文章编号:1673-9590(2026)03-0036-09

随机 MAC 机制下基于 Wi-Fi 探测请求的人流 OD 识别方法

朱成娟,赵星吉, 陈爱同
(大连交通大学 交通工程学院,辽宁 大连 116028)

摘 要:针对传统基于真实 MAC 地址的人流 OD 监测方法面临的隐私保护与数据失效问题,提出一种基于 Wi-Fi 探测请求信息元素的匿名人流 OD 识别方法,通过分析探测请求中 HT Capabilities 和 VHT Capabilities 等字段的稳定特征,设计改进的 DBSCAN 聚类算法,实现随机 MAC 设备的精准区分。针对跨区域轨迹匹配问题,提出基于信息元素分组筛选与距离矩阵计算的 OD 关联策略。结果表明,在开放设施中,设备聚类精度达到 97.87%,OD 匹配精度达到 81.8%,验证了方法的有效性。与现有技术相比,该方法在保护用户隐私的同时,兼顾了高精度与低计算复杂度,为公共安全、交通规划及应急管理提供了新思路。

关键词:人流 OD 识别;Wi-Fi 探测请求;随机 MAC 地址;聚类算法;隐私保护
文献标识码:A **DOI**:10.13291/j.cnki.djdxac.2026.03.004

随着城市化进程的加速,城市公共活动的频次和公共场所人流密度逐年增大,催生了对人流精准 OD(Origin-Destination)数据的需求。人流 OD 记录人流移动的起点和终点,是表征人群移动规律的核心数据,在公共安全、交通管理、城市规划、灾难和危机管理以及公共活动组织等领域有重要应用价值。传统的人流 OD 监测方法主要包括问卷调查和视频监控,但这些方式面临数据质量、隐私保护、实施成本三重困境。问卷调查可以直接获取用户的出行信息,但样本量有限、数据更新慢。视频监控能够实时获取大量数据,但涉及用户隐私,存在法律风险。特别是随着欧盟“General Data Protection Regulation Laws”[1]的出台,视频监控等数据的使用受到严格限制。因此,如何在保护个人隐私的前提下,设计监测人流的方法,获取人流量和移动模式的匿名统计信息,成为当前研究的重点。

随着智能手机等具有 Wi-Fi 功能的移动设备的普及,一些学者基于手机等移动设备的 Wi-Fi 探测请求(Probe Requests, PR)中包含的设备 MAC 地址设计了人流 OD 监测系统,用于实时监测人流量和流动模式,以支持公共安全、应急管理和交通管理[2-3]。早期手机等移动设备的 Wi-Fi 探测请求包含设备真实 MAC 地址,具有唯一性,许多研究基于这一特性设计人流监测方案,如实时计算大面积局部客流[4]、公交客流 OD 矩阵推算[5]、大型场馆内人群定位和流量分析[6]、城市出行轨迹提取[7]、估计乘客 OD 行程和站台等待时间等[8]。

由于 Wi-Fi 设备真实 MAC 的唯一性,通过长期监测可以对设备使用者进行定位和跟踪,存在泄露个人隐私的风险。为了保护设备使用者的隐私安全,现代移动设备普遍采用了 MAC 地址随机化技术,在预关联状态(即设备扫描可用网络但尚未连接时)使用随机生成的 MAC 地址,可以实现将多个探测请求与同一设备关联,从而保护用户隐私,为基于设备 Wi-Fi 的探测请求监测公共场所人流量及移动模式提供保障。

在移动设备采用随机 MAC 地址机制的场景下,基于 Wi-Fi 探测请求的人流监测面临两个核

收稿日期:2025-03-12
基金项目:国家自然科学基金面上项目(72171033);辽宁省教育厅科学研究计划项目(JDL2019037)
第一作者:朱成娟(1981—),女,讲师。E-mail:cjzhu@djtu.edu.cn

心挑战:①设备区分——通过聚类分析将同一区域内具有相同行为特征的随机 MAC 地址归类,以准确统计该区域人数;②轨迹关联——建立跨区域随机 MAC 地址的逻辑映射,从而识别完整的人员移动路径。针对设备区分挑战,现有研究主要有两类方法论。第一类基于时序特征分析,通过挖掘设备探测请求的时间模式差异实现区分。该方法由 Franklin 等[9]提出,Matte 等[10]进一步通过帧间到达时间分析提升了识别精度。第二类方法聚焦探测请求中的信息元素(Information Elements, IEs)解析,Vanhoef 等[11]验证了 IEs 特征对设备识别的有效性。近年来,研究趋势呈现多特征融合的特点,Uras 等[12]采用 OPTICS 聚类算法,联合分析 IEs 和时序特征,显著提升了 MAC 地址去随机化效果。Pintor 等[13]引入机器学习模型,通过 IEs、时序特征及信号强度等多维数据协同分析,将设备识别准确率提升至新高度。

