无线网络用户的Wi-Fi指纹匿名化研究

韩秀萍[1] 王 智[1] 裴 丹[2]
(清华大学深圳研究生院信息科学与技术学部 广东 深圳 518055)[1]
(清华大学计算机科学与技术系 北京 100084)[2]

摘 要 如今,上亿的Wi-Fi热点被广泛部署,用于给人们提供Wi-Fi连网服务。为了加快Wi-Fi连接的速度,移动设备会发送探测请求帧来发现附近的无线热点,并且保存曾经连接过的AP的SSID,即首选网络列表(PNL)。已有研究表明,由探测请求帧发出的SSID构成的Wi-Fi指纹会泄露用户的隐私信息。基于对现实情况中Wi-Fi指纹所造成的隐私泄露程度的分析,提出了数据驱动的隐私保护方案。首先,针对4个城市中2700万用户连接400万Wi-Fi热点的行为进行了测量研究,并证明了在很多场景下Wi-Fi指纹都可以用来区分用户。基于对Wi-Fi指纹中SSID语义信息的研究,可以推断出这些用户的身份信息(如工作信息)。其次,提出了一种基于协同过滤的启发式方法,它通过给用户的PNL中添加伪SSID来模糊其信息,并使得附近的人彼此之间的PNL与Wi-Fi指纹都更加相似。最后,基于真实的Wi-Fi连接数据验证了上述策略的有效性,实验结果表明,修改PNL不仅能保护用户隐私,而且能保证快速的Wi-Fi连接。

关键词 无线网络,隐私泄露,保护,探测请求帧,用户行为
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.08.002

Study on Wi-Fi Fingerprint Anonymization for Users in Wireless Networks

HAN Xiu-ping[1] WANG Zhi[1] PEI Dan[2]
(Department of Computer Science and Technology, Graduate School at Shenzhen, Tsinghua University, Shenzhen, Guangdong 518055, China)[1]
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)[2]

Abstract Billions of Wi-Fi assess points (APs) have been deployed to provide wireless connection to people with different kinds of mobile devices. To accelerate the speed of Wi-Fi connection, mobile devices will send probe requests to discover nearby Wi-Fi APs, and maintain previously connected network lists (PNLs) of APs. Previous studies show that the Wi-Fi fingerprints that consist of probed SSIDs individually will leak private information of users. This paper investigated the privacy caused by the Wi-Fi fingerprints in the wild, and provided a data-driven solution to protect privacy. First, measurement studies were carried out based on 27 million users associating with 4 million Wi-Fi APs in 4 cities, and it was revealed that Wi-Fi fingerprints can be used to identify users in a wide range of Wi-Fi scenarios. Based on semantic mining and analysis of SSIDs in Wi-Fi fingerprints, this paper further inferred demographic information of identified users (e.g., people's jobs), telling "who they are". Second, this paper proposed a collaborative filtering (CF) based heuristic protection method, which can "blur" an user's PNL by adding faked SSIDs, such that nearby users' PNLs and Wi-Fi fingerprints are similar to each other. Finally, the effectiveness of the design was verified by using real-world Wi-Fi connection traces. The experiments show that the refined PNLs protect users' privacy while still provide fast Wi-Fi reconnection.

Keywords Wireless network, Privacy leakage, Protection, Probe request frame, User behavior

1 引言

如今,无线网络已经成为一种基础设施,根据思科公司的预测,全球有超过600万的公共热点在提供无线网络连接。为了加快Wi-Fi连接的速度,移动设备通常会保存它曾经连接过的Wi-Fi网络的服务集标志符(SSID),而这些SSID会通过探测请求帧发送出去[1]。据研究,移动设备每秒最多能发送50个探测请求帧,而其中多达98%的帧中都包含SSID信息[2]。

本文将通过探测请求帧发送的SSID的集合称为用户的Wi-Fi指纹,它主要会造成两种隐私泄露。一方面,SSID通常包含语义信息,比如"小明的Macbook""Starbucks-WiFi""Company XXX net"等,通过这些字符往往能推断出用户曾经的去处。Chernyshev等[3]提出,49%的SSID都是可识别的,并且能提供这些设备的机主的潜在信息。另一方面,已有研究表明,用户的偏好[4]、身份识别[5]等都可以通过Wi-Fi指纹来推断,而攻击者也可以以此推断出用户的轨迹[6]。

到稿日期:2017-10-24 返修日期:2017-12-21

韩秀萍(1993-),女,硕士,主要研究领域为用户行为分析、数据挖掘,E-mail:hxp15@mails.tsinghua.edu.cn;王 智(1985-),男,博士,讲师,主要研究领域为多媒体内容分发、移动云计算、大范围多媒体系统,E-mail:wangzhi@sz.tsinghua.edu.cn(通信作者);裴 丹(1973-),男,博士,副教授,主要研究领域为计算机网络,E-mail:peidan@tsinghua.edu.cn。

