步、近场通信、唤醒、电池、任务进程, 详情如补充材料表 S2 所示.

定义1 (敏感权限) 存在安全或隐私危害的权限统称为敏感权限. 敏感权限集合用 $\mathcal{P}$ 表示, P_i 表示第 i 个敏感权限.

3.2 基于权限的数据泄露特征

以本文真实移动应用程序数据集为例, 从两个维度来研究基于权限的数据泄露特征. (1) App 类别. 不同类别 App 因其固有功能的需要, 必须请求相应的系统权限. 相反, 若 App 请求超出其服务范围外的权限则被认为是不合理的, 甚至是有风险的, 例如手电筒类应用不应请求与联系人相关的权限. 参考谷歌应用商店的分类体系对市场应用类别信息进行整理, 共得到 21 个类别: 安全类、办公类、儿

2) http://pip.tc260.org.cn/assets/wz/2020-03-07/ef2dab88-cd9d-4748-814a-a3eca027beba.pdf.
3) https://developer.android.google.cn/reference/android/Manifest.permission?hl=en.

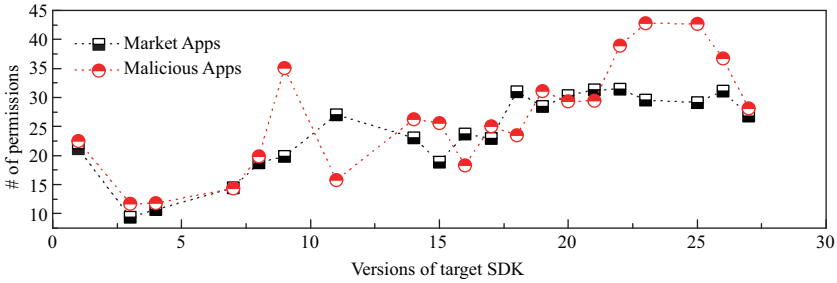

图 2 (网络版彩图) 恶意应用权限请求数量特征

Figure 2 (Color online) The number of requested permissions under different target SDK versions

童类、房产家居类、工具类、购物类、健康医疗类、教育类、理财类、旅游出行类、社交类、摄影图片类、生活助手类、视频类、通讯类、新闻类、音乐类、游戏类、娱乐类、阅读类、运动类. (2) 恶意应用程序, 主要是 “恶意收集用户信息” App. 正常应用和恶意应用在权限使用的种类、频次及组合上均有明显差异 [34], 恶意 App 通常因自身破坏功能的需要而具有较为异常的权限请求模式.

分析实验数据发现: (1) 各类别 App 数据收集特征. 各类别移动应用程序对其服务运行所需的基本权限请求比率均较高, 如网络、Wi-Fi 等权限. 不同类别移动应用程序间权限请求模式具有差异, 如通讯类应用多请求通讯录类和短信类权限, 运动类应用多请求位置、摄像机类权限. (2) 恶意应用数据收集特征. 如图 2 所示为不同目标 SDK 版本下恶意应用和市场应用的权限请求数量分布, 大部分版本中恶意应用的权限请求数量均明显大于或近似于市场应用.

3.3 权限分析原则

由隐私风险评估定义知, 隐私风险主要与 “数据泄露量” 和 “数据隐私危害程度” 两个因素相关. 基于此, 考虑权限数据泄露特征, 归纳相关文献得到 6 条权限分析原则. 其中前 5 条原则体现 “满足应用功能, 数据泄漏量最小” 思想, 第 6 条原则体现 “数据隐私程度不一” 思想.

(1) **最小权限 (数量) 原则.** App 在满足固有功能下应请求尽可能少的权限. 恶意或山寨应用往往请求权限数量较多, DroidRisk [4], Stowaway [10], Sensiticity_Score [35] 方法均采用这种思想.

(2) **特定权限或权限组合原则.** 该原则认为相同类别或主题 App 具有特定的权限或权限使用组合, 如摄影类应用会申请 CAMERA 权限. SOM [9], Kirin [36], K-Means [37] 等方法均验证了该原则.

(3) **权限与功能一致原则.** 鉴于不同功能需要特定的权限支持, 该原则认为权限请求应与 App 功能保持一致. CRCP 风险信号 [5] 和 PrivacyPalisade [8] 方法采用此原则.

(4) **权限与 API 调用一致原则.** 该原则认为权限请求应与 API 函数保持一致. 鉴于 API 函数运行须对应权限授权, Stowaway [10] 和 PScout [17] 方法均基于此对 API 和权限进行了映射分析.

(5) **权限与应用描述一致原则.** 该原则认为权限请求应与应用描述或隐私政策一致, 是权限与功能一致原则的扩展. 应用描述是 App 功能的载体, 其文字内容覆盖了具体的功能特性 [38].

(6) **权限风险分级原则.** 由于各权限映射用户数据不同, 因此具有不同的安全或风险级别. 在对权限风险进行分析或量化时, 不同权限应分级进行. DroidRisk [4] 和 EBIOS [29] 等评估方法均据此展开.

4 隐私风险量化模型

在移动应用程序数据泄露场景中, 已知敏感权限请求可代表用户泄露数据, 因此本节通过 App 请

求的敏感权限及权限被允许后对用户隐私产生的危害评估 App 具体隐私风险程度. 基于 “数据泄露量” 和 “数据隐私危害程度” 两类风险影响因素, 全面考虑权限的各类参数, 包括特定类别权限组合模式、隐私敏感程度、特定类别权限使用情况和权限调用者数量, 提出基于权限的分类别移动应用程序隐私风险量化模型. 目前隐私风险评估方法没有明确的评价指标, 基于 Peng 等 [3] 所提风险评估函数的 3 条要求, 本文提出适用于移动应用程序隐私风险量化的 3 条基本评估准则: (1) 同一类别下, App 隐私风险值与权限数量正相关. 请求的隐私权限越多表示其可获取并泄露的数据或系统资源越多, 即其隐私风险值越大. (2) 恶意应用隐私风险比正常应用高. 恶意应用通常具有较为异常的权限请求模式, 因获取数据或系统资源较异常而具有较高的隐私风险值. (3) 风险量化方法的可解释性.