当前研究存在两个显著局限,首先,相较于单区域设备区分研究,随机 MAC 机制下的跨区域人流 OD 识别缺乏系统性探索;其次,现有方法多针对无外干扰的封闭环境设计[14],对开放区域信号干扰等实际约束的适应性仍未解决。为此,本研究构建分层关联框架,基于设备区分领域的最新成果(包括时序特征分析与信息元素解析),设计跨区域随机 MAC 映射算法,重点解决开放环境中时空维度的设备关联难题,最终实现隐私保护前提下的人流轨迹识别。

1　问题描述

1.1　Wi-Fi 探测请求与随机 MAC 机制

基于移动设备 Wi-Fi 探测请求实现人流监测的技术基础是 IEEE 802.11 标准协议。根据该标准规范,当启用 Wi-Fi 无线通信接口的智能设备处于未关联状态(未连接到任何无线接入点)时,设备将周期性地执行主动扫描以搜索可用的 Wi-Fi 网络。在此过程中,设备按照预定义的信道切换顺序(如 2.4 GHz 频段的 Ch1—Ch14 信道),在每个信道上广播发送探测请求(Probe Request, PR)数据帧。若在设定的等待时间内(通常为 100~500 ms)未接收到来自接入点的探测响应(Probe Response),则自动切换至下一个信道继续扫描流程。通过此机制,具备 Wi-Fi 功能的移动设备可被部署在公共场所的监测探针发现,并记录其发送的探测请求信息。依据 IEEE 802.11 标准定义帧格式规范(详细字段见表 1),探测请求帧的 MAC 头部包含源地址(Source Address, SA)、目的地址(Destination Address, DA)等关键字段信息。

表 1　Wi-Fi 探测请求帧包含的信息

Tab. 1　Information in Wi-Fi probe request frames

字段名称	内容
帧控制字段(Frame Control)	协议版本、帧类型(管理帧)、子类型(探测请求)、标志位(To DS、From DS 等)
持续时间字段(Duration)	以微秒为单位的时间值
目的地址(Destination Address)	通常为广播地址(FF:FF:FF:FF:FF:FF)
源地址(Source Address)	设备的真实 MAC 地址或随机化 MAC 地址(如 iOS 设备的私有地址)
BSSID 字段(Basic Service Set Identifier)	通常为广播地址(FF:FF:FF:FF:FF:FF)
SSID 字段(Service Set Identifier)	特定网络的名称或空字符串(搜索所有网络)
支持速率字段(Supported Rates)	一组速率值,单位为 0.5 Mbps
扩展支持速率字段(Extended Supported Rates)	额外的速率值
能力信息字段(Capability Information)	支持 ESS/IBSS、加密(如 WEP)、短前导码、PBCC 等
信息元素(Information Elements, IEs)	DS Parameter Set、HT Capabilities、VHT Capabilities、Vendor Specific 等
信号强度(RSSI)	信号强度值,用于定位和聚类分析

当前实现高精度随机 MAC 设备区分的主流算法主要依赖探测请求帧内嵌信息元素(Information Elements, IEs)的聚类分析。表 2 列出了 Wi-Fi 设备探测请求中包含的典型信息元素参数,包括元素标识符(Element ID)、数据长度(Length)以及数据内容(Data)等字段的十六进制编码示例。分析表中所示的探测请求帧结构可以发现,在源地址(SA)字段中,当设备采用传统固定 MAC 地址时,其地址结构遵循 IEEE EUI-48 标准规范,前 3 个字节为组织唯一标识符(Organizationally Unique Identifier, OUI),由国际电气电子

工程师学会(IEEE)统一分配给各设备制造商;后3个字节为网络接口控制器(Network Interface Controller, NIC)编号,由制造商自行分配以确保全球唯一性。