针对用户隐私信息的保护,目前主要有两种策略。1)减少探测请求帧中 SSID 的发送。Bonne 等[7]设计了一种通过限制移动设备发送 SSID 的数量来保护隐私的方案。这类方法通常需要修改无线协议[8-9]。2)采用 MAC 地址随机化的方法。比如,iOS 在 Wi-Fi 扫描阶段采用了 MAC 地址随机化的方法来避免攻击者找到用户真实的 MAC 地址。

上述方法的缺陷在于:在某些情况下,攻击者总是能或多或少地获得用户曾经连接过的真实的 SSID,从而推断出用户的隐私。鉴于此,本文提出一种启发式的想法,即向用户设备的 PNL 中添加伪 SSID,这样探测请求帧中就会包含真实的与虚假的 SSID,使得攻击者即使获得了用户的 PNL,也依然无法获得用户真实的信息。文中称该方法为用户 Wi-Fi 指纹的匿名化。这个方法看似简单,但在实际设计与实施时存在很大挑战:1)如何挑选出能最大化降低用户隐私泄露的合适 SSID 并添加到目标用户的 PNL 中;2)面对用户在不同地点间不断移动的情形,应如何修改用户的 PNL。

针对上述问题,本文提出如下解决方案:首先,利用大量数据分析了由 Wi-Fi 指纹造成的隐私问题,并对隐私保护方案的可行性进行了细致分析;其次,根据测量的结果,提出一种基于协同过滤的保护算法,通过向用户的 PNL 中添加伪 SSID 来提高用户与邻近用户 PNL 的相似度;最后,通过真实数据来验证算法的有效性;最后总结全文。

本文第 2 节概述了相关工作;第 3 节介绍了具体的背景知识、数据集以及相关的测量工作;第 4 节提出了基于协同过滤的保护策略;第 5 节分析了不同的影响因素对所提算法的影响,并验证了其有效性;最后总结全文。

2 相关工作

2.1 SSID 泄露

近年来,很多学者从不同角度研究了 SSID 泄露的情况,譬如隐私、社交网络、人类行为等。具体的研究成果包括识别用户的设备[10-11]、推断用户的社交关系[12-13]、刻画用户的偏好[4]、分析用户群体的信息[14]等。这些研究提供了很多 SSID 信息的分析与应用,但是他们并没有提出任何相关的隐私保护策略。嗅探 SSID 能导致各种各样的攻击,比如双面恶魔攻击[15]、中间人攻击[16]等。但对于用户而言,删除其移动设备中的 PNL 极其不便,也是不切实际的。

2.2 隐私保护策略

很多研究者致力于 Wi-Fi 探测请求帧的研究,并提出了不同的保护策略。Bonne 等[7]设计了一个系统来限制移动设备中 SSID 的发送,即当设备处于深度睡眠模式时,禁止设备发送包含射程之外的 SSID 的探测请求帧。Lindqvist 等[8]提出了一种新的 AP 发现协议,该协议采用加密的方法来实现探测请求帧与探测相应帧的交互,显然这一对无线协议的修改会导致巨大的成本。Kim 等[9]利用 GPS 信息来减少 SSID 的发送,移动设备只能发送附近 AP 的 SSID,但这种方法在室内的环境下可能会失效。

此外,一些运营商采用了 MAC 地址随机化的方法来保护用户隐私。iOS 9 在其设备进行 Wi-Fi 扫描的阶段使用了 MAC 地址随机化的方法[17];当驱动与硬件支持时,Android 6.0 也采用了 MAC 地址随机化的方法[18];Windows 10[19],Linux kernel version 3.18[20]也都采用了该方法。然而,这些随机化的方法并不相同,即目前还没有一个 MAC 地址随机化的标准规则,这使得其有效性遭到质疑;同时,即使采用了 MAC 地址随机化,移动设备在一些场景下也依然会发送含有真实 MAC 地址与真实 SSID 的探测请求帧(比如,当 iOS 设备收到附近的曾经连接过的 AP 发出的信号时)。攻击者也可以通过字典攻击[21]来诱导移动设备泄露其 PNL。为了解决这个问题,本文提出在 MAC 地址随机化的基础上向 PNL 中添加伪 SSID 来模糊其 PNL,从而保护用户的隐私。

3 背景与测量研究

3.1 背景

在 Wi-Fi 连接的过程中,每个无线热点被其 BSSID 与 SSID 所标识,而每个移动设备被其 MAC 地址标识。移动设备发现附近的 AP 有两种方式:主动扫描与被动扫描。在被动扫描中,AP 主动地、周期性地发送信号帧,移动设备通过在每个信道上捕捉这些帧来建立连接。而主动扫描中,移动设备主动发送探测请求帧来寻找附近可连接的 AP,AP 也会通过探测响应帧来回复,经过验证之后,该 AP 与该设备就建立了连接。

从 SSID 的角度来说,被动扫描相对来说更不易泄露隐私,但如今主动扫描更加流行。其原因是,被动扫描会消耗更多的连接时间,而且隐藏 AP 只能通过主动扫描来发现。目前看来,抛弃隐藏 AP 与主动扫描这一机制的可行性并不高。

