

App潜在数据共享行为智能挖掘

刘凡嘉, 陈裕鑫, 王杨德, 唐鹏, 邱卫东

(上海交通大学计算机学院(网络空间安全学院、密码学院), 上海 200240)

摘要: 移动应用软件成为人们日常生活的重要组成部分, 同时也带来了数据隐私安全风险。一些软件将用户数据通过后台分享给第三方, 或者从第三方获得用户数据, 进行精准的个性化推荐。这种数据共享渠道具有隐蔽性, 在当前监管体系下难以被有效追踪。针对这一问题, 设计了一种自动化检测App之间潜在数据共享行为的智能挖掘方案。该方案构建了基于自动化测试的检测框架, 通过模拟用户行为, 利用关键词注入用户偏好数据, 收集个性化推荐界面截图; 引入基于人工智能模型的信息提取和关联性分析技术, 利用光学字符识别模型、多模态视觉大模型和大语言模型, 完成对App截图中的文本信息和图像内容的智能提取和理解, 实现对关键词和图像内容描述之间的关联性判断; 对30个App完成了用户偏好注入, 建立了含8 700张截图的数据集, 获得了1 466条数据共享行为证据。在30个App中, 有23个检出存在向第三方提供数据的行为, 有29个存在从第三方接收数据并进行个性化推荐的行为。构建了App数据共享行为的网络结构, 为App数据共享行为的自动化检测提供了新的解决方案。

关键词: 移动应用软件; 数据共享检测; 隐私保护; 自动化测试

中图分类号: TP393.0

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.AQ26003

Intelligent mining of App's potential data sharing behavior

Liu Fanjia, Chen Yuxin, Wang Yangde, Tang Peng, Qiu Weidong

School of Computer Science (School of Cyber Science and Engineering, School of Cryptography),
Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: Mobile App have become an essential part of people's daily lives, but they also brought about security risks to data privacy. Some Apps share user data with third parties in the background or obtain user data from third parties to deliver personalized recommendations. Due to the stealthiness of this data sharing channel, it is difficult to be effectively tracked under the current regulatory system. In response to this issue, an intelligent mining solution was designed and implemented to automatically detect potential data sharing behaviors between Apps. The detection framework was constructed based on automated testing, which injects user preference data with keywords by simulating user behavior, and then collects screenshots of personalized recommendation interfaces. Information extraction and correlation analysis technologies based on AI models were introduced, using optical character recognition models, multimodal visual large models, and large language models to complete the intelligent extraction and understanding of text information and image content in App screenshots, and to make judgments on the correlation between keywords and image content descriptions. The injection of user preference data was performed on 30 Apps, resulting in a dataset of 8 700 screenshots, and yielding 1 466 pieces of evidence of data sharing. Among the 30 Apps, 23 were found to provide data to third parties, and 29 were found to receive data from third parties and engage in personalized recommendations. A network structure of App data sharing behavior was constructed, provid-

收稿日期: 2026-01-08; 修回日期: 2026-04-28

通信作者: 邱卫东, qiuwd@sjtu.edu.cn

ing a new solution for the automated detection of App data sharing behavior.

Key words: mobile application, data sharing detection, privacy protection, automated testing

0 引言

随着数字化时代的到来,移动应用软件(application, App)已成为人们日常生活中不可或缺的一部分,其服务范围覆盖了在线购物、影音娱乐、社交通信、生活服务等各个领域,极大地丰富和便利了人们的日常生活。然而,这些App在为用户提供便捷服务的同时,也引发了用户对个人隐私安全的广泛关注和深刻担忧。

当前,国内外App在数据收集和共享行为上呈现出复杂多样的态势。许多App在未经用户明确同意的情况下,过度索取手机系统权限,甚至非法收集、存储和传输用户的个人信息^[1]。这种数据滥用行为不仅侵犯了用户的隐私权,也为不法分子提供了可乘之机,严重威胁用户的信息安全。

为了应对这一挑战,世界各国政府纷纷出台了严格的法律法规来保护个人信息。如欧盟的《通用数据保护条例》(general data protection regulation, GDPR)^[2]和中国的《中华人民共和国个人信息保护法》^[3]等,均对App的数据收集、使用、存储和共享等提出了明确要求,并规定了相应的处罚措施。

然而,这些法律法规在执行过程中仍然存在一些监管漏洞。部分App开发者通过后台数据共享的方式,绕过监管,将用户数据分享给第三方,或者从第三方购买并使用用户数据,进行精准广告推送、个性化推荐等商业行为。App后台数据共享的逻辑框架如图1所示。这种数据共享行为一般通过接入第三方提供的软件开发工具包(software development kit, SDK)或应用程序编程接口(application programming interface, API)来实现,由于App仅须调用接口来获得所需要的数据,不需要在用户设备本地执行个人信息处理,因此无法通过代码审查的方式被发现,在当前监管体系下难以被有效追踪和监控。

因此,本文的研究旨在利用智能化的信息挖掘和关联性分析手段,探索一种App之间潜在数据共享行为的自动化检测方案,实现从数据注

入,到页面检测,再到关联性分析,最后得出结论的智能化全流程。本文的主要贡献和创新点如下。

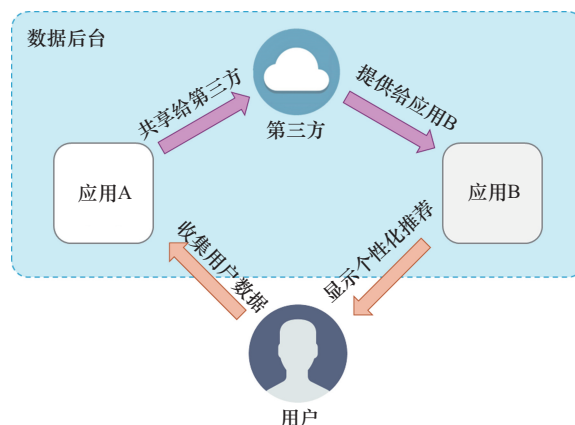


图1 App后台数据共享的逻辑框架

1) 提出了一种基于App前端用户界面呈现内容的潜在数据共享行为检测思路。不同于依赖流量分析、代码审查或权限分析的方法,本文从用户可直接感知的个性化推荐界面出发,对跨App数据共享行为进行间接判定,规避了中间传输过程中可能存在的加密或混淆机制。

2) 设计并实现了一套App内自动化行为模拟与个性化推荐界面挖掘方案。该方案能够在不需要大量人工操作的情况下,批量完成用户偏好注入与推荐页面采集,具备较高的自动化程度和可扩展性,为大规模App检测实验提供了可行的技术支撑。

3) 构建了一种基于人工智能模型进行信息提取和关联性分析的方法,从文本匹配、图像语义理解和深层语义关联等多个维度综合分析推荐内容与用户偏好之间的关系,并通过赋权机制对不同类型证据进行量化整合,评估潜在数据共享行为的可能性大小。

4) 从研究意义上看,本文为App潜在数据共享行为的智能挖掘提供了一种新的技术路径,丰富了移动应用隐私保护领域的研究方法;从实践角度看,所提出的方法可为监管部门、应用开发者及安全研究人员提供辅助检测手段,有助于提