表 2 Wi-Fi 设备探测请求中的信息元素

Tab. 2 Information elements in Wi-Fi device probe requests

ID	信息元素名称	数据长度	内容
0	SSID(Service Set Identifier)	可变长度(0~32 字节)	特定网络的名称或空字符串(搜索所有网络)
1	Supported Rates(支持速率)	1~8 字节	一组速率值,单位为 0.5 bit/s
50	Extended Supported Rates(扩展支持速率)	1~8 字节	额外的速率值
3	DS Parameter Set(DS 参数集)	1 字节	信道编号(如 1~14 表示 2.4 GHz 频段,36~165 表示 5 GHz 频段)
45	HT Capabilities(高吞吐量能力)	26 字节	包含 MIMO、信道绑定、短保护间隔等能力信息
191	VHT Capabilities(极高吞吐量能力)	12 字节	包含 MU-MIMO、信道宽度(80 MHz、160 MHz)等能力信息
221	Vendor Specific(厂商特定信息)	可变长度(至少 4 字节:3 字节 OUI + 1 字节数据)	厂商 ID(OUI)和自定义数据(如设备型号、操作系统版本等)
48	RSN(Robust Security Network)	可变长度(至少 2 字节:版本 + 组密码套件)	加密类型(如 AES)、认证类型(如 PSK、802.1X)等信息
42	ERP(Extended Rate PHY)	1 字节	包含是否支持 ERP-OFDM、是否使用保护机制等信息
33	Power Capability(功率能力)	2 字节	最小和最大传输功率值(单位为 dBm)
36	Supported Channels(支持信道)	可变长度(至少 2 字节:第一个信道 + 信道数量)	一组信道编号(如 1~14 表示 2.4 GHz 频段,36~165 表示 5 GHz 频段)
46	QoS(Quality of Service)	22 字节	包含 EDCA 参数、流量类别(TID)等信息
107	Interworking(互联互通)	可变长度(至少 1 字节:访问网络选项)	包含网络类型(如家庭、企业、公共热点)等信息
111	Roaming Consortium(漫游联盟)	可变长度(至少 1 字节:漫游联盟数量)	包含漫游联盟标识符(OI)列表
131	Time Advertisement(时间广告)	可变长度(至少 10 字节:时间戳 + 时区)	包含时间戳、时区等信息
221	WMM(Wi-Fi Multimedia)	可变长度(至少 7 字节:WMM 参数集)	包含流量优先级、EDCA 参数等信息

为应对日益严格的隐私保护法规要求(如欧盟《通用数据保护条例》GDPR),现代智能设备操作系统(如 iOS 14+和 Android 10+)普遍部署了 MAC 地址随机化技术方案:当 MAC 地址首字节的第 7 位(即 U/L 位,Universally/Locally administered bit)被设置为二进制“1”时(随机 MAC 地址形式见图 1),表明该地址为设备在预关联阶段动态生成的临时随机值。这一技术变革导致传统基于固定 MAC 地址的设备追踪与行为分析模型完全失效。

图 1 随机 MAC 地址形式

Fig. 1 Random MAC address format

1.2　人流 OD 识别的关键问题

在基于随机 MAC 地址的 Wi-Fi 探测请求数据进行人流 OD 识别的过程中,需要系统性解决以下三个层次的技术难题:

(1)设备定位准确性问题

定位问题的核心在于根据探测请求帧中携带的接收信号强度指示(Received Signal Strength Indicator, RSSI)参数,准确判断目标设备是否位于预设的监测区域内。然而,由于无线信号在传播过程中普遍存在多径效应(Multipath Effect)、障碍物遮挡衰减(如墙体对 2.4 GHz 信号的穿透损耗可达 10~20 dB)以及同频段设备干扰(如蓝牙设备与 Wi-Fi 信道重叠)等因素的综合影响,实测 RSSI 值往往呈现显著波动特性。

(2)设备聚类精确性问题

设备聚类问题的核心挑战在于如何对同一监测区域内捕获的具有时变 MAC 地址的移动设备进行准确分组。具体难点是:①同一品牌同型号设备由于硬件设计与驱动软件的一致性,其探测请求中的信息元素参数(如支持的频段、MIMO 配置等)相似度高,会显著增加聚类算法的误判风险。②在低人流量时段或设备处于休眠状态(如屏幕关闭)时,单个设备可能仅产生 1~2 条探测请求,导致特征提取不充分。

(3)跨区域 OD 匹配有效性问题

OD 匹配问题的核心目标是通过分析不同监测区域捕获的设备聚类结果,建立跨空间区域的设备移动轨迹关联。该问题的主要技术障碍表现为:①路径复杂性。实际场景中人员移动路径具有高度多样性,设备可能经过多个非连续监测区域(如从教室 A 移动到走廊再进入教室 B),导致简单的起点-终点(O-D)匹配模型无法准确还原真实轨迹。②时间异步性。由于各区域监测探针的时钟可能存在毫秒级偏差,且设备在不同区域的停留时间存在差异,难以直接通过时间戳对其实现精确匹配。

2　人流 OD 识别的算法框架

在实际环境中,Wi-Fi 探针除捕获手机、平板等移动终端的无线信号外,亦会同步接收智能家电(如摄像头、空调等)及搭载 Wi-Fi 模块的工业设备产生的探测请求。由于此类非移动设备与信号残缺数据会干扰人流分析,需建立多级数据清洗机制。基于随机 MAC 地址特性过滤非移动终端,通过信号质量评估剔除噪声数据,形成可用于定位、设备识别及 OD 匹配的净化数据集。研究发现,智能设备与移动终端的 MAC 地址生成机制存在显著差异,前者普遍采用固定全局 MAC 地址,而后者为保护用户隐私启用随机 MAC 地址动态生成策略。因此,可以将全局 MAC 地址作为非移动设备的判别特征,直接排除在人流分析流程之外。在数据清洗的基础上,人流 OD 识别的算法框架见图 2。