对于攻击者来说,当移动设备发送探测请求帧时,他在其周围利用无线网络嗅探的工具来捕捉这些帧并获取用户曾经连接过的 SSID,最终获取用户的轨迹等隐私信息。

3.2 数据集

本文的实验主要基于两个数据集:1)Wi-Fi 连接的数据集;2)Wi-Fi 地址数据集。这两个数据集通过一个在线的 Wi-Fi 连接软件(类似 Wi-Fi 万能钥匙)在中国的 4 个城市收集了为期一个月的信息。其中,Wi-Fi 连接数据集主要包括连接的时间戳、匿名化的用户 ID、SSID、BSSID 等。因为 PNL 即为曾经连接的 Wi-Fi 列表,所以用该软件收集的用户在一个月内连接的 SSID 作为用户的 PNL。因为攻击者只能在某个地点的某个时间内获取用户部分的 PNL,所以本文利用一天内连接过的 SSID 来模拟 Wi-Fi 指纹(即设备泄露的 SSID 集合),而它与真实世界中获取的 SSID 集合的大小是一致的[3,13]。在本文的数据集中,97% 的用户的 PNL 都不超过 20,其 Wi-Fi 指纹的大小不超过 4。

Wi-Fi 连接数据集包括 0.27 亿用户的共 2.5 亿条 Wi-Fi 连接记录,而 Wi-Fi 地址数据集则包括 400 万 Wi-Fi 热点的具体地理位置,具体包括经纬度、详细的街道地址以及位置类型(如医院、酒店等)。在该数据集中,共有 16 种位置类型。表 1 列出了用户 PNL 中至少包含一个该类型 SSID 的用户比例。

表 1 数据集中的位置类型

Table 1 Location types in dataset

No.	位置类型	用户/%	No.	位置类型	用户/%
1	企业	17.53	9	组织机构	4.57
2	医疗保健	32.97	10	汽车	2.66
3	基础设施	6.66	11	生活服务	3.96
4	娱乐休闲	14.10	12	美食	4.80
5	房产小区	78.00	13	购物	46.25
6	教育	21.02	14	运动	2.01
7	文化	6.25	15	酒店	23.78
8	旅游景点	19.69	16	金融	0.34

其中，“房产小区”是最广泛的，有 78.00％的用户都曾连接过该类型的 AP（房产小区可能是住宅小区或者是商务楼宇）；有 46.25％的用户连接过购物中心的 AP。本文利用 PNL 中位置类型的多样性来刻画用户画像（profile）的丰富度。

3.3　算法的动机及可行性分析

已有研究成果已经证明，Wi-Fi 指纹可以用来刻画并区分用户。本文旨在提出一种保护隐私泄露的方法，即通过添加伪 SSID 来模糊用户的 PNL，实现对 Wi-Fi 指纹的匿名化。

SSID 可能包含丰富的语义信息，比如 HuaweiCompany-1F 表示这个用户在该公司工作，这表明即使没有其他信息，攻击者也依然可以仅根据几个 SSID 就获得用户的偏好（比如你喜欢去什么酒吧）。通过测量发现，超过 90％的用户在其 PNL 中包含不超过 6 个位置类型，54.9％的用户仅包含 1 至 2 个位置类型，这表明用户的画像缺乏多样性，且这些地址常对应着用户的住处或工作场所[22]。若向用户 PNL 中添加更多的 SSID，则可以丰富用户画像，并模糊用户的真实偏好与身份信息。

通过上述两个数据集，本文分析了 SSID 造成的隐私泄露。在本文中，当一个用户的 Wi-Fi 指纹与周围其他人存在不同时，该用户可以被区分开。通过分析发现，54.60％的用户能够被 Wi-Fi 指纹区分开，随着时间的延长，用户也会泄露更多的 SSID，从而导致被区分的用户比例上升至 90.58％。

综上可以看出，用户隐私的泄露（用户被刻画或被区分开）主要源于其 PNL 的差异性。因此，若通过添加伪 SSID 使不同用户的 PNL 完全相同，则攻击者就无法利用探测请求帧去窥探用户的隐私。然而，PNL 的长度是有限制的，并且会动态更新[14]，因此这个想法是不可行的。

如果选择降低用户 PNL 的差异性，而非消除其差异性，可使上述想法变得更加可行。换言之，通过添加伪 SSID 到用户的 PNL 中来提高用户彼此之间的相似度，从而降低用户被 Wi-Fi 指纹区分开的概率。为了验证该想法，图 1 中分析了用户 PNL 相似度与其被 Wi-Fi 指纹区分开的概率之间的关系。这里，PNL 的相似度利用 Cosine-IDF[6] 来定义，公式如下：

$$C(S_u,S_v)=\frac{\sum\limits_{s\in S_u\cap S_v}(\log(\frac{1}{f_s}))^2}{\sqrt{\sum\limits_{s\in S_u}(\log(\frac{1}{f_s}))^2}\sqrt{\sum\limits_{s\in S_v}(\log(\frac{1}{f_s}))^2}} \tag{1}$$

$$f_s=\frac{|U|}{|U_s|} \tag{2}$$

其中，S_u 是用户 u 的 SSID 集合（即 PNL），而 f_s 是 SSID s 的流行度，$|U|$ 是当前地区用户的总数，而 $|U_s|$ 是曾经连接过 SSID s 的用户数目。该相似度主要考虑了不同用户 PNL 的交集与每个 SSID 的流行度。如果一个用户与他人都连接过同一个 SSID，则表明二者的相似度大于 0，称其为“邻居”用户。而有着最高相似度的邻居被称为“兄弟”用户。这里主要关注用户与其兄弟之间的 PNL 相似度。