升对 App 数据共享行为的可观测性和评估能力,促进移动应用生态的合规与健康发展。

1 相关工作

1.1 App 收集个人信息现状

当前, App 在用户数据处理和权限索取上的深层问题成为国内外研究者关注的焦点,一些对用户数据共享行为的调查显示了人们对这一问题的普遍疑虑。文献[4]进行了一项调查,结果显示有九成用户怀疑自己在不同 App 间被跟踪,有 93% 的用户认为一些 App 在未经用户同意的情况下分享其个人信息。大多数用户认为游戏、娱乐、社交和购物等类型的 App 收集的数据更多,而工具和教育类 App 则收集得较少。文献[5]基于对中国 4G 网络的 280 亿匿名访问日志的分析发现,用户被广泛地跨应用跟踪,并形成了专门的跟踪社群。有 10% 的用户面临个人身份信息 (personally identifiable information, PII) 泄露风险,但 90% 的 PII 跟踪流量限于中国境内。

App 不仅广泛收集用户信息,还在用户不知情的情况下将他们的个人信息与第三方共享,这种情况已成为普遍现象。文献[6]使用中间人代理记录 App 的网络流量,测试了 110 个流行的免费 App。结果表明,很大一部分 App 在未通知用户的情况下与第三方共享用户的个人数据和搜索词等信息,并且安卓应用比 iOS 应用更有可能与第三方共享个人信息。

此外,还有一些研究着眼于揭示 App 收集个人数据的动机和利益。如文献[7]设计了一种推断数据收集目的的技术,解决了用户对 App 收集敏感数据目的不理解的问题。文献[8]探讨了引入第三方数据是如何影响市场竞争的,并指出服务提供商通常会根据平台特点选择不同的追踪方法。

1.2 现有检测 App 数据收集和共享的方法

为了判断哪些 App 以何种方式、何种规模实施了用户数据共享的行为,研究者采取了各种方法来识别。文献[9]开发了一种自动化方法,通过真实世界的网络流量数据来检测第三方广告跟踪服务,结果识别出 2 121 个此类服务,其中 233 个之前未被广告跟踪黑名单所发现。文献[10]提出了一种基于流量特征的隐私泄露评估方案 Hos-

tRisk, 利用词频-逆文本频率模型和层次聚类,通过分析 App 内域名的行为特征来计算其与业务的相关性,识别出高风险的第三方域名。

有研究者结合法律法规来检测 App 的隐私保护情况。文献[11]考察了移动健康类 App 中 GDPR 合规性的现状,提出健康隐私合规检测系统 (HPDROID) 来识别 3 种违例行为: 隐私政策不完整、数据收集不一致和数据传输不安全。结果显示, 77.9% 的 App 包含至少一种不一致的数据收集行为。文献[12]也基于 GDPR 强调了个人数据跨境转移的风险。文献[13]面向 GDPR 的合规需求,构建了一套包含隐私政策收集、层级结构还原及概念分类器的智能化检测框架,搭建 GDPR 知识图谱,显著提升了对隐私政策中特定概念识别的效果。文献[14]则构建了面向《中华人民共和国个人信息保护法》的多级隐私政策知识图谱及中文语料库,开发了中文隐私政策分类模型 (CPP-BERT),实现了对隐私政策中细粒度概念的高效识别与合规性分析。

而在技术层面,一些研究聚焦于开发具有针对性的系统和工具,评估 App 存在的隐私风险,帮助用户发现并避免此类 App 的行为。文献[15]提出了一种根据应用描述和 API 使用情况来识别其所需要的最低权限的方法,并通过检查 App 的额外权限来评估隐私风险。文献[16]利用机器学习技术对网络流量进行分类,识别出多种 PII 泄露行为,能够揭示并让用户阻止 PII 的泄露。文献[17]和文献[18]分别通过动态和静态的污点分析方法,监控手机上第三方应用的信息流流向,以检测敏感数据的使用和共享情况。文献[19]提出了一种隐私风险评估方案,通过统计第三方获取的权限并构建权限组合的非线性影响模型,量化评估隐私泄露风险并给出权限管理策略。此外,还有很多研究者^[20-24]通过堆栈跟踪、联邦学习、静态数据流分析等方式,发现了个人信息被收集、共享的多种渠道,并采取措施通知用户。

近年来,已有研究关注移动应用生态中的数据共享与隐私风险问题。文献[25]针对跨应用内容共享问题开展了系统性研究,重点分析了内容在不同 App 间传播过程中所引发的隐私风险。该研究通过构建跨应用内容传播路径,对用户在不

同平台中的行为关联进行建模,揭示了群体隐私泄露及跨应用信息推断的潜在风险机制。文献[26]从用户行为与使用场景的角度出发,通过分析不同应用场景下用户数据的交互方式,探讨了多应用环境中的隐私信息流转特征,并构建了基于场景差异的隐私交换行为分析模型。结果表明,不同场景对隐私数据的采集与共享有显著影响,为理解应用间数据流动提供了行为层面的解释框架。

研究者们通过流量分析、污点分析、权限识别等多种手段来发现个人信息的滥用,然而前端页面是用户在使用App并与之交互的过程中最直接的方式。对于大多数一般用户而言,App在屏幕上呈现的图像和文字内容是他们察觉到自己的个人信息被收集并共享的唯一方式。一项调查显示^[27],有82%的受访者遇到过在某一个App搜索的内容在另一个App中被推荐的情况。

然而,由于很多App在用户毫无察觉的情况下通过后台向第三方共享用户数据,用户难以追踪也难以阻止信息被收集,给人们的信息安全带来了潜在的隐患。目前,国内外的研究仍然缺乏从前端呈现这一用户体验最直观的角度来进行实验的研究,也缺乏对这一行为进行自动化、智能化挖掘的系统。本文从用户可感知的前端推荐内容出发,在不需要访问底层数据流或系统权限信息的情况下,对App之间的潜在数据共享行为进行建模与分析,为该问题提供了一种新的研究视角。

2 App潜在数据共享行为自动化挖掘方案

在现实生活中,用户偏好往往是通过一系列零散、长时间、潜移默化的互动行为逐渐形成的。这种形成过程的复杂性和不确定性使得手动模拟用户行为变得既不现实也不准确。为了解决上述问题,本文提出了一种App内自动化行为模拟和界面呈现挖掘方案。该方案通过运行脚本程序,对批量App进行自动化的目标数据注入和个性化推荐检测操作。

2.1 自动化框架与流程

本文的实验思路基于如下假设:如果一个App记录了用户的特定偏好,且存在与其他App的潜在数据共享行为,那么这些偏好可能会在其

他App的个性化推荐内容中体现出来。基于这一假设,本文设计了一个分为两个阶段的实验流程。首先,通过在其所选App内执行特定内容的搜索、浏览、互动等用户行为,诱导App将该类内容作为用户偏好记录下来。如果该App确实存在数据共享行为,那么某些其他App会接收到该用户的偏好数据。于是,多次收集其他所有App个性化推荐页面的截图,在后续流程中检测其是否确实体现了之前注入过的用户偏好。

本文选择了30个被测App,并采用“一次注入+一轮检测”的模式进行实验。对于每个App,逐一执行5个预设关键词的目标数据注入操作。在完成一个App的注入后,对其他29个App执行一轮检测,观察它们是否在个性化推荐页面中展示了与注入关键词相关的信息。整个流程的形式化如算法1所示,自动化挖掘方案整体流程如图2所示。

算法1 App数据共享自动化挖掘算法

输入 待测试App集合 $A = \{a_1, a_2, \dots, a_n\}$, 用户偏好关键词集合 $K = \{k_1, k_2, \dots, k_m\}$
输出 个性化推荐页面截图集合 S
 (1) 初始化 $S = \emptyset$.
 (2) 对于每个待测试App, $a_i \in A$:
 (3) 对每个关键词 $\{k_l \in K | i \leq l \leq i + 5\}$:
 (4) 执行关键词 k_l 的数据注入。
 (5) 对其余待测试App, $a_j \in A - \{a_i\}$:
 (6) 执行个性化页面收集,得到本轮截图集合 S_{ij} , $S = S \cup S_{ij}$.