对第 t 个移动应用程序 $A_t=(c_t,P_t)$ 进行隐私风险量化, 该移动应用程序所属类别 $c_t\in C$ ($|C|=21$), 权限请求列表 $P_t=[P_{t,1},P_{t,2},\ldots,P_{t,m}]$, ($m=39$, 即敏感权限数量), 第 i 个权限 $p_{t,i}=1$ ($i\in[1,m]$) 表示该移动应用程序请求了第 i 个权限. 本文把 A_t 的隐私风险程度量化为一个具体值 pr_{A_t}, 即隐私风险 (privacy risk, PR), 其基本模型可表示为

$$\mathrm{PR}_A(A_t)=\mathrm{PR}_A(c_t,P_t)=\mathrm{pr}_{A_t}, \tag{1}$$

其中 $\mathrm{pr}_{A_t}\in[0,1]$, 即隐私风险值最终归一化到 $[0,1]$ 区间. App 隐私风险程度越高, pr_{A_t} 值越大.

本文全面考虑特定类别权限组合模式及使用情况、权限的隐私敏感程度及调用者数量等影响因素, 以权限及其组合为对象, 分别定义权限敏感度、权限类别异常度、权限类别使用率和权限调用者数量参数构建 App 隐私风险量化评估模型, 具体构建过程分为以下 6 步.

第 1 步, 基于特定权限或权限组合原则, 考虑权限组合对用户隐私产生额外危害. 单个恶意行为有时需要同时申请多个权限, 比如完成 “发送短信” 行为必须请求联系人 “READ_CONTACT” 和短信 “SEND_SMS” 两个权限. 但鉴于 Ilyas 等 [39] 已证明在联合分布中两两关系最重要且近似于整体分布, 本文只考虑两个权限组合 (即权限对) 产生的隐私风险. 权限 $p_{t,i}$ 与权限 $p_{t,j}$ 组成权限对 p_{t,i_j}.

第 2 步, 基于权限风险分级原则, 考虑权限的隐私敏感度参数.

定义2 (权限敏感度) 某权限或权限对被允许后对用户隐私造成的危害大小.

这种隐私危害具体指对个人身份识别产生的正向影响, 即增加个人身份被识别的几率. 权限 p_i (权限对 p_{i_j}) 的权限敏感度记为 s_i (s_{i_j}). 权限的敏感度越高, 表示其泄露后对用户的隐私危害越大.

第 3 步, 基于权限与功能一致原则, 考虑移动应用程序特定功能类别下的权限特征.

定义3 (权限类别使用率) 某权限或权限对在某类别移动应用程序中被请求比率.

类别 c_t 中权限 $p_{t,i}$ (权限对 p_{t,i_j}) 的类别使用率记为 $r_{t,i}$ (r_{t,i_j}). 该参数考虑移动应用程序请求的某个权限在其类别下的特征. 权限类别使用率高, 表示该权限或权限对是其类别内基于共有功能而被普遍申请的, 隐私风险相对较小.

定义4 (权限类别异常度) 某移动应用程序请求的敏感权限组合相对于其所在类别特定请求模式的异常程度.

移动应用程序 A_t 权限类别异常度记为 O_t. 权限类别异常度越高, 表示该 App 请求的权限组合与其类别权限模式相差越大.

第 4 步, 基于最小权限 (权限数量) 原则, App 隐私风险来源于其请求的各权限 (权限对) 的隐私风险. 移动应用程序 A_t 所请求的单个权限 $p_{t,i}$ (权限对 p_{t,i_j}) 的隐私风险大小与其权限敏感度和类别使用率相关, 权限敏感度越高, 类别使用率越低, 其隐私风险值越大. 因此单个权限 $p_{t,i}$ (权限对 p_{t,i_j})

的隐私风险大小 $\mathrm{pr}_{p_{t,i}}$ ($\mathrm{pr}_{p_{t,i\_j}}$) 计算如下:

$$\mathrm{pr}_{p_{t,i}} = p_{t,i} \times s_{t,i} \times (1 - r_{t,i}), \tag{2}$$

$$\mathrm{pr}_{p_{t,i\_j}} = p_{t,i\_j} \times s_{t,i\_j} \times (1 - r_{t,i\_j}), \tag{3}$$

第 5 步, 基于最小权限 (权限数量) 原则, 对各权限 p_i (权限对 p_{i_j}) 的隐私风险进行累加, 得到各权限隐私风险总和. 已知数据泄露情况下, 数据泄露范围越大, 获取到用户数据的收集者越多, 产生的用户隐私危害越大. 由图 1 知应用本身开发者和内嵌的服务第三方均可请求权限来获取用户数据, 且服务第三方和应用本身具有同样的权限调用能力 [2], 潜在可获取同样的数据. 因此本文不区分应用本身和服务第三方的权限使用差异, 但通过定义权限调用者数量来考虑数据泄露范围带来的隐私危害.

定义5 (权限调用者数量)　可调用移动应用程序权限或权限对的应用开发者或服务第三方的数量.

移动应用程序 A_t 权限调用者数量记为 u_t. 由各权限隐私风险总和及权限调用者数量得到 A_t 的权限隐私风险值 pr_{p_t}, 如式 (4) 所示. 此外, 本文使用 sigmoid 函数对权限隐私风险值 pr_{p_t} 进行正则化处理, 使其值分布 $[0,1]$ 区间内, 如式 (5) 所示, 其中 $\omega = \sum_{t=1}^{n} \ln(\mathrm{pr}_{p_t})/n$ (n 为数据集中移动应用程序数量), 用来调节 sigmoid 函数的映射区间.

$$\mathrm{pr}_{p_t} = u_t \times \left(\sum_{i=1}^{m} \mathrm{pr}_{p_{t,i}} + \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathrm{pr}_{p_{t,i\_j}} \right), \tag{4}$$

$$\mathrm{pr}'_{p_t} = \frac{1}{1 + \mathrm{e}^{-\ln(\mathrm{pr}_{p_t}) + \omega}}, \tag{5}$$

第 6 步, 基于权限类别异常度与权限隐私风险值, 计算移动应用程序的隐私风险值. 权限类别异常度越高, 权限隐私风险值越大, 该移动应用程序的整体隐私风险越大. 本文假设两个因素影响权重相等, 令 $\alpha = 0.5$, 则可得出移动应用程序 A_t 的隐私计算模型, 如下所示:

$$\mathrm{pr}_{A_t} = \alpha \times O_t + (1 - \alpha) \times \mathrm{pr}'_{p_t} = \frac{O_t + \frac{1}{1 + \mathrm{e}^{-\ln(\mathrm{pr}_{p_t}) + \omega}}}{2}. \tag{6}$$

以下具体求解隐私风险量化模型中移动应用程序 $A_t = (c_t, p_t)$ 的各项权限参数: 权限敏感度 s_i (s_{i_j})、权限类别异常度 O_t、权限类别使用率 $r_{t,i}$ (r_{t,i_j}) 和权限调用者数量 u_t.