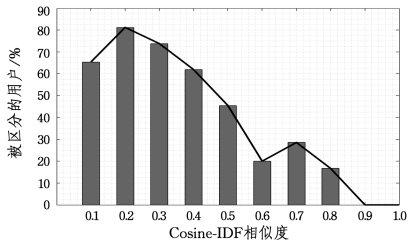


图 1　用户与其兄弟用户之间的 PNL 相似度 vs. 用户被区分的概率

Fig. 1　PNL similarity among siblings vs. probability of being identified

图 1 中，用户依其 PNL 的相似度被划分成 10 类，从[0，0.1]到[0.9，1]。可以观察到，与他人有着高相似度的用户更不容易被区分开，即用户的 PNL 相似度与其被区分开的概率之间有着强相关且成反比。当相似度高于 60％时，21.75％的用户被区分开。而当相似度低于 30％时，73.48％的用户被区分开，也即与之前的区分度相比提高了 3 倍。这表明，通过增加用户 PNL 的相似度来降低用户被区分开的概率是可行的。需要注意的是，PNL 的提升并不能决定 Wi-Fi 指纹是否被区分开，因为攻击者每次获得的该用户的 Wi-Fi 指纹不一定是相同的，即使两个 PNL 完全一致，其对应的 Wi-Fi 指纹也有可能是不一样的。但是，如果两个 PNL 越相似，那么 Wi-Fi 指纹被用来区分用户或者刻画用户的可能性就越低。下面将使用用户 PNL 的相似度来描述用户隐私泄露的程度。

此外，用户被区分的概率随着地点的变化也会发生变化。比如，医院用户的区分度可达 57.86％，而在大学图书馆中用户的区分度则只有 22.35％。这表明，位置类型对于用户关系的差异性也有影响，因此在实验部分，算法的评估将从几个典型的不同地点进行。

用户的活动范围跨度很大，从数十公里到上百公里，因此利用用户活动范围内的全部数据（全局信息）来分析并修改用户的 PNL 是不可行的，因为全局范围内的计算量过大，且用户所处环境变化也很大。鉴于此，本文采用一个用户处于某个地点时周围用户的信息来进行分析判断（局部信息），这也与现实中攻击者只能在某个地点的有限时间内去获取用户 Wi-Fi 指纹的场景一致。图 2 分析了用户与其兄弟用户之间距离与相似度的关系。这里定义一个用户最频繁访问的 SSID 的位置作为他最频繁访问的地点。可以看出，当相似度大于 0.6 时，有 75％以上的用户彼此间的距离小于 2 km。空间上距离接近的用户其 PNL 往往也非常相似，这也证明了利用周围相似用户的 PNL 来修改目标用户的 PNL 的可行性。

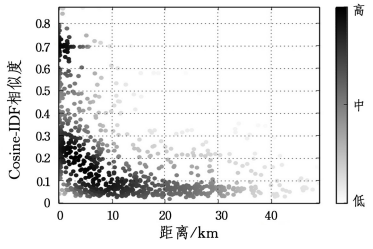


图 2　距离与相似度的关系

Fig. 2　Relation between distance and similarity

已有研究表明，主动扫描的延迟主要是由于密集的 AP 覆盖与隐藏 AP 造成的[23]。由于平时用户通过连接新的 AP 来向 PNL 中添加 SSID，并未影响用户的 Wi-Fi 连接体验，因此向用户的 PNL 中添加伪 SSID 并不会对用户的连网体验造成影响。

总而言之，高达 90.90%的用户都可以通过 Wi-Fi 指纹被区分开，而与邻居用户具有高相似度的 PNL 的用户更难被区分，同时也与邻居用户在空间上更加接近。因此，利用周围人的 PNL 来模糊目标用户的 PNL 是可行的，同时也会使得用户间的 PNL 相似度增加，从而降低隐私泄露的风险。

4 无线隐私保护策略

本节首先描述了本文需要解决的问题，即伪 SSID 的选择问题，然后针对该问题提出了一种基于协同过滤的保护算法，并利用该算法给用户的 PNL 中添加其未曾连接过的 SSID。该保护策略被设定为在上述 Wi-Fi 连接软件中运行。