需要注意的是,不能在短时间内完成多个轮次的App数据注入和推荐页面收集,否则App后台可能尚未进行数据共享,无法在用户的个性化推荐界面检测出来,也就无法判断是否存在数据共享行为。实验流程采用上述多个轮次间歇交替执行的模式,在不同时间点设置多轮检测,保证数据有充足的流转时间。

2.2 实验设计

2.2.1 实验平台与工具

本文选择MuMu模拟器作为实验平台,为了实现宿主主机程序与模拟器内的安卓系统之间的交互,选择Appium作为主要的自动化工具。

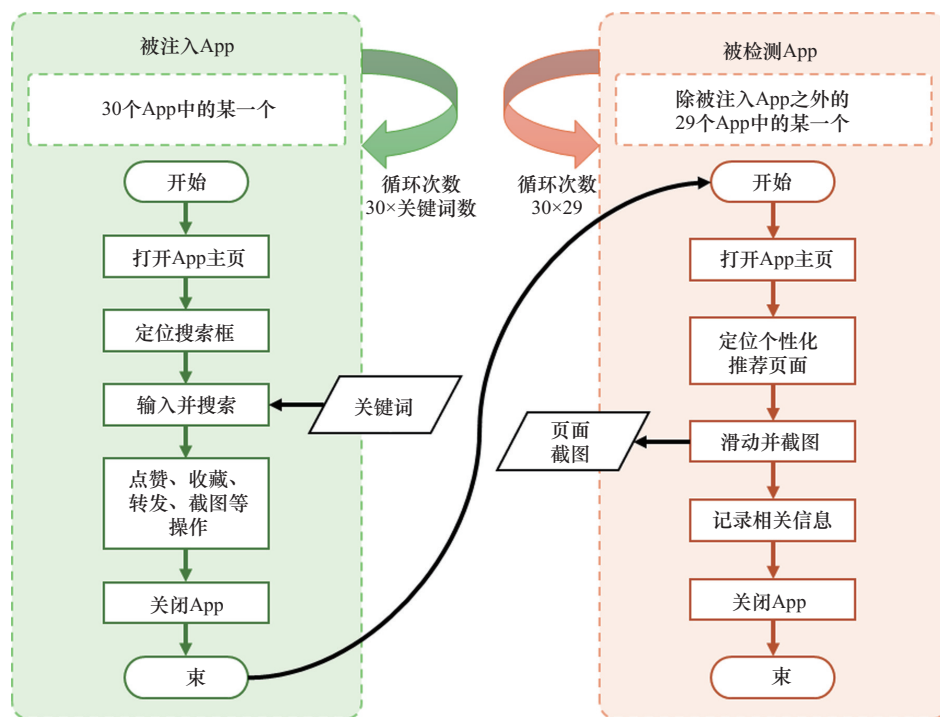


图2 自动化挖掘方案整体流程

Appium 是一个开源的、跨平台的自动化测试框架，允许测试人员使用 WebDriver 协议与移动平台进行交互。Appium 的作用原理是基于客户端-服务器架构，在进行自动化测试时，将 Appium 服务器作为中介，接收来自客户端的超文本传输协议（hyper text transfer protocol, HTTP）请求，并将其转换为相应的操作，发送给模拟器或真机上的安卓系统。客户端内置的 jar 文件将包含配置细节和测试用例的自动化脚本转换为基于 JavaScript 的轻量级的数据交换（JavaScript object notation, JSON）格式发送到服务器，服务器识别命令后与相应的终端设备建立连接，并触发终端设备上测试用例的执行。终端设备以 HTTP 的形式向 Appium 发送响应。

2.2.2 被测试 App 选取

选取被测试的 App 时，为了保证实验具有较强的可行性和现实意义，需要设置一些具体的标准，如流行度广、知名度高、活跃用户规模大；具有搜索功能和个性化推荐页面，具有用户偏好收集倾向；涵盖尽可能多的类别和领域，对模拟器环境排斥要尽可能小。根据上述标准，本文参考应用市场的排行榜，选择了若干用户群体广泛的 App 作为实验对象，从应用市场下载上述 App

的安装包，并在 MuMu 模拟器中安装。

在实验过程中，部分 App 对模拟器环境或自动化操作行为具备一定的检测与防护机制，发现异常后会触发相应的安全策略，如弹出验证码、限制登录、强制退出、中断网络连接等，一定程度上会对自动化流程的连续执行产生干扰。为此，本文在实验初期对 App 进行了可行性评估：如果触发安全机制，首先判断能否通过一次性的人工干预恢复正常使用；若干预后自动化流程仍无法稳定运行，则认为该 App 在当前实验环境下不适合开展自动化测试，并将其排除在实验对象之外。最终，本文筛选了 30 个被测 App，均能在模拟器环境中完成完整的自动化注入与检测流程。

2.2.3 关键词设计

为了确保关键词的设计能够有效地服务于实验目的，本文在设计时遵循了以下标准：符合 App 的服务领域；选择实体、物品、人物等类别的名词，而非抽象概念、形容词、动词等；特异性和指向性强，粒度尽可能细。避开当前有显著热度的公共话题以及与模拟器虚拟定位地点相关的关键词，降低热点或地域性推荐机制对实验结果的干扰。综合考虑效果和时间成本，本文为每

个App选择了5个关键词,同时对一些关键词设置了若干近义词。

2.2.4 实验账号设置

考虑到App通常依赖账号相关的键值信息对用户数据进行存储与标识,实验中须保证账号条件的一致性,以模拟同一用户同时使用多款App的实际场景。在与第三方进行数据共享时,App可能通过手机号、国际移动设备识别码(international mobile equipment identity, IMEI)等唯一特征将来自外部的数据与本地用户画像进行匹配关联。因此,为避免因账号差异引入不必要的干扰变量,实验采用统一的账号注册条件。此外,鉴于用户偏好建模具有累积特性,过往使用记录可能对个性化推荐结果产生影响。为避免历史行为对实验结果的干扰,本文使用新申请的手机号进行注册,确保所有被测App均在同一安卓系统环境下完成注册和使用。通过上述设置,方案能在控制账号与设备变量的前提下,更集中地考察App之间潜在的数据共享行为。

2.2.5 元素定位

要实现Appium自动化测试,需要准确地与App界面中的元素进行交互,也就需要找到这些元素在用户界面(user interface, UI)中的位置。本文使用了Python的Selenium库,使用元素选择器来唯一地定位元素,即使UI发生不太大的变化,这种方法也能确保找到正确的元素。主要可以通过元素的id、xpath、text、content-desc等有意义的属性来定位。