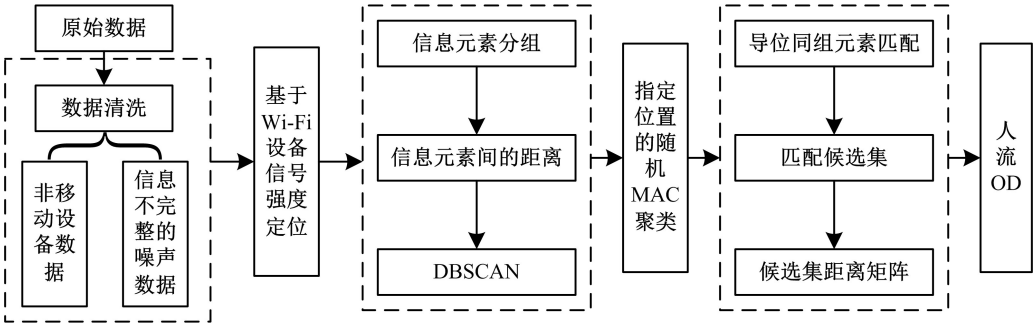

图 2　算法框架

Fig. 2　Algorithm framework

在解决 Wi-Fi 设备的定位问题时,Wi-Fi 信号强度(RSSI)易受多径效应、障碍物衰减及环境干扰影响,导致单一探针接收信号存在显著波动(相同位置设备 RSSI 标准差达±8 dBm)。现有研究主要通过 3 类方法应对定位误差问题:①多探针空间协同。部署冗余探针节点,基于信号到达时间差(TDoA)或三角定位实现较高精度的定位,但硬件成本较高。②信号强度-距离映射。通过场景实测建立 RSSI 与物理距离的经验模型,适用于结构规则的小型封闭空间。③机器学习优化。

利用 KNN、随机森林等算法学习环境特征与信号衰减的非线性关系,可动态适应复杂环境,但依赖大规模标注数据。本研究受限于试验场景与设备部署条件,选择构建信号强度-距离映射模型:通过校准测试确定 RSSI 有效区间为[-90 dBm, 0 dBm],并依据路径损耗公式推算设备相对位置,实现低成本单探针定位。

$$L(d) = L_0 + 10\gamma\log_{10}(d/d_0) \tag{1}$$

式中:γ 为环境衰减因子;d_0 为参考距离。

针对指定区域内 Wi-Fi 探测请求数据的设备区分问题,现有方法普遍采用对信息元素进行处理后,再调用聚类算法(如 DBSCAN、OPTICS 等)进行聚类[13-14]。此类方法在 22 个设备的公开数据集上聚类后生成 21 个类簇,虽能实现设备数量统计层面的较高准确性(误差为±4.5%),适用于区域人流量估算,但其类簇纯度存在显著缺陷:约 45.5%的类簇(10/21)混杂了多个设备的探测请求,导致跨区域类簇相似度比对时 OD 识别误差率提升至 31.2%。为解决上述问题,本研究对公开数据中各设备的探测请求数据进行分析,发现 HT/VHT 信息元素具有较好的稳定性。具体表现在:(1)HT Capabilities 字段稳定性(所有移动 Wi-Fi 设备都具有该信息元素)。对于同一个 Wi-Fi 移动设备,其 HT Capabilities(HT)字段前 16 位(HT_CAP_INFO)在 MAC 随机化过程中保持恒定。(2)VHT Capabilities 字段一致性(部分设备具有信息元素)。若存在,则同一 Wi-Fi 移动设备的 VHT Capabilities(VHT)字段前 24 位(VHT_CAP_INFO)同样具有跨请求持久性。基于此,提出改进的分组聚类流程:

Step 1 特征分组。依据 HT/VHT 字段前缀值将探测请求划分为互斥子集,消除跨设备干扰。

Step 2 组内距离计算。在每组内按照本文设计的代码计算探测请求信息元素间的距离。

Step 3 组内聚类。基于组内的探测请求信息元素间的距离,使用 DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 算法得到聚类结果。

计算探测请求信息元素间距离的距离矩阵算法流程如下:

Step 1 预处理。提取所有 MAC 地址构成列表,为每个 MAC 解析其对应的信息元素(IE);初始化各 MAC 间的距离矩阵。

Step 2 计算每对 MAC 地址间的距离。相同地址时,dist = 0;不同地址时,分 4 种情况:①取公共信息元素 IE;②基础距离设为 dist = STARTING_DST;③分五类处理公共 IE;④dist = min{min_DST, dist}。