4.1 权限敏感度

权限敏感度主要有两类计算方法: (1) 将权限敏感度分两类: 敏感、不敏感. 默认敏感权限的敏感度为 1 [6] 或把其设为未知变量, 进行目标函数最优求解 [4]; (2) 将权限敏感度划分多个等级, 使用 EBIOS 方法评估 [7, 28]. 本文选择对权限敏感度进行多级区分, 因此采用法国数据保护机构国家信息与自由委员会制定的标准风险评估方法 EBIOS [29] 计算权限敏感度.

EBIOS 方法主要通过 3 个因素确定数据风险量级: 数据产生的隐私问题、可能引起的隐私风险及是否有解决方法. 数据产生的隐私问题越多, 可能引起的隐私风险越大, 解决办法越少, 则其隐私风险量级越高. 数据产生的隐私问题, 包括数据的可识别程度及危害程度. 本文基于 EBIOS 方法把单个权限敏感度划分为 4 个级别: 可忽略、有限、显著、严重, 分别对应权限敏感度分值为 1, 2, 3, 4. 划分权限的可识别程度及危害程度为同样分值的 4 个等级, 将两者分值相加转换为对应的权限敏感度, 如补充材料表 S3 所示. 由此本文对 39 个敏感权限的可识别程度和危害程度分别评估后得到最终的权限敏感度, 其具体量化如补充材料表 S4 所示. 对于权限对敏感度 s_{i_j}, 其值等于单个权限敏感度的加和.

4.2 权限类别异常度

已知各类别 App 具有特定的权限请求模式, 本小节计算 A_t 所请求权限 P_t 相对于类别 c_t 特定权限请求模式的异常程度. 权限请求特征 P_t 的异常程度可通过异常检测算法求解, 但有以下要点: (1) 权限请求列表为高维数据; (2) 无监督, 且数据集中存在部分异常点. Liu 等 [30] 提出的孤立森林方法是适用于高维数据的无监督的异常检测算法, 具有线性时间复杂度和较好的算法效果. 在一个高维数据集样本中, iForest 算法把异常点看作 "分布稀疏且与高密度群体距离较远" 的孤立点, 采用二分树 iTree 递归地随机选取特征对数据集进行切割, 直至每个数据点都被孤立, 最后使用数据点的切割路径长度 (即切分次数) 表示该点偏离整体数据集的程度. 高密度区域的数据点相比异常点需要更多次数的切割才能被孤立, 因此异常点的切割路径长度通常较小.

已有研究 [8] 在使用该方法计算权限类别异常度时, 主要存在两点弊端: (1) 以权限请求列表作为输入特征, 高维度使得异常检测算法准确率降低; (2) 权限请求数量小的少数 App 的异常度较高. 基于最小权限原则, 同类别下 App 的权限请求数量越少, 其异常水平应该越低. 基于图 3 的各类别 App 数量在各权限请求数量下的分布趋势可知, 权限请求数量较少与较多的 App 在整体中均占少数, 它们偏离整体数据集程度均较大, 因此直接使用该算法会导致权限请求数量小的 App 具有高异常程度.

为解决这两点问题, 本文提出适用于计算 App 权限类别异常度的算法 App_iForest. 该算法首先处理权限特征. (1) 特征降维, 以分组的权限数量替代整体权限列表作为特征. 基于权限请求的种类和数量, 通过累加并合并 18 组隐私数据内权限请求数量作为新的特征, 将特征维度降至 14. (2) 特征扰动, 以维度平均值替代较小值作为特征. 已知各维度中较大或较小的特征数值均占少数, 通过以各维度特征数值平均值替代比其小的特征数值, 从而增加具有较小特征数值点的切割路径.

其次, 该算法基于 iForest 思想来计算移动应用程序 $A_t = (c_t, P_t)$ 的权限类别异常度. 使用 $h(A_t)$ 表示 A_t 在所有 iTree 上切割路径长度的平均值, $h(A_t)$ 值越小, 表示该点被切割次数越少, A_t 异常程度越大. n_{c_t} 表示类别 c_t 中所有移动应用程序数量. 首先求解类别 c_t 中移动应用程序数据点在所有 iTree 上的平均路径长度 $s(n_{c_t})$, 用来标准化数据点 A_t 的平均路径长度 $h(A_t)$. $s(n_{c_t})$ 计算如下:

$$s(n_{c_t}) = 2H(n_{c_t} - 1) - \frac{2(n_{c_t} - 1)}{n_{c_t}}, \tag{7}$$

$$H(n_{c_t} - 1) = \ln(n_{c_t} - 1) + \xi, \tag{8}$$

其中, $H(n_{c_t} - 1)$ 是调和数, 使用 $\ln(n_{c_t} - 1) + \xi$ 近似估计. ξ 为欧拉常数, 值为 0.5772156649. 基于值 $s(n_{c_t})$ 和 $h(A_t)$, 若 $h(A_t)$ 值越小, A_t 异常程度越大, 即 O_t 值越大. O_t 可表示为

$$O_t = 2^{-\frac{h(A_t)}{s(n_{c_t})}}. \tag{9}$$

4.3 权限类别使用率

移动应用程序 A_t 的权限 $p_{t,i}$ (权限对 p_{t,i_j}) 的类别使用率 $r_{t,i}$ (r_{t,i_j}) 指在类别 c_t 中请求权限 $p_{t,i}$ (权限对 p_{t,i_j}) 的移动应用程序占类别 c_t 中全部移动应用程序的比例. 考虑到类别 c_t 内某些权限类别异常度较高的移动应用程序可能影响权限类别使用率的计算结果, 计算该参数时仅使用类别 c_t 中权限类别异常度小于阈值 θ 的移动应用程序.