实验记录了每个App的UI自动化操作所需要的元素选择器,其中有些是每个App都需要的,包括搜索界面入口、搜索输入框、搜索确认按钮、若干内容条目入口等。此外,根据各App不同的界面布局和功能,还有一些元素为可选项,包括页面跳转按钮、搜索结果过滤按钮、点赞按钮、收藏按钮、加入购物车按钮、个性化推荐界面入口等。

2.3 目标数据注入

目标数据注入旨在让App记录预先设计的关键词,将该关键词所代表的标签作为用户偏好上传到第三方,并与其他App共享,这样则有可能在其他App中检测出来。目标数据注入步骤如下。

步骤1 启动App。利用Appium根据被注入应用的appPackage和appActivity的值唤起该应用的启动进程。

步骤2 进入搜索界面。如果App首页存在搜索按钮,则直接点击;如果没有,程序会尝试寻找并点击能跳转到有搜索按钮的界面的元素,再点击搜索按钮,最终进入搜索界面。

步骤3 输入关键词并搜索。在搜索界面中找到输入框,输入预先设计的关键词。如果存在搜索确认按钮,程序会自动点击;否则,程序会唤起输入法,按下软键盘上的搜索按钮,同样能达到确认搜索的目的。

步骤4 处理搜索结果。进入搜索结果页面后,如果某些App的搜索结果比较杂乱,需要过滤,程序会尝试点击具有过滤结果作用的按钮,从而只显示某一类型的条目,并使其布局更加整齐,方便后续的条目定位。

步骤5 浏览内容并进行互动。逐一打开搜索结果中的文章、帖子、商品、视频等内容,根据App是否支持该操作,对内容做出点赞、喜欢、收藏、加入购物车等互动。这些操作可以增加App认为该内容是用户偏好的概率。对每条内容操作完毕后返回上一级,继续打开下一条内容,直到遍历完搜索结果页面中的大部分内容。

步骤6 搜索其余关键词。一个关键词的上述操作执行完毕后,返回搜索界面,继续搜索下一个关键词,即重复步骤3~步骤5,直到所有的关键词都搜索完毕。

步骤7 退出App。

2.4 收集个性化推荐页面

收集个性化推荐页面的目的是记录App展示的个性化推荐内容,获得待检测页面的截图,在之后的流程中分析这些App是否接收了来自第三方的用户偏好数据。个性化推荐页面收集的具体流程如下。

步骤1 启动App。与目标数据注入中的方法相同。

步骤2 进入个性化推荐界面。对于大多数App来说,启动后的主页即个性化推荐内容的展示界面。对于个别个性化推荐内容不在首页的

App, 程序将找到相应的入口并点击, 然后进入个性化推荐界面。

步骤3 浏览内容并截图。为了尽可能多地浏览推荐内容, 需要不断刷新屏幕中展示的信息。对于大多数App来说, 可以通过下滑的方式向下浏览; 而对于一些App, 需要通过下拉后松开的方式刷新推荐内容, 或者退出个性化推荐界面后再重新进入。

综上所述, 每次刷新截取一张屏幕截图, 每个App的每次检测共收集10张内容基本不重复的截图。由于每轮对29个App进行检测, 共进行30轮, 因此最终将得到总计 $30 \times 29 \times 10 = 8\,700$ 张待检测的截图。

3 App潜在数据共享行为分析方案

本文采用光学字符识别 (optical character recognition, OCR) 技术来提取页面截图中的文本信息, 使用多模态视觉大模型对截图内容进行理解, 得到对图像的描述性文本信息。对于上述两种来源的文本信息, 根据数据注入时的关键词进行搜索匹配, 若找到则记录相关信息。此外, 使用大语言模型判断语义关联性, 对关键词和图像描述文本进行深度匹配。最后处理上述3种方式得到的关联信息, 生成App之间的潜在数据共享关系, 如算法2所示。信息提取和关联性分析流程如图3所示。

算法2 App数据共享分析算法

输入 个性化推荐页面截图集合 S

输出 数据共享行为证据集合 E

- (1) 初始化 $E = \emptyset$ 。
- (2) 对于每张截图 $s_i \in S$:
- (3) 执行OCR文本提取, 得到其中的文字内容 $t_i = \text{OCR}(s_i)$;
- (4) 执行图像内容理解, 得到图像内容描述文本 $d_i = \text{ContentUnderstanding}(s_i)$;
- (5) 执行关键词搜索匹配, 若 $\exists k_l \in K$ 在 t_i 或 d_i 中出现:
- (6) 生成数据共享证据 e_i , $E = E \cup \{e_i\}$ 。
- (7) 执行语义关联性判断, 若 $\exists k_l \in K$ 与 d_i 有关联:
- (8) 生成数据共享证据 e_i , $E = E \cup \{e_i\}$ 。

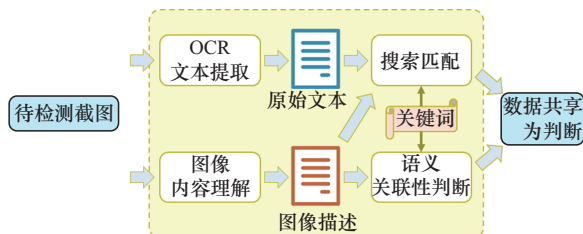


图3 信息提取和关联性分析流程

3.1 OCR文本提取

页面截图中的文字能直观地体现该App推荐的内容, 其中的文本信息如果与事先在其他App中注入的数据产生雷同, 将是App间潜在数据共享最直接的证据。在OCR文本提取任务上, 本文使用PaddleOCR的PP-OCRv4^[28-29]开源模型, 提取待检测截图中的所有文本。本文对全部待检测截图进行了遍历, 将每张截图识别结果分别保存到一个txt文件中, 就完成了图像OCR数据集的构建, 共计8 700个文本文件。

3.2 图像内容理解

App个性化推荐界面中显示的图像信息同样包含了大量的潜在数据共享线索。基于传统机器学习技术的图像物体识别模型通常只能识别出有限的类别, 且粒度较粗, 因此本文选择多模态视觉大模型来完成图像内容理解任务。多模态大模型结合了计算机视觉、自然语言处理等多种技术, 能够对图像中的物体、商品、品牌、人物、风景、行为、专有标识、指示关系等进行细粒度的识别和理解, 减少遗漏细节或过度宽泛的现象。

本文选取开源多模态大模型InternVL 1.5^[30]完成图像内容理解任务, 具体操作是在云主机中部署InternVL 1.5的本地模型, 通过编写Python程序调用该模型, 提供提示词 (Prompt) 依次处理先前收集的图片数据。图4展示了一个图像内容理解模型输出示例。图4左侧为输入的原始图像, 右侧为模型的输出文本, 可见模型的输出能够较好地描述图像内容, 不仅正确识别了图像展示的主体和场景, 还能准确把握图像中的文字、人物穿着、行为、背景等细节信息, 基本将能体现该App推荐内容倾向的信息提取出来。除了推荐条目的图像信息, 模型还能解读App界面布局, 也有助于统计数据共享行为的表现形式。



图4 图像内容理解模型输出示例

本文对全部待检测的8 700张截图进行了遍历, 使用 InternVL 1.5 模型对每张截图进行推理。模型的输出被分别保存到一个txt文件中, 形成了一个包含8 700个文本文件的图像内容理解数据集。

3.3 语义关联性判断

根据App的推荐算法原理, 大多数App的用户偏好是以标签形式分类储存的, 而具体的推荐内容可能属于某一标签、某一主题, 但细节表述有所不同, 其推荐内容与用户偏好标签之间可能不存在直接的文本匹配, 因此还需要引入语义关联性判断, 以识别推荐内容与用户偏好标签之间的深层次联系。