分五类处理公共 IE 的情况又分为:

(1)If 公共速率信息(SUPP/EXT_KEY)不空,则 dist =dist+计算支持速率的对称差集元素数 × DATA_RATE_SCALE。

(2)If 公共网络能力(HT/VHT_CAP_KEY)不空,则 dist = dist+对比 16 进制值的指定位段差异。

(3)If 公共扩展能力(EXT_CAP/INTERWORKING)不空,则 dist =dist+数据长度不同时加固定惩罚值,否则计算位级差异。

(4)If 公共厂商信息(VENDOR_SPEC_KEY)不空,则又分为 3 种情况:① dist =dist+固定权重;②dist =dist+计算厂商集的对称差元素数 × 惩罚系数;③dist =dist+对共有厂商对比数据类型的位差异。

(5)If 公共扩展标签(EXT_TAG_KEY)不空,则 dist =dist+计算类型位差异。

Step 3 输出 MAC 间的距离矩阵。

针对不同区域的 Wi-Fi 设备探测请求聚类后的再匹配问题,按照如下步骤进行匹配:

Step 1 基于信息元素一致性,选取人流 OD 匹配候选集。基于起点 O 区域和终点 D 区域的 Wi-Fi 设备各自的聚类结果,根据探测请求中信息元素包含的种类进行初步筛选,仅保留信息元素种类一致的数据对,作为匹配的候选数据。

Step 2 利用距离矩阵算法得到两组数据间的距离并进行匹配。对于信息元素一致的数据对,调用距离矩阵算法,计算数据对之间的距离。如果两组数据之间的距离小于设定的阈值(本文设置值为 100),并且 O 区域类簇数据被探测到的最后时间与 D 区域类簇数据被探测到的最早时间之间的差值大于从 O 到 D 的平均时间,则视为匹配成功,将两组数据匹配到同一条记录中。

3 验证试验

3.1 试验与数据说明

试验选取两个相距 30 m 的开放教室作为测试场景:教室 1 中初始人数为 47 人,面积约 100 m^2,在中心点部署单台 Wi-Fi 探针(有效覆盖半径约为 200 m),数据采集持续 30 min;教室 2 中面积与布局同教室 1,初始状态含 2 人,探针部署位置对

称,数据采集时长 5 min。为模拟真实环境下移动设备携带情况,所有参与者未预先告知试验目的,其携带设备类型包括手机、平板及笔记本电脑,对设备的 Wi-Fi 启闭状态不做干预,同时周边教室正在上课。试验过程中,随机选取 11 人从教室 1 移动至教室 2,以构建 OD 轨迹真值标签。对公开数据[13-14]中各探测请求间的距离进行统计。

由统计可知,大多数探测请求之间的距离不小于 100,故聚类参数设置如下:采用 DBSCAN 算法,邻域半径 ε 为 100(定义邻域空间最大距离);MinPts 取 2(核心点判定阈值,即邻域内最少点数)。试验数据通过探针捕获 IEEE 802.11 协议探测请求帧(原始数据为 16 进制编码)。

3.2　数据的聚类分析

在教室 1 场景中,共捕获 30 480 条探测请求(Probe Request, PR)数据,数据清洗流程后形成有效数据集,数据清洗过程如下:

(1)非移动设备过滤。基于 MAC 地址随机性检测,剔除固定 MAC 设备数据(如监控摄像头等)。

(2)噪声数据剔除。移除信号强度异常(RSSI ∉ [-90 dBm, 0 dBm])及字段残缺记录。

(3)跨区域干扰排除。依据路径损耗模型定位设备物理位置,排除教室外设备数据,最终保留 18 472 条有效探测请求。

教室 1 长时间存在的随机 MAC 见表 3。

表 3　教室 1 长时间存在的随机 MAC

Tab. 3　Long-term random MAC addresses in classroom 1

序号	MAC	信号强度均值/dBm	持续时段
1	06:14:4b:6d:3e:8b	-58.2	16:29:42—17:05:20
2	06:14:4b:6d:40:f7	-58.9	16:29:40—17:05:14
3	06:69:6c:bf:3b:6b	-86.7	16:29:42—17:05:20
4	4a:69:6c:89:1d:f5	-55.3	16:29:42—17:04:48
5	ea:69:6c:89:1d:ff	-32.5	16:29:40—17:05:20
6	56:06:8a:0e:e9:e6	-86.3	16:38:44—16:54:01
7	12:87:ec:d9:33:8e	-23.2	16:29:40—17:01:38
8	6a:47:36:55:25:39	-73.5	16:29:40—17:02:14
9	3a:45:45:fd:70:59	-64.2	16:29:40—17:01:40
10	72:47:e9:40:f3:37	-71.4	16:30:12—17:04:48
11	5e:24:69:d5:e2:36	-52.1	16:29:40—17:03:48
12	1e:fb:48:4a:45:30	-81.4	16:29:40—16:57:31
13	a6:14:35:12:76:82	-52.3	16:29:40—17:01:44

表 3　(续表)

Tab. 3　(Cont.)