$$r_{t,i} = \frac{|\{A_k\}_{c_t=c_k, O_k<\theta, p_{t,i}=1}|}{|\{A_k\}_{c_t=c_k, O_k<\theta}|}, \tag{10}$$

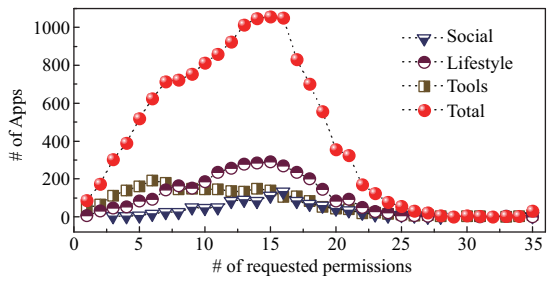

图 3 (网络版彩图) 各权限请求数量的应用程序数量分布

Figure 3 (Color online) The app number of different permissions

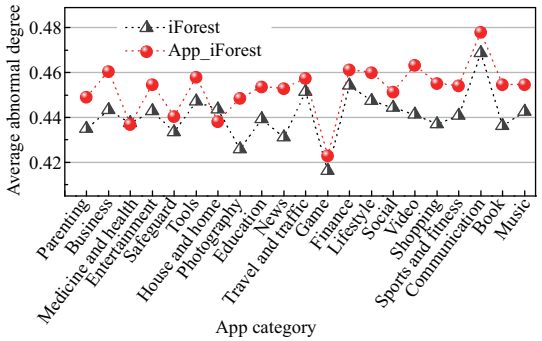

图 4 (网络版彩图) 各类别应用程序平均权限类别异常度

Figure 4 (Color online) Average abnormal degree of each app category

$$r_{t,i\_j} = \frac{|\{A_k\}_{c_t=c_k, O_k<\theta, p_{t,i\_j}=1}|}{|\{A_k\}_{c_t=c_k, O_k<\theta}|}. \tag{11}$$

4.4 权限调用者数量

已知应用本身及嵌入的服务第三方 SDK 均可调用权限而获取用户数据. 基于静态代码分析方法, 本文反向编译安卓应用程序包 (Android application package, APK) 得到源代码文件, 包括安卓虚拟机使用的字节码文件 smali. 通过分析 smali 文件所在路径组成的包名, 手工筛减无效包名并核查服务第三方后, 得到 App 数据集对应的 6097 个权限调用者, 如广告类 SDK 中的腾讯、有米广告等. 由此, A_t 内权限调用者列表记为 $U_t = \{U_{t,1}, U_{t,2}, \ldots, U_{t,g}\}$ $(g \leqslant 6097)$, 其权限调用者数量为 $u_t = |U_t|$.

5 实验与结果

本节对所提模型进行科学性和有效性的验证, 分别研究改进后的权限类别异常度算法的合理性和隐私风险量化模型对 3 条评估准则的满足性, 并探讨第三方应用商店下载机制的隐私风险预警效果.

5.1 数据集

移动应用程序包括市场应用和恶意应用. 市场应用数据集分为两部分: 一是具有元数据信息的移动应用程序, 从各第三方移动应用商店网站[4] 抓取到元数据信息, 包括类别、下载量、用户评分等, 数量达 326785 个; 二是具有应用程序包 APK 文件的移动应用程序, 经 Apktool 工具[5] 反编译解析出包名、版本、目标 SDK 版本和权限信息. 为了保证市场应用信息准确性, 筛掉信息不完全和下载量小于 100 的移动应用程序, 得到 8980 个符合实验条件的移动应用程序. 恶意应用样本数据集来自恶意软件网站 VirusShare[6], 下载并解析得到 12170 个恶意应用文件信息. 对比市场应用发现, 8980 个市场应用中存在 364 个已检测出的恶意应用版本, 这说明市场应用中存在部分异常应用, 甚至恶意应用.

4) 包括应用宝、豌豆荚、360 手机助手、小米应用商店、安智网、应用汇.
5) https://ibotpeaches.github.io/Apktool/documentation/#introduction.
6) https://virusshare.com.

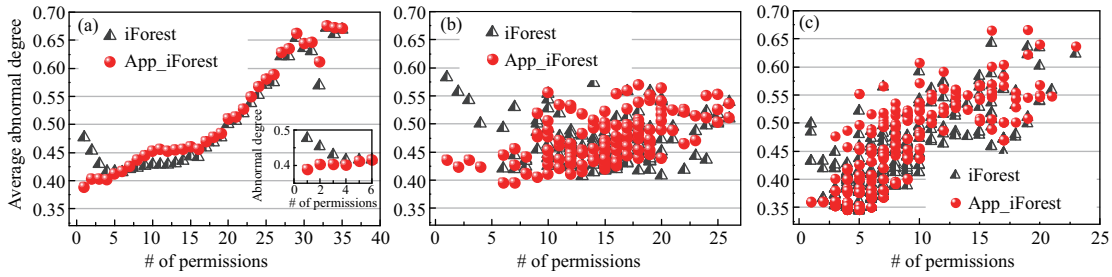

图 5 (网络版彩图) 权限请求数量与权限类别异常度关系

Figure 5 (Color online) Relation between the number of permissions and abnormal degree of permission list. (a) Market Apps; (b) communication Apps; (c) game Apps

5.2 权限类别异常度

通过 App_iForest 算法计算 App 权限类别异常度, 并得到各类别平均权限类别异常度, 结果如图 4 所示. 灰色三角线代表直接以权限列表为特征的 iForest 算法结果, 红色圆圈线则代表使用 App_iForest 算法的计算结果. 由图可知, 两种算法结果在各类别间趋势基本一致, 表明本文所提 App_iForest 异常度算法是合理的. 平均权限类别异常度越高的类别说明该类别内 App 间权限请求模式差异越大, 存在较多权限请求异常的 App. 生活助手类、办公类、理财类、视频类、通讯类是平均权限类别异常度最高的 5 个类别, 游戏类、医疗健康类、房产家居类、安全类、摄影图片类则是异常度最低的 5 个类别.