语义关联性判断的基本方法是通过 Prompt 提供一段图像内容描述和一个备选关键词列表, 要求模型判断提供的图像内容描述中是否体现了某个关键词。其中用到的 Prompt 是在程序运行的过程中, 根据当前输入的待检测内容动态生成的, 由3个部分构成: 任务指示要求、图像内容描述和备选关键词列表。具体如下。

(1) 任务指示要求为固定格式, 明确了判断任务的要求和输出格式, 如“根据下面这段手机应用界面的文字描述, 判断应用所显示的内容是否与关键词列表中的某个关键词有联系。若是, 回答与列表中的哪一个关键词有关, 只输出一个关键词, 不要输出任何其他内容; 若否, 仅回答0, 不要输出任何其他内容。”这样的要求能够确保模型只输出简洁明确的判断结果, 通过程序直接进行下一步处理, 避免了不必要的干扰信息。

(2) 图像内容描述即为图像内容理解任务的

输出结果。需要检测哪张截图, 就将其在图像内容理解数据集中对应的文本信息填入这一区域。

(3) 备选关键词列表也是动态变化的, 根据待检测截图产生的轮次确定, 只包括该轮之前注入的关键词。

本文通过调用大语言模型 GLM-4^[31-32] 的公开 API, 对图像内容理解数据集里待检测的8 700个图像内容描述文件进行遍历, 得到了相应的语义关联性判断结果, 为后续数据共享行为的判断提供了充足的证据支持。

3.4 数据共享行为判断

在实验流程上, 考虑时间先后顺序的因果约束, 仅对发生在个性化推荐页面截图之前注入的关键词进行关联性匹配; 对于某一关键词尚未注入就已在个性化推荐页面中展示的偶然情况, 不纳入数据共享行为的判断范围。

首先进行关键词的直接搜索匹配。对于 OCR 数据集和图像内容理解数据集内的每一个文本文件, 本文用 Python 程序遍历其对应截图生成时间之前的关键词列表, 进行精准查找。在搜索匹配时, 如果某个关键词本身搜索不到, 会尝试用它的子关键词进行搜索。子关键词包括原关键词的同义词、近义词、缩写、别称等。即使原关键词没有直接出现在文本中, 其相关表述也可能被识别出来。对于语义关联性判断的结果, 若遍历输出的文本文件数量不为0 (为0代表无任何关联), 则根据其中的关键词找出对应的数据源App。

上述流程全部完成后, 实验从 OCR 数据集、图像内容理解数据集和语义关联性判断结果共26 100条数据中, 找到了1 466条存在数据共享行为嫌疑的证据。

4 实验结果及分析

实验中用到的安卓模拟器是安装在操作主机上的MuMu模拟器。该设备的系统为Android 12, 处理器、显卡、内存和硬盘都与操作主机共用, 分辨率为1 080×1 920 (DPI 480), 模拟手机型号为Samsung Galaxy S22 Ultra。

4.1 数据去重和赋权

在个性化推荐页面收集的流程中, 本文希望每次对一个App收集10张内容不相同的截图。但是在刷新页面内容时, 页面中的有些部分可能

不会随着向下滑动、下拉刷新和退出重进而发生改变。此外,由于存在偶然的网络延迟,页面内容可能并未完全刷新就被截图。这种情况可能造成同样的文本或图像在多张截图中重复出现,在结果中产生连续的多条相似记录,干扰最终结果。为了去除结果中的重复项,实验添加了一个限制条件:在每个 App 每组的 10 张截图中,同一关键词的检出情况最多只记录一次,通过 Python 程序对记录检出结果的 Excel 表格进行自动化批量操作,删除重复的记录。

此外,在进行 App 数据共享行为的判断时,单纯根据检出记录的数量来判断是不合适的,应该考虑不同类型检出记录的可信度和说服力。实验根据检出来源和被检出的关键词类别,将检出记录的类型划分为 5 种。为了体现各种检出类型之间可信度的差异,并且使不同类型的检出记录进行统一表达和对比,对每种类型设定了不同权重,作为从定性到定量的映射机制。各种检出类型的数量和赋予的权重见表 1。如在 OCR 结果中直接搜索到了原关键词,是最有力的证据,其权重就设定得较高;语义关联性判断能发掘关键词和页面内容之间的深度联系,也是很有效的判据,权重也应较高;对于未找到原关键词,只匹配到了子关键词的情况,其说服力就偏弱,权重就较低。这样处理的好处在于能够更加准确地反映出不同检出情况在判断数据共享行为时的相对重要性,提高了检测的可信度和可靠性。在判断 App 数据共享行为时,实验根据各条记录附带的权重进行加和计算,得到一个分数 Score,作为得出结论的最终判断指标,其计算方法如下:

$$\text{Score} = \sum_i w_i \cdot |e_i|$$

其中, $|e_i|$ 为某检出类型的数量, w_i 为对应的权重。

为进一步分析各类检测方法在整体结果中的作用,表 1 统计了不同类型检出记录的数量及其占比。不同检出类型在数据集中均有一定覆盖,表明各类检测方法均发挥了作用,而非由单一来源主导。从结果分布角度看,高权重类型(如原关键词直接匹配和语义关联)对应的检出数量虽相对较少,但其关联的确定性更强;低权重类型

数量较多,但单条证据的判别能力相对有限。该分布特征与赋权设计中“高置信度低频、低置信度高频”的基本假设一致,在一定程度上从数据层面支持了当前赋权方案的合理性。

表 1 各种检出类型的数量和赋予的权重

检出来源	匹配到的关键词类型	数量(占比)	赋予权重
OCR	原关键词	142 个 (9.7%)	3
	子关键词	465 个 (31.7%)	1
图像内容理解	原关键词	108 个 (7.4%)	2
	子关键词	461 个 (31.4%)	1
语义关联	/	290 个 (19.8%)	3

4.2 检测结果验证

要评估模型的输出结果正确与否,需要提供每条数据的真实标签。然而,检测结果的数据量过于庞大,因此本文采取随机抽样的方式,从数据集中随机抽取 600 条样本进行人工打标判断,将模型的判断结果按照二分类的指标进行映射,检测结果验证样本集的分类统计见表 2。

表 2 检测结果验证样本集的分类统计

检测结果	人工判断结果	对应类别	样本数量/个
关键词 X	关键词 X	真正例	258
	无关联或关键词 Y	假正例	42
无关联	无关联	真负例	295
	关键词 X	假负例	5

根据二分类混淆矩阵的数值计算,本文检测方法的检测结果验证样本集的评估指标见表 3。可见检测方案的整体效果较好,在大多数情况下都能做出正确的预测。模型的召回率很高,说明对确有关联的数据共享行为漏报较少,这在需要高度敏感性的应用场景中更为重要。验证样本集中出现了少量假正例,原因可能是截图内容存在与多个关键词相关的语义信息,模型难以完全区分特定关键词,或者图像过于复杂导致模型产生误判。这些假正例在整体数据集中占比较小,多源证据融合的赋权机制也有助于减弱单条假正例对最终判断的干扰,对检测结果影响有限。