序号	MAC	信号强度均值/dBm	持续时段
14	e2:b4:64:2f:b2:30	-54.6	16:29:40—17:01:18
15	8a:d6:67:46:87:88	-82.4	16:44:27—17:05:20
16	ee:bb:da:ce:45:42	-84.5	16:29:40—17:05:12
17	d6:26:38:75:11:55	-63.2	16:29:40—17:01:18
18	be:c2:09:8b:86:ff	-64.3	16:29:40—17:03:00
19	2e:23:63:45:b0:0e	-66.7	16:29:48—17:00:56
20	56:86:ad:89:69:d8	-70.1	16:29:54—17:01:26
21	52:7a:99:7b:f3:f9	-52.1	16:31:16—16:52:09
22	aa:94:45:36:81:e8	-67.4	16:36:52—17:01:42
23	aa:d2:b5:14:73:9f	-64.3	16:32:32—17:01:30
24	ae:c8:d5:25:86:00	-62.4	16:39:14—16:57:37
25	b6:3d:0b:49:95:91	-52.1	16:31:00—16:59:37
26	d2:f2:13:f1:bf:a1	-65.2	16:32:04—16:55:45
27	da:a1:19:b3:ad:5f	-71.4	16:33:20—17:05:04

数据分析表明,27 个随机 MAC 地址呈现长时活跃特性(持续时间>25 min),结合其 HT/VHT 特征可判定为独立设备,对剩余短时随机 MAC 数据(18 472 条)进行分层聚类:

(1)特征分组。按 HT Capabilities 前 16 位(HT_CAP_INFO[0:15])与 VHT Capabilities 前 24 位(VHT_CAP_INFO[0:23])划分特征子空间。

(2)组内聚类。在各子空间内计算信息元素间的距离,应用 DBSCAN 算法(ε = 100, MinPts = 2)生成 19 个类簇(表 4)。可知,教室 1 共识别 46 个独立移动设备(27 长时 MAC+19 类簇),相较实际人数 47 人,识别准确率可达 97.87%(误差±1 人),验证了本文算法在持续监测场景下的有效性。

表 4　教室 1 短时随机 MAC 的聚类结果

Tab. 4　Clustering results of short-term random MAC addresses in classroom 1

类簇	属于的信息元素分组	信号强度均值/dBm	持续时段
1	HT_ad 和 VHT_92e19033	-72.00	16:29:56—17:03:24
2	HT_ad 和 VHT_92e19033	-65.09	16:31:32—17:04:34
3	HT_ad 和 VHT_92e19033	-87.00	16:34:46—17:05:14
4	HT_ad 和 VHT_92f19033	-62.60	16:31:18—17:05:14
5	HT_ef 和 VHT_92f19033	-88.96	16:33:24—17:03:54
6	HT_ad 和 VHT_92f19033	-61.21	16:42:19—17:05:16

表 4 （续表）

Tab. 4 （Cont）

类簇	属于的信息元素分组	信号强度均值/dBm	持续时段
7	HT_ef 和 VHT_92f19033	-70.95	16:29:40—17:06:25
8	HT_ef 和 VHT_92f99133	-65.11	16:30:10—17:04:12
9	HT_ef 和 VHT_92f99133	-66.35	16:30:22—17:05:18
10	HT_ef 和 VHT_92f99133	-70.00	16:34:06—17:06:53
11	HT_ef 和 VHT_92f99133	-75.40	16:35:44—17:04:54
12	HT_2d	-67.10	16:29:40—17:08:13
13	HT_2d	-65.34	16:31:14—17:02:39
14	HT_2d	-64.41	16:32:36—17:04:26
15	HT_2d	-64.40	16:33:44—17:08:05
16	HT_ef	-87.09	16:31:08—17:05:20
17	HT_No_HT_CAP	-68.51	16:30:02—17:05:08
18	HT_No_HT_CAP	-67.80	16:31:06—17:05:08
19	HT_ad	-70.48	16:31:32—17:03:06

教室 2 共探测 6 min，收集到 4 358 条探测请求数据，经相同预处理流程（非随机 MAC 过滤、RSSI 范围筛选、跨区域定位排除）后，检测到3个持续时间>4.5 min 的随机 MAC 地址（表5）和 112 条短时随机 MAC 数据。对短时随机 MAC 按 HT_CAP 与 VHT_CAP 分组与计算信息元素距离后，组内应用 DBSCAN（$\varepsilon = 100$，MinPts = 2）生成 8 个类簇（表 6）。最终 共识别 11 个独立设备，相较于实际人数 13 人（教室 2 初始 2 人+11 人迁入），准确率可达 84.6%（FN = 2，FP = 1）。