进一步从权限请求数量与权限类别异常度的对应关系上验证 App_iForest 异常度算法的有效性, 分别分析全部市场应用、异常度最高的通讯类应用、异常度最低的游戏类应用中权限请求数量与权限类别异常度的关系. 如图 5(a) 所示, 当权限请求数量在 $[0,5]$ 之间时, 原有 iForest 算法的平均异常度远大于 App_iForest 算法的平均异常度; 而在 $[5,40]$ 区间内, 两种方法得到的平均权限类别异常度趋势完全一致. App_iForest 算法解决了小数量权限请求的 App 具有高异常度水平的问题. 同时, 随着权限请求数量的增加, 平均权限类别异常度逐渐升高, 符合最小权限 (数量) 原则. 以通讯类和游戏类应用为例进行具体分析, 结果如图 5(b) 和 (c) 所示, 与全部市场应用数据集结果一致. 除 $[0,5]$ 区间, 两种方法计算的异常度水平相差很小, 再次证明本文所提的权限类别异常度算法的科学性和有效性.

5.3 移动应用程序隐私风险

本小节通过比较市场应用隐私风险值与其权限请求数量的关系研究隐私风险值分布, 结果如图 6 所示. 在一定程度上隐私风险值与权限请求数量呈正相关关系, 即隐私风险值随着权限请求数量的增多而升高, 满足风险评估准则第一条. 在相同权限数量下, 隐私风险值在最大值与最小值间浮动, 说明 App 请求的权限组合、各类别特定权限模式和权限调用者数量会对隐私风险产生影响. 进一步以通讯类和游戏类为例探讨 App 类别特定权限模式对隐私风险值的影响, 如图 6(b) 和 (c) 所示. 除再次验证了隐私风险与权限请求数量在一定程度上的正比关系外, 两类别数据中隐私风险值在相同权限请求数量上不同的增长趋势表明类别特定权限模式和权限调用者情况在量化模型中的加权作用.

将本文所提模型 PRS (privacy risk score) 与已有隐私风险量化方法 PNB (naive Bayes model with informative priors) [3], DroidRisk [4], Sensitivity_Score [6], PrivacyPalisade [8] 进行实验对比, 分别计算市场应用和恶意应用数据集内 App 隐私风险水平, 两个数据集上各方法结果作箱形图如图 7 和 8 所示 (为实验结果可对比, 本文对相关方法结果进行了最大最小归一化处理). 由图可知: (1) 本文 PRS 方法结果数据均匀集中在 [0.2,0.85] 间, 无异常点. 而基于权限使用率的 PNB 方法结果值多集中在

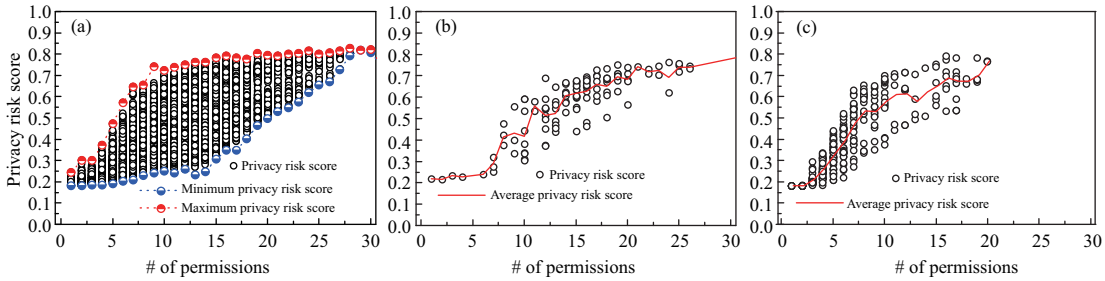

图 6 (网络版彩图) 移动应用程序的隐私风险与权限请求数量关系

Figure 6 (Color online) Relation between the number of permissions and privacy risk score of applications. (a) Market Apps; (b) communication Apps; (c) game Apps

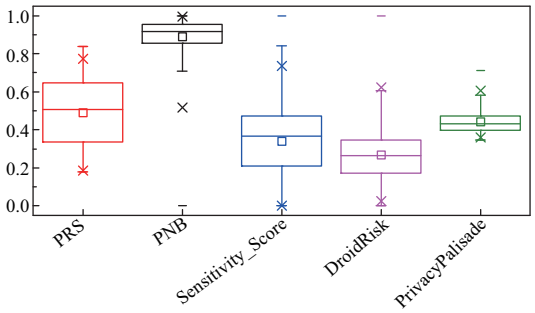

图 7 (网络版彩图) 市场应用隐私风险量化结果箱形图

Figure 7 (Color online) Boxplots of the risk levels of market Apps

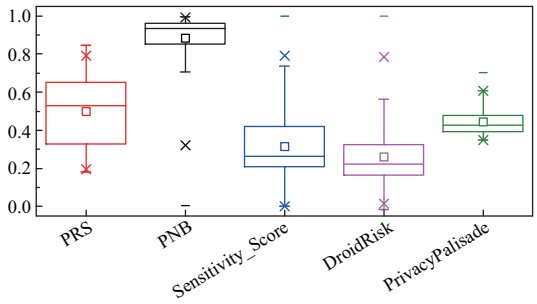

图 8 (网络版彩图) 恶意应用隐私风险量化结果箱形图

Figure 8 (Color online) Boxplots of the risk levels of malware Apps

[0.85,0.95] 区间内, 低值区间多有异常值; 基于权限敏感度的 Sensitivity_Score, DroidRisk 方法在高值区间分布少量异常点; 基于权限异常度的 PrivacyPalisade 方法以 0.5 为均值, 多集中在 [0.4,0.6] 区间, 高值区间含异常点. (2) 对比恶意应用和市场应用结果分布, 仅 PRS 和 PNB 方法的恶意应用隐私风险水平中位数值增大, 表明恶意应用的高隐私风险特征. 恶意应用识别度良好表明本文模型合理性, 并验证了本文风险量化方法在同类工作中更为有效.