表3 检测结果验证样本集的评估指标

评估指标	指标数值
准确率	92.17%
精确率 (precision)	86.00%
召回率 (recall)	98.10%
F1 分数 (F1 score)	91.65%
漏报率 (false negative rate, FNR)	1.90%
误报率 (false positive rate, FPR)	12.46%

4.3 挖掘结果统计分析

判断依据的原理如下: 在没有任何用户偏好的情况下, App展示的推荐内容应该是随机的、广泛的。如果没有外部因素的影响, 用户在不同App中看到的推荐内容应该是多样化的, 没有特定的偏向性。然而, 如果用户在某个App中提供了用户偏好, 并且这些偏好在另一个App的个性化推荐内容中得到了体现, 那么就有充分的理由相信这两个App之间存在数据共享行为。

通过对比关键词注入前后的推荐页面发现, 在某一关键词被注入前, 相关内容出现频率较低, 而被注入后在多个App中显著集中出现, 该现象难以完全由随机推荐或热门内容分发机制解释。

4.3.1 数据提供和接收倾向

在进行App潜在数据共享行为的检测时, 本文分别统计了数据提供和接收的情况。有些App可能主要作为数据提供者向其他App共享用户数据; 而有些App可能主要作为数据接收者, 从其他App获取用户数据。

图5展示了App提供数据倾向的分数, 该值越大表示向其他App提供数据越多。可见, 前4个App的此类行为较多, 包括社交平台类的微博和百度贴吧、影音娱乐类的快手和西瓜视频。

图6展示了App接收数据倾向的分数, 该值越大表示从其他App获取数据越多。可见, 新闻阅读类App的此类行为较为频繁, 如央视新闻、今日头条等。

数据提供和接收的情况对于评估App的隐私保护措施和合规性具有重要意义。如果一个App向其他App共享了大量的用户数据, 该App需要特别注意遵守隐私政策, 确保用户数据的共享是合法、合规的, 并且得到了用户的明确同意。如果一个App从其他App获取了大量的用户数据, 该App需要特别注意遵守第三方共享清单, 确保接收的用户数据在清单范围之内, 并且在使用这些数据时遵守相关的隐私保护规定。

为了更深入地揭示App数据共享行为与其所属类别之间的关系, 本文将30个被测App分为社交平台、影音娱乐、新闻阅读、在线购物和生活服务5个类别, 分别统计各类别App提供和接收数据的总和。由于各类别内的App数量并不一致, 为了保证分析时的可比性, 本文计算了每个类别的平均带权分数, 即总分数除以该类别内的App数量。按类别划分的App提供和接收数据倾向的平均分数及所占比例如图7所示。

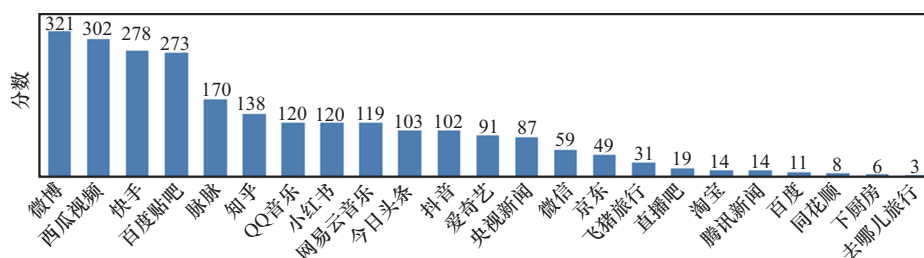


图5 App提供数据倾向的分数

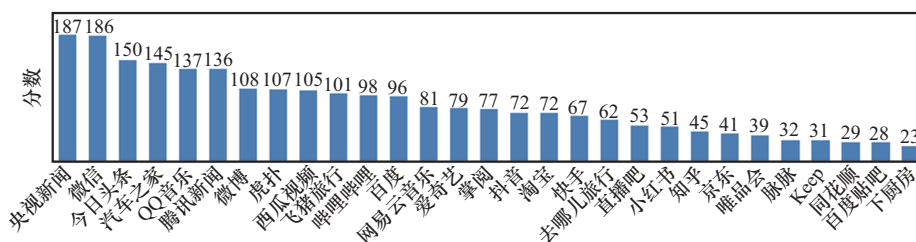


图6 App接收数据倾向的分数

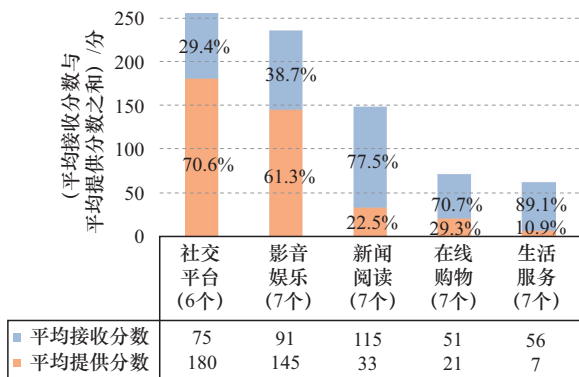


图7 按类别划分的App提供和接收数据倾向的平均分数

根据实验结果可知,各个类别App的数据共享活跃度存在差异,其中社交平台和影音娱乐类App的潜在数据共享行为更为频繁,而在线购物和生活服务类App的此类行为较少。在数据提供和接收的差异方面,社交平台和影音娱乐类App更倾向于收集自己App的用户偏好数据,然后共享给第三方;而新闻阅读、在线购物和生活服务类App则更倾向于从第三方处获取用户数据,然后在自己的App中向用户展示相关的推荐内容。各类别App在数据提供和接收倾向上的差异,主要源于各自性质和功能不同,也反映了App在数据安全和隐私保护方面的不同考量。

4.3.2 数据共享方向和关系

除了对单个App的分析,还需要关注用户数据在两个App之间的流向。单向的数据传递指的是数据从一个App流向另一个App,而没有反向流动,一般发生在App之间存在明确的数据共享协议或者一个App需要依赖另一个App提供的资源时。App单向数据传递倾向的分数如图8所示,微博、微信等社交类App存在将用户的社交行为数据传递给新闻阅读和生活服务类App的可能性,使后者能提供更精准的广告投放。

双向的数据共享指的是数据在两个App之间双向流动,通常发生在App之间存在合作关系或者共同开发服务时。App双向数据共享倾向的分数如图9所示。由图9可知,西瓜视频和今日头条属于同一公司的App,可能存在高度的数据互通;网易云音乐和QQ音乐同属音乐类App,在内容风格方面较为相似,可能存在数据关联。

4.3.3 数据共享网络

App之间的数据共享行为错综复杂,构成了一个庞大的网络。这个网络由许多数据共享路径和节点组成,每个节点代表一个App,而边则代表App之间的数据共享关系,构成一个有向图。一个App可能与多个其他App共享数据,而这些App之间也