4.1 伪 SSID 的选择问题

本文旨在找到若干 SSID 添加到目标用户的 PNL 中，并最大化其与周围用户的 PNL 相似度：

$$\underset{\mathscr{A}\in S}{\text{maximize}} \sum_{u\in U}\sum_{v\in U\setminus u} Sim(S_u+A_u, S_v+A_v) \tag{3}$$

$$\text{subject to } \mathscr{A}=\bigcup_{u\in U} A_u \tag{4}$$

$$|A_u|\leqslant k, \forall u\in U \tag{5}$$

其中，A_u 是添加到用户 u 的 PNL 中的伪 SSID 集合，v 是当前区域内的其他用户，$\mathscr{A}$ 是所有伪 SSID 集合的集合，而 S 是当前所有用户 U 的所有 SSID 的集合。添加到每个用户 PNL 中的 SSID 个数不能多于 k 个，以避免造成过重的计算负担。

为了描述用户与 SSID 之间的关系，图 3 给出了一种双层网络模型 $G(V_u, V_s, E)$，它包括两种节点，其中椭圆节点表示用户(V_u)，方形节点表示用户曾连接过的 $SSID(V_s)$，用户与 SSID 之间的连线表示这个 SSID 出现在了该用户的 PNL 中。这里，用 S_u 表示用户 u 曾经连接过的 SSID 集合(即 PNL)，而用 U_s 表示曾经连接过 SSID s 的用户集合。在该关系网络中，如果两个用户至少拥有一个共同的 SSID，说明他们的 PNL 相似度大于 0，即他们是相连的邻居用户。将一个用户的邻居统称为 N_u^h，其表示经过 h 跳步连接起来的邻居用户。比如，N_u^1 表示那些与用户 u 至少共享一个 SSID 的邻居用户。

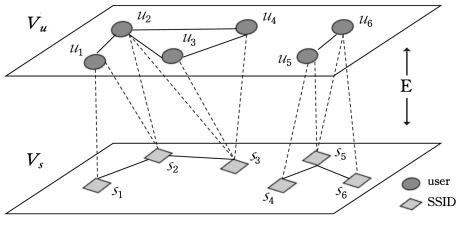


图 3 基于 Wi-Fi 连接的用户关系图

Fig. 3 Diagram of user relationships based on Wi-Fi connections

为了分析用户与 SSID 之间的相关性，本文利用 tf-idf 机制来衡量 SSID s 对用户 u 的重要性：

$$\omega_{u,s}=f_{u,s}\times\log\left(\frac{|U|}{|U_s|}\right) \tag{6}$$

其中，$f_{u,s}$ 是用户 u 连接 SSID s 的频率。根据常识，被频繁连接的 SSID 通常对用户很重要(比如工作场所的 Wi-Fi)。因此，这里假设如果一个用户频繁连接 SSID，且连接过该 SSID 的用户人数很少(比如家里的 Wi-Fi)，即认为该 SSID 对该用户很重要。为了衡量用户 u 与其邻居用户(包括直接相连的邻居与间接相连的邻居)的相似度，本文给出如下定义：

$$Sim(S_u, S_v)=\begin{cases} C(S_u, S_v), & \text{if } v\in N_u^1 \\ Ave(Sim(S_u, S_t)\times C(S_t, S_v)), & \text{otherwise} \end{cases} \tag{7}$$

其中，t 是连接用户 u 与 v 的中间用户，属于 N_u^1。如果两个用户直接相连，则他们的 PNL 的相似度即为 Cosine-IDF 相似度；否则，其 PNL 相似度即为该用户与中间用户 t 的相似度乘积的平均值。

4.2 基于协同过滤的保护算法

基于以上概念，本文提出了一种基于协同过滤的算法来从周围用户的 PNL 中选择伪 SSID 并添加到目标用户的 PNL 中。协同过滤[24]是一种基于他人的兴趣偏好来为用户进行推荐的技术，即通常会为用户推荐与他具有相似偏好的其他用户的商品，如果他接受了该推荐，则二者偏好会更加相似。因此，本文通过给用户推荐相似用户的 PNL 中的 SSID，来增加用户 PNL 之间的相似度。将用户与 SSID 的相关性作为用户对 SSID 的评分，通过如下公式来最终计算一个用户对他的未连接过的 SSID 的“评分”。

$$R_{u,s}=\overline{\omega_u}+\frac{\sum_{v\in U}(\omega_{v,s}-\overline{\omega_v})*Sim(S_u, S_v)}{\sum_{v\in U} Sim(S_u, S_v)} \tag{8}$$

其中，$R_{u,s}$ 是用户 u 对其未曾连接过的 SSID s 的评分，$\omega_{v,s}$ 是用户 v 对其连接过的 SSID 的评分，$\overline{\omega_u}$ 是用户 u 对其原有 SSID 的评分均值。该公式主要考虑了用户 PNL 的相似性与 SSID 对用户的重要程度。所谓重要的 SSID，即用户与 SSID 的相关性较高，这些 SSID 通常会被用户连接得更为频繁且与用户有更紧密的关系，也更容易被移动设备发出，因此也更容易被添至用户的 PNL 中。对于不重要的 SSID，则会被最近连接过的 SSID 替换[14]。