本文试验在教室 1 与教室 2 采用差异化数据采集时长设计，在教室 1 持续监测 30 min，捕获设备全周期活跃状态；在教室 2 短时监测 5 min，模拟一般场景下的快速人流分析需求。从现有研究可知，手机等移动 Wi-Fi 设备在不同工作模式下发送 Wi-Fi 探测请求频率并不相同。在亮屏模式下，设备发送 Wi-Fi 探测请求的频率高，息屏模式下发送的频率低。因此，较长的探测时间能够全面收集各个设备发送的 Wi-Fi 探测请求，对应聚类分析结果的精度较高，如教室 1 的结果。教室 2 的数据收集只有 5 min，在这个时间段内，有些设备可能未发送 Wi-Fi 探测请求，因此设备区分的数量与实际数量的差距较大。但从教室 1 的区分结果来看，试验结果仍充分验证了本文方法在随机 MAC 地址生成机制下对移动设备的高效区分能力。

表 5 教室 2 长时间存在的随机 MAC

Tab. 5 Long-term MAC addresses in classroom 2

序号	MAC	持续时段	RSSI 值
1	a6:14:35:12:76:82	17:02:23—17:04:37	-88.25
2	52:7a:99:7b:f3:f9	17:01:51—17:04:49	-88.10
3	e2:b4:64:2f:b2:30	17:01:29—17:04:07	-89.08

表 6 教室 2 短时随机 MAC 聚类结果

Tab. 6 Clustering results of short-term random MAC addresses in classroom 2

类簇	属于的信息元素组	信号强度均值/dBm	持续时段
1	HT_ad 和 VHT_92e19033	-71.40	17:01:12—17:04:45
2	HT_ad 和 VHT_92f99133	-80.67	17:00:39—17:04:25
3	HT_ef 和 VHT_92f19033	-86.67	17:00:02—17:04:53
4	HT_ef 和 VHT_92f99133	-89.00	17:01:12—17:04:45
5	HT_2d	-83.50	17:00:24—17:04:47
6	HT_2d	-72.62	17:00:44—17:04:24
7	HT_No_HT_CAP	-89.00	17:00:12—17:04:08
8	HT_ef 和 VHT_92f19033	-82.00	17:00:32—17:04:39

3.3 人流 OD 匹配分析

为了收集从教室 1 到教室 2 的人流 OD 数据，将教室 2 得到的长时间存在的 MAC 与教室 1 的进行对比。对比发现，教室 2 的 3 个长时 MAC 分别对应于教室 1 中的序号为 13、14 和 21 的长时 MAC，即有 3 名试验人员从教室 1 移动到教室 2。对教室 2 中短时随机 MAC 进行聚类分析，得到 8 个类簇，通过匹配算法对其进行匹配，具体过程如下：

（1）将两个教室中属于相同信息元素分组的类作为匹配候选对。如表 6 中类 1 的匹配候选对是表 4 中类 1、类 2 和类 3，即表 6 中类 1 应与表 4 中的类 1、类 2 和类 3 作匹配。

（2）根据第 2 节距离矩阵计算算法得到各匹配对中各探测请求间的距离，若平均距离不大于 100 则属于同类。

两教室人流 OD 匹配结果见表 7。从表中可知，在 5 min 的探测时间段内，本文设计的人流 OD 匹配算法识别出有 9 个移动设备由教室 1 移动到教室 2，此外表 6 中属于教室 2 的两个设备未被匹配。推测表 6 未被匹配的设备是教室 2 原有 2 人的设备，不是由教室 1 移动至教室 2 的设备。从匹配的结果来看，精度为 81.8%。

表 7　两教室人流 OD 匹配结果

Tab. 7　Pedestrian OD matching results between two classrooms

教室 2 类簇	属于的信息元素分组	教室 1 类簇	相对差距
1	HT_ad 和 VHT_92e19033	1	94.3
2	HT_ad 和 VHT_92f99133	—	360.2
3	HT_ef 和 VHT_92f19033	4	88.0
4	HT_ef 和 VHT_92f99133	8	88.0
5	HT_2d	—	380.1
6	HT_2d	14	88.0
7	HT_No_HT_CAP	17	96.2
8	HT_ef 和 VHT_92f19033	7	97.4

注:表中“—”表示未匹配成功。

4　结论

本文针对随机 MAC 机制下的人流 OD 识别问题,提出了一种基于 Wi-Fi 探测请求信息元素(如 HT Capabilities 和 VHT Capabilities)的聚类与匹配算法。通过分析探测请求中信息元素的稳定特征(如 HT 和 VHT 字段的前缀值),结合改进的 DBSCAN 聚类算法,实现了高精度的设备区分。在跨区域 OD 匹配中,通过信息元素分组筛选和距离矩阵计算,解决了随机 MAC 机制下设备轨迹关联的难题。验证试验表明,在开放教室场景中,设备聚类精度可达 97.87%,OD 匹配精度可达 81.8%,显著优于现有方法。本文方法在保护用户隐私的前提下,为公共安全、交通管理和应急响应提供了可靠的人流监测技术支持。未来研究可进一步优化开放环境下的定位精度,并探索动态需求场景的实时处理能力。

参考文献:

[1] COVACI A I. Wi-Fi MAC address randomization vs crowd monitoring[D]. Enschede: University of Twente, 2022.