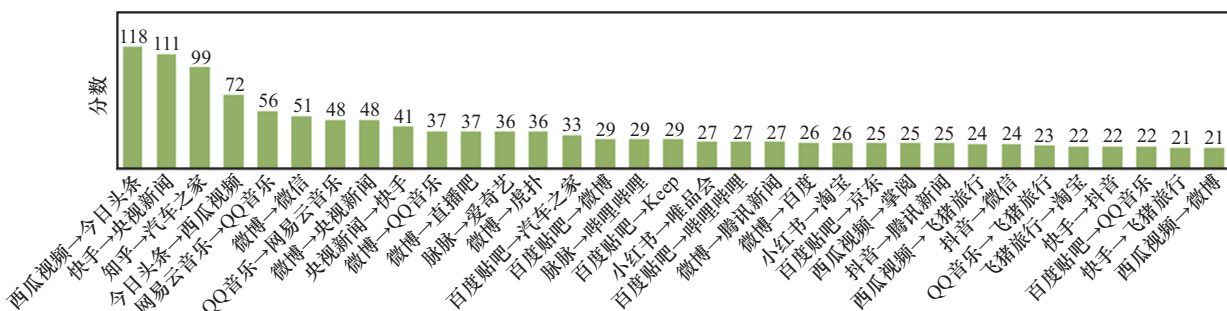


图8 App单向数据传递倾向的分数

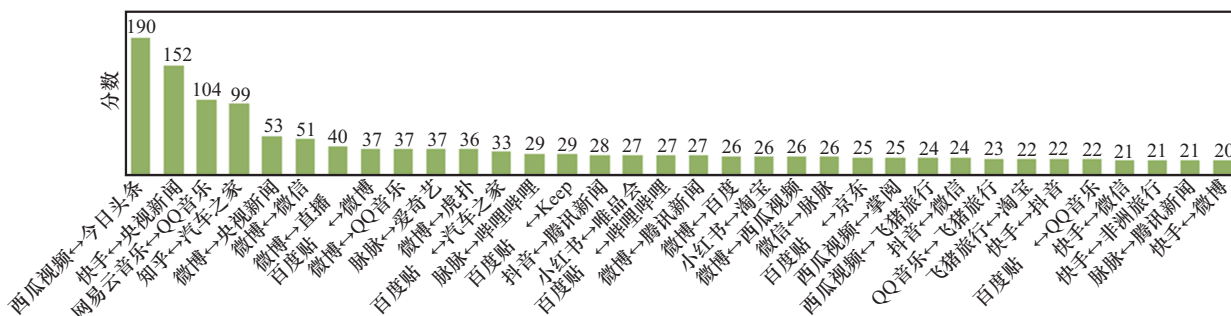


图9 App双向数据共享倾向的分数

可能存在数据共享关系。实验得到的App数据共享网络如图10所示。图10中节点的大小和颜色表示节点的度,与其他App检出数据共享越多,节点越大、颜色越深;边的粗细和深浅表示边的权重,两个App之间数据共享的分数越高,边就越粗。为了图像清晰起见,本文在绘图时省略了权重较小的边。根据图10可知,百度贴吧、微博、微信这3个社交平台类App在网络中与其他App建立联系的现象最多,可看作是数据共享网络中的关键节点。

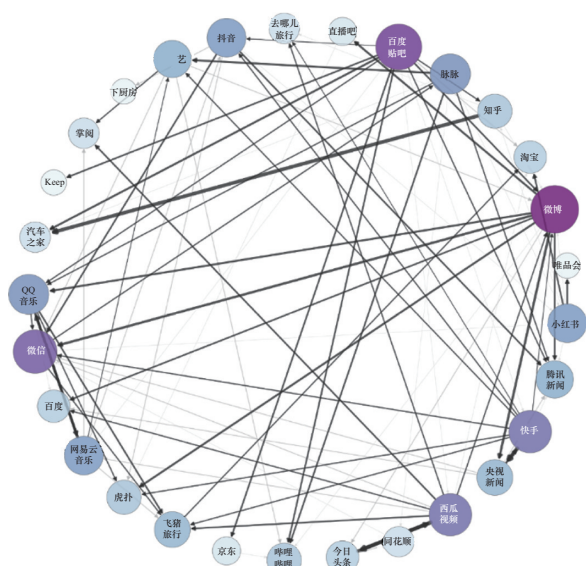


图10 App数据共享行为网络

4.4 可扩展性分析

本文仅在一定范围内的App上进行实验,但所提出的App数据共享行为挖掘和分析方案具有良好的可扩展性,主要体现在以下几个方面。

(1) 规模具有扩展性。自动化框架基于程序驱动方式实现,可通过批量任务调度对大批量App进行并行测试,理论上可扩展至更大规模的App集合,仅受限于计算资源与执行时间。

(2) 环境具有适应性。本文方法仅采集前端界面,不依赖特定设备或底层系统接口,因此可在模拟器与各种型号的真实设备中运行。

(3) 模型具有替换与升级能力。在信息提取与分析环节,所采用的OCR与大语言模型均为模块化设计,可随着底层模型发展替换为性能更优或适配性更强的模型,提升整体检测效果。

4.5 研究价值与应用场景

需要说明的是,尽管许多App在隐私政策或

用户协议中对第三方数据共享行为进行了概括性声明,但相关声明通常以宽泛条款呈现,难以反映数据行为的实际情况。因此,隐私政策中的披露内容并不能替代对应用实际行为的技术性验证。本文所提出的检测方法可作为对App隐私声明的补充手段,用于评估其前端呈现行为是否与声明内容保持一致,从而为隐私合规性审查提供可观测、可验证的依据。此外,用户普遍感知到的“在某一App中出现的兴趣内容会在其他App中被推荐”现象,更多源于长期使用过程中的主观体验与直觉判断,缺乏系统性的实验验证和量化分析。本文对上述现象进行了结构化建模与客观检验,从技术层面对用户体验性认知给出了可重复的实证解释。

从应用价值角度看,本文方法并非旨在简单证明App之间的数据共享现象是否存在,而是提供了一种在不依赖源码分析、流量分析或自主披露的前提下,对潜在数据共享行为进行自动化挖掘与量化评估的技术路径。该方法可为监管部门开展技术抽检、App开发企业进行合规自查以及隐私保护研究提供辅助分析工具,具备一定的现实适用性和研究价值。

5 结束语

本文针对App潜在数据共享行为的检测问题,提出了一种智能挖掘方法,通过App内的自动化行为模拟和呈现界面挖掘、深入的信息提取和关联性分析,实现了对App潜在数据共享行为的有效识别。本文的研究成果为移动应用隐私保护领域带来了新的视角和解决方案。

后续的研究将考虑继续引入人工智能方法,解决自动获取UI界面元素的问题,可通过引入真实设备或多设备环境,对不同运行条件下的检测结果进行对比分析;可实施控制检测时间点等变量的对比实验,找出检测数据流转的最佳时间窗口;可结合App的第三方信息共享清单,实现共享行为合规性检查的完整自动化流程;还可以探索将本文成果集成到在线隐私检测平台,尝试实现应用落地,为企业、监管部门、个人开发者等主体提供App自查和监察服务,辅助检测和评估App的第三方数据共享行为,提高全行业的隐私保护水平。

参考文献:

- [1] 中央网络安全和信息化委员会办公室. APP 违法违规收集使用个人信息专项治理报告(2019)[R]. 2020.
Office of the Central Cyberspace Affairs Commission. Report on the special governance of illegal and violative collection and use of personal information by Apps(2019)[R]. 2020.
- [2] The European Parliament and Council of the European Union. 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 Directive repealing 95/46/EC (General Data Protection Regulation)[EB]. 2016.
- [3] 中国人大网. 《中华人民共和国个人信息保护法》[EB]. 2021.
The Website of the National People's Congress of China. The Personal Information Protection Law of the People's Republic of China [EB]. 2021.
- [4] Quermann N, Degeling M. Data sharing in mobile apps: user privacy expectations in Europe[C]//Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). Piscataway: IEEE Press, 2020: 107-119.
- [5] Wang Z H, Li Z Y, Xue M H, et al. Exploring the eastern frontier: a first look at mobile app tracking in China[C]//Passive and Active Measurement. Cham: Springer, 2020: 314-328.
- [6] Zang J, Dummit K, Graves J, et al. Who knows what about Me? a survey of behind the scenes personal data sharing to third parties by mobile apps[J]. Technology Science, 2015, 30: 2.1.
- [7] Jin H J, Liu M Y, Dodhia K, et al. Why are they collecting my data: inferring the purposes of network traffic in mobile apps[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 2(4): 1-27.
- [8] Binns R, Zhao J, Van Kleek M, et al. Measuring third-party tracker power across web and mobile[J]. ACM Transactions on Internet Technology, 2018, 18(4): 1-22.
- [9] Razaghpanah A, Nithyanand R, Vallina-Rodriguez N, et al. Apps, trackers, privacy, and regulators: a global study of the mobile tracking ecosystem[C]//Proceedings 2018 Network and Distributed System Security Symposium. Internet Society, 2018.
- [10] 王竹, 贺坤, 王新宇, 等. Android 设备中基于流量特征的隐私泄露评估方案[J]. 通信学报, 2020, 41(2): 155-164.
Wang Z, He K, Wang X Y, et al. Traffic characteristic based privacy leakage assessment scheme for Android device[J]. Journal on Communications, 2020, 41(2): 155-164.
- [11] Fan M, Yu L, Chen S, et al. An empirical evaluation of GDPR compliance violations in Android mHealth apps[C]//Proceedings of the 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE). Piscataway: IEEE Press, 2020: 253-264.
- [12] Guamán D S, Del Alamo J M, Caiza J C. GDPR compliance assessment for cross-border personal data transfers in Android apps[J]. IEEE Access, 2021, 9: 15961-15982.
- [13] 李昕, 唐鹏, 张西珩, 等. 面向 GDPR 隐私政策合规性的智能化检测方法[J]. 网络与信息安全学报, 2023, 9(6): 127-139.
Li X, Tang P, Zhang X H, et al. GDPR-oriented intelligent checking method of privacy policies compliance[J]. Chinese Journal of Network and Information Security, 2023, 9(6): 127-139.
- [14] 张西珩, 李昕, 唐鹏, 等. 基于知识图谱的隐私政策合规性检测与分析[J]. 网络与信息安全学报, 2024, 10(6): 151-163.
Zhang X H, Li X, Tang P, et al. Privacy policy compliance detection and analysis based on knowledge graph[J]. Chinese Journal of Network and Information Security, 2024, 10(6): 151-163.
- [15] Xiao J M, Chen S Z, He Q, et al. An Android application risk evaluation framework based on minimum permission set identification [J]. Journal of Systems and Software, 2020, 163: 110533.
- [16] Ren J J, Rao A, Lindorfer M, et al. ReCon: revealing and controlling PII leaks in mobile network traffic[C]//Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services. New York: ACM, 2016: 361-374.
- [17] Enck W, Gilbert P, Han S, et al. TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones[J]. ACM Transactions on Computer Systems, 2014, 32(2): 1-29.
- [18] Liu Y, Liao L J, Song T. Static tainting extraction approach based on information flow graph for personally identifiable information [J]. Science China Information Sciences, 2020, 63(3): 132104.
- [19] 王新宇, 牛犇, 李凤华, 等. APP 隐私泄露风险评估与保护方案 [J]. 通信学报, 2019, 40(5): 13-23.
Wang X Y, Niu B, Li F H, et al. Risk assessing and privacy-preserving scheme for privacy leakage in APP[J]. Journal on Communications, 2019, 40(5): 13-23.
- [20] Shuba A, Markopoulou A. NoMoATS: towards automatic detection of mobile tracking[J]. Proceedings on Privacy Enhancing Technologies, 2020, 2020(2): 45-66.
- [21] Gálvez R, Moonsamy V, Diaz C. Less is More: a privacy-respecting Android malware classifier using federated learning[J]. Proceedings on Privacy Enhancing Technologies, 2021, (4): 96-116.
- [22] Arzt S, Rasthofer S, Fritz C, et al. FlowDroid: precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for Android apps[J]. ACM SIGPLAN Notices, 2014, 49(6): 259-269.
- [23] Bakopoulou E, Shuba A, Markopoulou A. Exposures exposed: a measurement and user study to assess mobile data privacy in context[PP]. arXiv Preprint, 2022, arXiv: 2008.08973.
- [24] Van Kleek M, Llicardi I, Binns R, et al. Better the devil you know: exposing the data sharing practices of smartphone apps[C]//Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2017: 5208-5220.
- [25] Wu J R, Nan Y H, Xing L Y, et al. Leaking the privacy of groups and more: understanding privacy risks of cross-app content sharing in mobile ecosystem[C]//Proceedings 2024 Network and Distributed System Security Symposium. Internet Society, 2024.
- [26] 韩震, 石磊, 赵彦霖. 考虑不同使用场景的 APP 用户隐私交换行为[J]. 系统管理学报, 2022, 31(2): 353-361.
Han Z, Shi L, Zhao Y L. Privacy exchange behavior of APP Users' Considering different usage scenarios[J]. Journal of Systems & Management, 2022, 31(2): 353-361.
- [27] 广东省消委会消费维权网. 广东省消委会发布“常见类型 App 非必要个人信息收集情况”调查结果, 呼吁经营者依法依规收集使用消费者个人信息[EB]. 2021.
Website for Consumer Rights Protection of Guangdong Province

Consumer Council. The Consumer Council of Guangdong Province has released the results of a survey on the collection of non-essential personal information by common types of Apps, calling on operators to collect and use consumers' personal information in accordance with the law and in a standardized manner[EB]. 2021.

- [28] Li C X, Liu W W, Guo R Y, et al. PP-OCRv3: more attempts for the improvement of ultra lightweight OCR system[PP]. arXiv Preprint, 2022, arXiv: 2206-01300.
- [29] PaddlePaddle.PP-OCRv4 技术报告[EB]. 2025.
PaddlePaddle.PP-OCRv4 technical report[EB]. 2025.
- [30] Chen Z, Wu J N, Wang W H, et al. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks[C]// Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 24185-24198.
- [31] Du Z X, Qian Y J, Liu X, et al. GLM: general language model pre-training with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Kerrville: Association for Computational Linguistics 2022: 320-335.
- [32] Zeng A H, Liu X, Du Z X, et al. GLM-130B: an open bilingual pre-trained model[C]//Proceedings of the Eleventh International Conference on Learning Representations. Kigali: ICLR, 2023.

[作者简介]



刘凡嘉 (2002-), 男, 上海交通大学硕士生, 主要研究方向为人工智能安全、隐私保护。



陈裕鑫 (2002-), 男, 上海交通大学硕士生, 主要研究方向为隐私安全、人工智能安全。



王杨德 (1987-), 男, 上海交通大学博士生, 主要研究方向为计算机取证、口令安全。



唐鹏 (1992-), 男, 上海交通大学助理研究员, 主要研究方向为数据要素流通共享安全、隐私保护。



邱卫东 (1973-), 男, 博士, 上海交通大学教授、博士生导师, 主要研究方向为密码分析、数据要素流通共享安全、隐私保护。