下面给出了算法的具体执行过程。首先，计算用户与其原有的 SSID 的相关性及用户与其他用户的相似度。然后，用户对原来不曾连接过的 SSID 进行评分，将评分进行递减排序后选出评分最高的前 k 个 SSID 添加到目标用户的 PNL 中。这里，使用宽度优先搜索来得到这 k 个伪 SSID。如果用户 u 的直接相连邻居用户 N_u^1 所提供的新的 SSID 的个数不少于 k，则通过遍历直接相连的用户的 PNL 可以得到 A_u，否则，遍历 $N_u^h(h>1)$中的用户，直到得到前 k 个 SSID。本次实验中，由于用户的 PNL 的长度远大于 k 值，因此绝大多数情况下遍历的用户都为 N_u^1。

此外，如果一个用户所有的 SSID 都没有被其他人连接过，则周围没有相似的 PNL，他的 PNL 就不会被上述过程修改。针对这种情况，算法选择添加当前区域最流行的 SSID 到用户的 PNL 中。用户 u 被修改过的 PNL 会由用户原来的 PNL S_u 与新添加的伪 SSID 集合 A_u 所更新。

以上为一个单步更新过程，本文将用户的连接过程分为两类。一类是用户从一个地方移动到另一个地方(如从家里到医院)；另一类是用户待在某处不动，而其周围环境不断变化。这里假设当用户到达一个新地方时，他会主动连接 Wi-Fi，在连接过程中，将会执行上述算法来修改用户的 PNL。

当用户待在一处而环境不断变化时，算法将会周期性地更新以修改 PNL，它不会移除用户原有的 SSID，只是用新的伪 SSID 替代旧的伪 SSID。当设备已经连接上 Wi-Fi 时，上述过程不会影响其连接状态。

5　实验与评估

5.1　实验设置

本文在不同的测试场景下验证算法的有效性。所选择的 3 个不同测试场景为火车站、大学、商场，这些地点是根据上述数据集中的位置类型、街道地址与经纬度确定的。这里通过详细的街道地址找到相应的地点，并通过位置类型与经纬度进行验证。因为数据具有稀疏性，把在上述地点中出现了一天的用户作为测试人群，并通过一个月的 Wi-Fi 连接记录产生他们的 PNL，即使用一个月的 Wi-Fi 连接记录模拟 PNL。在这 3 个场景中，测试人群总人数分别为有 76，116 和 119。实验通过对比本文算法与随机选择算法的结果，验证了算法的有效性；同时还分析了添加的 SSID 个数 k 与原有 PNL 的长度 n 对算法结果的影响；最后分析了算法对于模糊 PNL 的效果。

5.2　实验结果

图 4 与图 5 对比了所提算法与随机选择算法的结果，横纵坐标分别表示当 $k=0$ 与 $k=5$ 时用户之间的相似度，横坐标一定时，纵坐标越大，算法的结果就越好。

在随机选择算法中，从总 SSID 的集合 V_s 中随机选择了 k 个 SSID 添加到用户的 PNL 中。当被添加个数 $k=5$ 时，在上述 3 个地点，用户与他们直接相邻的邻居 N_u^1 的平均 Cosine-IDF 相似度从 0.11，0.07，0.16 提升到 0.16，0.11，0.20，说明该随机方法对于用户相似度的提升几乎没有影响。而本文算法将该相似度提升到了 0.23，0.15，0.38。通过分析发现，每个用户的邻居都非常多，而且大多是相似度很低的弱关系，因此导致了平均相似度的提升不太显著。

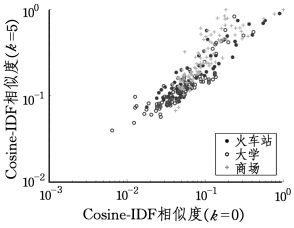


图 4　本文算法的相似度提升

Fig. 4　Similarity increasement of our algorithm

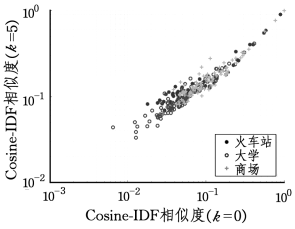


图 5　随机选择算法的相似度提升

Fig. 5　Similarity increasement of random-selection algorithm

算法中，添加到用户 PNL 中的 SSID 个数的上限为 k。图 6 分析了 k 对用户与其邻居用户的相似度的影响。当 k 增加时，用户的相似度也增加，而当 k 从 0 变为 1 时，相似度的增加最为显著，3 个地点的平均相似度分别提高了 53.86%，69.39%，61.06%；当 $k=7$ 时，平均相似度分别提高了 119.21%，140.52%，147.35%。从图 6 中可以看出，大学的相似度低于其他两处，这与直觉不符，经过研究发现，这是因为大学里的用户有远超其他各处用户的邻居数目。大学里的用户邻居数平均为 59.81，而火车站与商场的用户邻居数分别为 16.32，9.13，从而表明大学用户与周围用户的关系大都是弱关系，会导致平均相似度较低，这也是本文算法与随机选择方法相比提升效果不明显的原因。为此，图 7 给出了 k 对用户及其兄弟用户相似度的影响，结果表明，大学里的用户最初的相似度明显高于其他二者，从而导致了增长幅度较小。