5.4 恶意应用检测

基于权限特征进行恶意应用检测是一种最为常用的静态特征检测方法 [36,37,40]. 本小节通过恶意应用检测验证所提模型的有效性. 已知市场应用中存在部分异常应用, 为验证模型在恶意应用检测上的稳定性, 本小节将 8980 个市场应用分为两个数据集分别进行实验: (1) Market_0 数据集, 选取市场应用中权限类别异常度小于其类别平均权限类别异常度的 App, 包括 6316 个 App 版本; (2) Market_1 数据集, 选取市场应用中与对应类别恶意应用等量的最小类别异常度的 App, 包括 465 个 App 版本.

基于隐私风险值检测恶意应用的基本原理是通过定义隐私风险阈值比例 β, 把隐私风险值位于前 β 的 App 归为恶意应用, 其余为正常应用. 选取不同的阈值比例 β, 可得到恶意应用检测性能 ROC 曲线. 通过对比本文所提权限类别异常度算法 App_iForest、隐私风险值量化方法 PRS 方法与其他 3 种基于权限的异常应用检测方法 [3,5,36] 在的恶意应用检测效果来验证模型有效性, 结果如图 9 所示. 在 Market_0 数据集上, 本文所提出的权限类别异常度算法与隐私风险值量化模型效果相差不大, 但优于

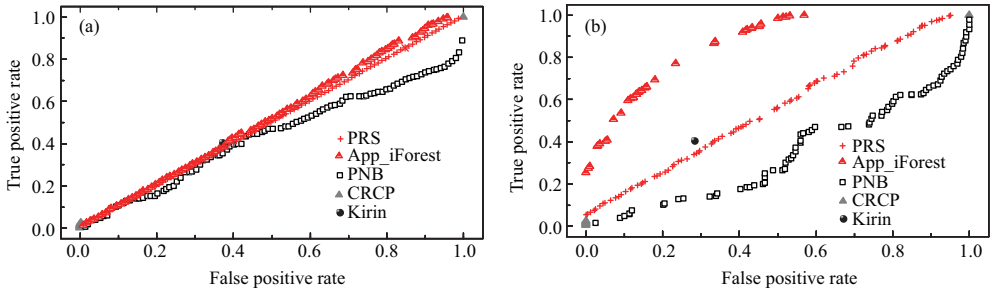

图 9　(网络版彩图) 恶意应用检测 ROC 曲线

Figure 9　(Color online) ROC curve for malicious application detection. (a) Market_0; (b) Market_1

PNB 方法 [3] 和 CRCP 风险信号方法 [5]. 在异常应用含有量较少的 Market_1 数据集上, 各模型效果差异拉大, 类别异常度检测方法远优于其他方法, 基于隐私风险值的检测方法效果次之. 其中 Kirin 检验规则 [36] 效果最差, 因为其中 9 条恶意应用检测规则目前只有 7 条可以使用. 随着恶意应用的演化, 部分检验规则失效. CRCP 风险信号方法只关注各权限的使用率, 其效果不如考量了权限敏感度的本文方法和 PNB 方法. PNB 方法也考虑了权限的多样特征, 其模型缺陷主要是结果值与权限数量的单调性会在权限使用率大于 50% 时被破坏. 以上实验结果证明本文模型在同类恶意应用检测算法中效果较优.

5.5　风险预警机制

当用户在第三方移动应用商店下载并安装 App 时, 除权限请求和隐私政策条款, 用户没有其他途径获知该 App 隐私风险, 然而权限请求和政策条款的高度集成语言不易被用户理解. 第三方移动应用商店仅直观地显示 App 的用户评分和下载量, 辅助用户选择是否下载该 App. 本小节通过分别分析第三方移动应用商店 App 下载量、用户评分 (评分值在 [0,5] 区间) 与隐私风险值的关系, 进一步研究应用商店在 App 隐私风险管理中的作用, 主要探讨其下载机制是否能满足 App 隐私风险预警功能需要.

首先观察各 App 隐私风险值与其下载量的关系, 把市场应用的下载量划分为如图 10(a) 中 X 轴显示的 8 个区间, 并对各区间 App 个数和平均隐私风险值进行统计. 随着下载量增加, 对应下载量区间内 App 个数先增加后减少, [10k, 100k) 区间内 App 数量最多. 平均隐私风险值和平均权限请求数量随着下载量的增大而增大, 这说明用户下载 App 时并未考虑该其隐私风险. 其次, 观察各 App 隐私风险值与对应评分间关系, 统计 $[0,5]$ 区间内各评分下 App 数量及平均隐私风险值. 图 10(b) 结果显示隐私风险值与用户评分之间没有明显的线性关系, 各评分对应的平均隐私风险值在 0.5 左右浮动. 其中评分在 4.5 以上和 1.5 以下的 App 具有较高的隐私风险值, 各评分下的平均隐私风险水平出现 “两极化”. 这说明评分机制并不能起到有效的风险预警作用, 评分较高的 App 可能具有较高的隐私风险值. 这主要是因为评分机制并不是针对用户隐私风险值而设置, 它更多反应 App 功能及运行效果.

综合上述, 在下载 App 时, 用户通常不关心其隐私风险情况, 或 “妥协” 于其功能服务需要. 用户评分并不是一个可行的隐私风险预警机制. 鉴于此, 建立有效的 App 隐私风险预警机制是一项紧急且必要的工作. 基于本文隐私风险量化方法, 在用户安装或使用 App 过程中显式通知其隐私风险或在 App 详情页上显式标注 App 隐私风险值, 是简单有效的保护用户隐私的方法.

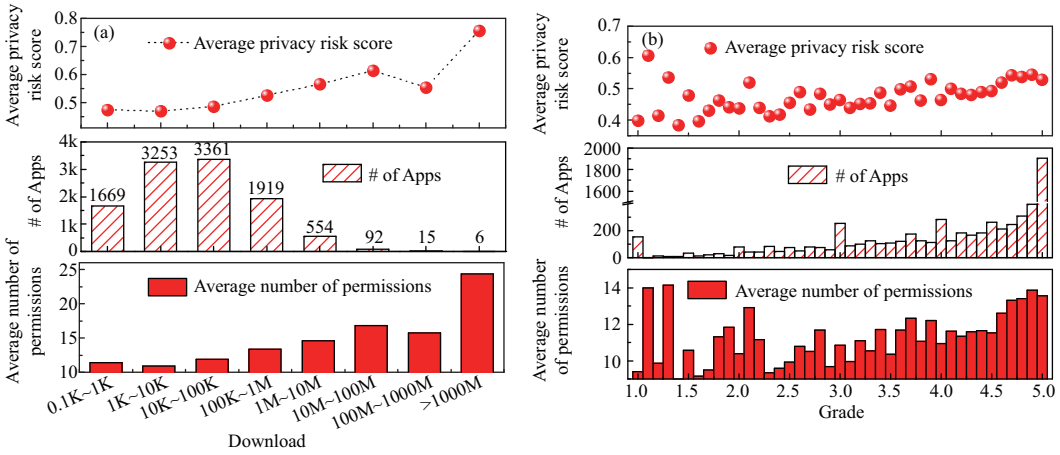

图 10　(网络版彩图) 移动应用程序市场的风险预警机制

Figure 10　(Color online) Privacy risk early-warning mechanisms of mobile application market. (a) Download; (b) grade

6　总结

为了对 App 隐私进行主动监测和评估, 本文提出了一种基于权限的分类别 App 隐私风险量化模型. 考虑权限与用户隐私数据的映射关系, 综合定义了权限变量的参数 – 权限敏感度、权限类别异常度、权限使用率和权限调用者, 最后结合权限分析原则构建隐私风险量化模型. 实验证明该模型满足隐私风险评估的 3 条基本准则, 且风险量化评估方法在同类方法中效果较优. 对于恶意应用检测, 正常应用和恶意应用风险量化值的明显区分度表明了模型可用性, 与同类异常应用检测算法的结果对比也表明了模型的有效性. 实验同时显示目前国内第三方移动应用商店的下载量及用户打分评级功能并不能起到有效的隐私风险警示作用, 直观展示 App 隐私风险值是简单有效地保护用户隐私的方法.

本文基于静态的权限分析方法有效评估了 App 收集用户数据造成的最大隐私风险. 下一步, 我们将基于动态分析方法研究用户使用 App 时的实时隐私风险变化. 首先组建用户隐私数据库, 通过动态信息流追踪等方法分析用户操作 App 时的数据流动情况, 进而构造出操作行为与隐私数据间的映射关系. 然后选取隐私数据参数, 提出并验证 App 用户的动态隐私风险量化模型.

补充材料　表 S1–S4. 本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

1　Meng X F, Zhang X J. Big data privacy management. J Comput Res Dev, 2015, 52: 265–281 [孟小峰, 张啸剑. 大数据隐私管理. 计算机研究与发展, 2015, 52: 265–281]

2　Demetriou S, Merrill W, Yang W, et al. Free for all! Assessing user data exposure to advertising libraries on Android. In: Proceedings of Annual Network and Distributed System Security Symposium, 2016

3　Peng H, Gates C, Sarma B, et al. Using probabilistic generative models for ranking risks of Android Apps. In: Proceedings of ACM Conference on Computer and Communications Security, 2012. 241–252

4　Wang Y, Zheng J, Sun C, et al. Quantitative security risk assessment of Android permissions and applications. In: Proceedings of IFIP Annual Conference on Data and Applications Security and Privacy, 2013. 226–241

5　Sarma B, Li N, Gates C, et al. Android permissions: a perspective combining risks and benefits. In: Proceedings of the 17th ACM Symposium on Access Control Models and Technologies, 2012. 13–22

6 Liccardi I, Pato J, Weitzner D J. Improving user choice through better mobile Apps transparency and permissions analysis. J Priv Confidentiality, 2014, 5: 1

7 Hamed A, Ayed H K-B, Machfar D. Assessment for Android Apps permissions a proactive approach toward privacy risk. In: Proceedings of the 13th International Wireless Communications and Mobile Computing Conference, 2017. 1465–1470

8 Quattrone A, Kulik L, Tanin E, et al. PrivacyPalisade: evaluating App permissions and building privacy into smartphones. In: Proceedings of IEEE International Conference on Information Communication and Signal Processing, 2015

9 Barrera D, Kayacik H G, van Oorschot P C, et al. A methodology for empirical analysis of permission-based security models and its application to Android. In: Proceedings of CCS'10, 2010. 73–84

10 Felt A P, Chin E, Hanna S, et al. Android permissions demystified. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, 2011. 627–638

11 Chia P H, Yamamoto Y, Asokan N. Is this app safe?: a large scale study on application permissions and risk signals. In: Proceedings of the 21st International Conference on World Wide Web, 2012. 311–320

12 Frank M, Dong B, Felt A P, et al. Mining permission request patterns from Android and facebook applications. In: Proceedings of IEEE International Conference on Data Mining, 2012. 870–875

13 Wei M K, Gong X, Wang W Y. Claim what you need: a text-mining approach on Android permission request authorization. In: Proceedings of IEEE Global Communications Conference, 2015

14 Slavin S, Wang X Y, Hosseini M B, et al. Toward a framework for detecting privacy policy violations in Android application code. In: Proceedings of International Conference on Software Engineering, 2016. 25–36

15 Story P, Zimmeck S, Sadeh N. Which Apps have privacy policies? An analysis of over one million google play store Apps. 2018. https://usableprivacy.org/static/files/Story_APF_2018.pdf

16 Yu L, Luo X P, Liu X L, et al. Can we trust the privacy policies of Android Apps? In: Proceedings of the 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2016. 538–549

17 Au K W Y, Zhou Y F, Huang Z, et al. PScout: analyzing the Android permission specification. In: Proceedings of ACM Conference on Computer and Communications Security, 2012. 217–228

18 Gordon M I, Kim D, Perkins J H, et al. Information flow analysis of Android applications in DroidSafe. In: Proceedings of Annual Network and Distributed System Security Symposium, 2015