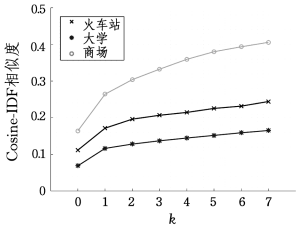


图 6　k 对用户及其邻居用户的相似度的影响

Fig. 6　Impact of k on similarity between users and their neighbors

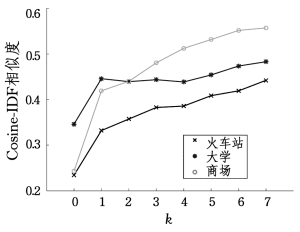
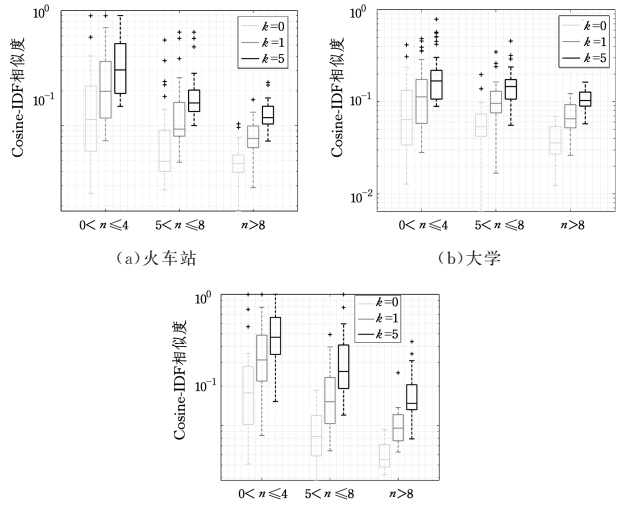


图 7　k 对用户及兄弟用户的相似度的影响

Fig. 7　Impact of k on similarity between users and their siblings

图 8 给出了 PNL 相似度与用户最初的 PNL 长度 n 之间的关系。根据 n 的大小，用户被分为 3 组：$0<n\leqslant 4$，$4<n\leqslant 8$，$n>8$。第一组中，3 个地点的用户之间的平均相似度分别为 0.19，0.09，0.27，当 $k=5$ 时，分别提高了 95.57%，139.17%，103.41%。在最后一组（$n>8$）中，相似度提高了 154.86%，166.75%，200.97%。由此可知，PNL 较长的用户由于其最初的相似度较低，通常会随 k 有明显的提升。



(a)火车站　(b)大学　(c)商场

图 8　n 对用户及其邻居用户的相似度的影响

Fig. 8　Impact of n on similaritiy between user and their neighbors

图 9 阐述了用户 PNL 中 SSID 的位置类型变化，以此来说明算法对用户画像丰富度的影响。从图 9 中可以看出，当 $k=5$ 时，用户 PNL 中位置类型的平均个数从 3.03，2.70，2.39 增加到 4.28，4.02，3.13，从而说明了用户画像的多样化。

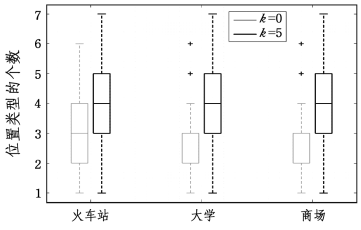


图 9　用户 PNL 中位置类型的变化

Fig. 9　Changes on location types in users' PNLs

总而言之，PNL 修改算法有效地提升了用户与周围用户的 PNL 的相似度，这表明在不同场景下用户被刻画或者被区分开的概率都将减小，它将有效阻止攻击者窃听用户的隐私。

结束语 在 Wi-Fi 连接过程中，主动扫描经常会泄露用户的隐私。为了理解这一问题可能带来的安全隐患，文中在两个包含数百万 Wi-Fi 连接记录的数据集上分析了主动扫描可能造成的 SSID 泄露风险。本文发现，仅通过部分 PNL 就可以辨别超过 90%的用户，利用 SSID 包含的语义信息，攻击者甚至可以得到用户在社会学意义上的隐私信息。为了保护用户隐私，阻止攻击者对用户进行识别和刻画，文中提出了一种基于用户间相似度和协同过滤的 SSID 保护策略。该策略通过添加伪 SSID 来模糊用户的真实 PNL，实现对用户 Wi-Fi 指纹的匿名化。相对于传统方法，它的成本非常低，而且容易实施。本文方法降低了用户被攻击者识别的可能性，减少了用户的信息泄露程度。为验证算法的有效性，文中在多个不同场景下进行了模拟实验。在未来，我们将针对用户和地点的独特性来优化算法的性能。

参 考 文 献

[1] DAI Z，DINO A，MACAULAY S A. Attacking Automatic Wireless Network Selection[C]// Proceedings of the Sixth Annual IEEE SMC Information Assurance Workshop. IEEE，2005：365-372.