19 Zimmeck S, Wang Z Q, Zou L Y, et al. Automated analysis of privacy requirements for mobile Apps. In: Proceedings of Annual Network and Distributed System Security Symposium, 2017

20 McReynolds E, Hubbard S, Timothy L, et al. Toys that listen: a study of parents, children, and internet-connected toys. In: Proceedings of Conference on Human Factors in Computing Systems, 2017. 5197–5207

21 Ren J J, Rao A, Lindorfer M, et al. ReCon: revealing and controlling privacy leaks in mobile network traffic. In: Proceedings of ACM SIGMOBILE MobiSys, 2016. 361–374

22 Reyes I, Wijesekera P, Reardon J, et al. "Won't somebody think of the children?" Examining COPPA compliance at scale. In: Proceedings of Privacy Enhancing Technologies, 2018. 63–83

23 Enck W, Gilbert P, Chun B G, et al. TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. Commun ACM, 2014, 57: 99–106

24 Le A, Varmarken J, Langhoff S, et al. AntMonitor: a system for monitoring from mobile devices. In: Proceedings of ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data, 2015. 15–20

25 You W, Liang B, Shi W C, et al. TaintMan: an ART-compatible dynamic taint analysis framework on unmodified and non-rooted android devices. IEEE Trans Depend Secure Comput, 2020, 17: 209–222

26 Reardon J, Feal A, Wijesekera P, et al. 50 ways to leak your data: an exploration of Apps' circumvention of the Android permissions system. In: Proceedings of USENIX Security Symposium, 2019. 603–620

27 Kiran K V D, Mukkamala S, Katragadda A, et al. Performance and analysis of risk assessment methodologies in information security. Int J Comput Trends Technol, 2013, 4: 7–26

28 Mylonas A, Gritzalis D, Tsoumas B, et al. A qualitative metrics vector for the awareness of smartphone security users. In: Proceedings of the Trust, Privacy, and Security in Digital Business, 2013. 173–184

29 Agence nationale de la sécurité des systèms d'information: EBIOS — expression des Besoins et identification des objectifs de Sécurité. https://www.ssi.gouv.fr/guide/ebios-2010-expression-des-besoins-et-identification-des-objectifs-

de-securite/
30 Liu F T, Ting K M, Zhou Z H. Isolation-based anomaly detection. ACM Trans Knowl Discov Data, 2012, 6: 1–39
31 Meng W, Ding R, Chung S P, et al. The price of free: privacy leakage in personalized mobile in-Apps ads. In: Proceedings of Annual Network and Distributed System Security Symposium, 2016
32 Taylor V F, Beresford A R, Martinovic I. Intra-library collusion: a potential privacy nightmare on smartphones. 2017. ArXiv:1708.03520
33 Nath S. MAdScope: characterizing mobile in-app targeted ads. In: Proceedings of International Conference on Mobile Systems, Applications and Services, 2015. 59–73
34 Zhou Y J, Jiang X X. Dissecting Android malware: characterization and evolution. In: Proceedings of IEEE Symposium on Security and Privacy, 2012. 95–109
35 Liccardi I, Pato J, Weitzner D J, et al. No technical understanding required: helping users make informed choices about access to their personal data. In: Proceedings of MOBIQUITOUS'14, 2014. 140–150
36 Enck W, Ongtang M, Mcdaniel P. On lightweight mobile phone application certification. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, 2009. 235–245
37 Aung Z, Zaw W. Permission-based Android malware detection. Int J Sci Technol Res, 2013, 2: 228–234
38 Rahul R, Xiao X S, Yang W, et al. WHYPER: towards automating risk assessment of mobile applications. In: Proceedings of Usenix Conference on Security, 2013. 89–97
39 Ilyas I F, Markl V, Haas P J, et al. CORDS: automatic discovery of correlations and soft functional dependencies. In: Proceedings of International Conference on Management of Data, 2004. 647–658
40 Li J H, Qu C. Survey of Android malware detection methods. J Appl Res Comput, 2019, 36: 1–7

Privacy risk quantification of mobile application based on requested permissions

Minjie ZHU, Qingqing YE, Xiaofeng MENG* & Xin YANG

School of Information, Renmin University of China, Beijing 100872*, China*
* Corresponding author. E-mail: xfmeng@ruc.edu.cn

Abstract With the prevalence of mobile devices and mobile applications (Apps), service providers have become increasingly enthusiastic in collecting user data, which would cause huge privacy risk due to the invisibility of data collection. How to evaluate the maximum privacy risks of mobile Apps is a key challenge, which not only contributes to the regulation of App market, but also helps users to avoid potential privacy leakage. By investigating the maximum data leakage of an App, this paper proposes a privacy risk quantification model based on the requested permissions and the principles of App permission analysis. The proposed model introduces four important parameters, namely, permission sensitivity, anomaly degree of permission list, utilization rate of an App, and number of permission callers, to evaluate the potential privacy risk of an App. We conduct experiments of privacy risk evaluation and malicious App detection over real datasets, and the results show that our proposed model achieves better performance against state-of-the-art solutions, which demonstrates the effectiveness of this model. Further, analytical results also indicate that this privacy risk quantification model can serve as an effective privacy risk warning mechanism for user privacy preservation.

Keywords privacy protection, mobile application, privacy risk quantification, permission-based analytical method

Minjie ZHU was born in 1993. She received her M.S. degree in computer application from Renmin University of China in 2019. Her research interests include mobile privacy and privacy protection.

Qingqing YE was born in 1992. She is a research assistant professor in The Hong Kong Polytechnic University. She received her Ph.D. degree from Renmin University of China in 2020. She has received several prestigious awards, including National Scholarship and IEEE S&P Travel Award. Her research interests include data privacy and security, and adversarial machine learning.

Xiaofeng MENG was born in 1964. He is a professor and Ph.D. supervisor at Renmin University of China, as well as a fellow of the China Computer Federation. His main research interests include cloud data management, web data management, flash-based databases, and privacy protection.

Xin YAN was born in 1994. He received his M.S. degree in software engineering from Renmin University of China in 2020. His research interests include privacy protection and blockchain.