[2] FREUDIGER J. How Talkative is Your Mobile Device?：An Experimental Study of Wi-Fi Probe Requests[C]// Proceedings of the 8th ACM Conference on Security and Privacy in Wireless and Mobile Networks. ACM，2015.

[3] CHERNYSHEV M，VALLI C，HANNAY P. On 802.11 Access Point Locatability and Named Entity Recognition in Service Set Identifiers[J]. IEEE Transactions on Information Forensics and Security，2016，11(3)：584-593.

[4] FAN Y C，CHEN Y C，TUNG K C，et al. A Framework for Enabling User Preference Profiling Through Wi-Fi Logs[J]. IEEE Transactions on Knowledge and Data Engineering，2016，28(3)：592-603.

[5] XU Q，ZHENG R，SAAD W，et al. Device Fingerprinting in Wireless Networks：Challenges and Opportunities[J]. IEEE Communications Surveys and Tutorials，2016，18(1)：94-104.

[6] CUNCHE M，KAAFAR M A，BORELI R. Linking Wireless Devices Using Information Contained in Wi-Fi Probe Requests[J]. Pervasive and Mobile Computing，2014，11(4)：56-69.

[7] BONNE B，QUAX P，LAMOTTE W. Raising Awareness on Smartphone Privacy Issues with SASQUATCH，and Solving Them with PrivacyPolice[C]// Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems：Computing，Networking and Services. 2014：379-381.

[8] LINDQVIST J，AURA T，DANEZIS G，et al. Privacy-preserving 802.11 Access-point Discovery[C]// Proceedings of the Second ACM Conference on Wireless Network Security. ACM，2009：123-130.

[9] KIM Y S，TIAN T，NGUYEN L T，et al. Lapwin：Location-aided Probing for Protecting User Privacy in Wi-Fi Networks[C]// Proceedings of IEEE Conference on Communications and Network Security. 2014：427-435.

[10] PANG J，GREENSTEIN B，GUMMADI R，et al. 802.11 User Fingerprinting[C]// Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking. ACM，2007：99-110.

[11] DESMOND L C C，YUAN C C，PHENG T C，et al. Identifying Unique Devices Through Wireless Fingerprinting[C]// Proceedings of the First ACM Conference on Wireless Network Security. ACM，2008：46-55.

[12] CHENG N，MOHAPATRA P，CUNCHE M，et al. Inferring User Relationship from Hidden Information in Wlans[C]// Military Communications Conference. IEEE，2012：1-6.

[13] BARBERA M V，EPASTO A，MEI A，et al. Signals from The Crowd：Uncovering Social Relationships through Smartphone Probes[C]// Proceedings of the 2013 Conference on Internet Measurement Conference. ACM，2013：265-276.

[14] LUZIO A D，MEI A，STEFA J. Mind Your Probes：De-anonymization of Large Crowds through Smartphone Wi-Fi Probe Requests[C]// Proceedings of the 35th Annual IEEE International Conference on Computer Communications. IEEE，2016：1-9.

[15] SONG Y，YANG C，GU G. Who is Peeping at Your Passwords at Starbucks? -To Catch An Evil Twin Access Point[C]// Proceedings of IEEE/IFIP International Conference on Dependable Systems and Networks. IEEE，2010：323-332.

[16] CALLEGATI F，CERRONI W，RAMILLI M. Man-in-the-Middle Attack to the Https Protocol[J]. IEEE Security and Privacy，2009，7(1)：78-81.

[17] SKINNER K，NOVAK J. Privacy and Your App[C]// Apple Worldwide Dev. Conf.(WWDC). America，2015.

[18] Android 6.0 Changes [EB/OL]. https://developer.android.com/about/versions/ marshmallow/android-6.0-changes.html.

[19] WANG W. Wireless Networking in Windows 10[C]// Windows Hardware Engineering Community Conference (WinHEC). 2015.

[20] VANHOEF M，MATTE C，CUNCHE M，et al. Why Mac Address Randomization is not Enough：An Analysis of Wi-Fi Network Discovery Mechanisms[C]// Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. ACM，2016：413-424.

[21] VANHOEF M，MATTE C，CUNCHE M，et al. Why Mac Address Randomization is not Enough：An Analysis of Wi-Fi Network Discovery Mechanisms[C]// Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. ACM，2016：413-424.

[22] ZANG H，BOLOT J. Anonymization of Location Data does not Work：A Large-scale Measurement Study[C]// Proceedings of the 17th Annual International Conference on Mobile Computing and Networking. ACM，2011：145-156.

[23] XU C，TENG J，JIA W. Enabling Faster and Smoother Handoffs in Ap-dense 802.11 Wireless Networks[J]. Computer Communications，2010，33(15)：1795-1803.

[24] TERVEEN L，HILL W. Beyond Recommender Systems：Helping People Help Each Other[C]// HCI in the New Millennium. 2001：487-509.