[2] HONG H , SILVA G D , CHAN M C. Crowd probe: non-invasive crowd monitoring with Wi-Fi probe[J]. Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies, 2018, 2(3):1-23.

[3] GUILLEN-PEREZ A, BANOS M D C. A WiFi-based method to count and locate pedestrians in urban traffic scenarios[C]//2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). Limassol: IEEE, 2018: 123-130.

[4] TONG H J, ZHU L J, XIONG Y, et al. Modeling large passenger flow safety by simulation and testing[C]// 2017 3rd IEEE International Conference on Computer and Communications (ICCC). Chengdu: IEEE, 2017: 235-239.

[5] 徐艺梅, 胡兴华, 毕宇航, 等. 基于 Wi-Fi 和蓝牙数据的公交客流 OD 分析方法研究[J]. 电子设计工程, 2023, 31(22): 153-156.
XU Y M, HU X H, BI Y H, et al. Research on OD analysis method of bus passenger flow based on Wi-Fi and Bluetooth data [J]. Electronic Design Engineering, 2023, 31(22): 153-156.

[6] 徐洋, 孙建忠, 黄磊, 等. 基于 Wi-Fi 定位的区域人群轨迹模型[J]. 山东大学学报(理学版), 2019, 54(5): 8-20.
XU Y, SUN J Z, HUANG L, et al. Trajectory model of area crowd based on Wi-Fi positioning[J]. Journal of Shandong University (Natural Science), 2019, 54(5): 8-20.

[7] 廖嘉欣, 吴启用, 兰小机, 等. 基于 Wi-Fi 探针数据的城市出行轨迹提取[J]. 地球信息科学学报, 2021, 23(11): 1946-1955.
LIAO J X, WU Q Y, LAN X J, et al. Urban travel trajectory extraction based on Wi-Fi probe data[J]. Journal of Geo-Information Science, 2021, 23(11): 1946-1955.

[8] RYU S, PARK B B, EL-TAWAB S. WiFi sensing system for monitoring public transportation ridership: a case study[J]. KSCE Journal of Civil Engineering, 2020, 24(10): 3092-3104.

[9] FRANKLIN J, MCCOY D, TABRIZ P, et al. Passive data link layer 802.11 wireless device driver fingerprinting [C]//Proceedings of the 15th conference on USENIX Security Symposium. Berkeley: USENIX Association, 2006: 16-89.

[10] MATTE C, CUNCHE M, ROUSSEAU F, et al. Defeating MAC address randomization through timing attacks[C] //Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks. New York: Association for Computing Machinery, 2016: 15-20.

[11] 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. Xi′an: ACM,

2016: 413-424.
[12] URAS M, COSSU R, FERRARA E, et al. WiFi probes sniffing: an artificial intelligence based approach for mac addresses de-randomization [C]// 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD). Pisa: IEEE, 2020: 1-6.
[13] PINTOR L, ATZPRI L. Analysis of Wi-Fi probe requests towards information element fingerprinting[C]// GLOBECOM 2022-2022 IEEE Global Communications Conference. Brazil: IEEE, 2022: 38573862.
[14] SIMONCIC A, MOHORCIC M, MOHORCIC M, et al. Non-intrusive privacy-preserving approach for presence monitoring based on WiFi probe requests[J]. Sensors, 2023, 23(5): 2588.

Pedestrian Origin-Destination (OD) Identification Method Based on Wi-Fi Probe Requests Under Randomized MAC Address Mechanisms

ZHU Chengjuan, ZHAO Xingji, CHEN Aitong
(School of Transportation Engineering, Dalian Jiaotong University, Dalian 116028, China

Abstract: To address the privacy protection and data invalidation issues faced by traditional Origin-Destination (OD) crowd flow monitoring methods based on real MAC addresses, this paper proposes an anonymous OD crowd flow identification method based on Wi-Fi probe request information elements. By analyzing stable features in fields such as HT Capabilities and VHT Capabilities within probe requests, we design an enhanced DBSCAN clustering algorithm to accurately distinguish randomized MAC devices. To address cross-region trajectory matching, we introduce an OD association strategy leveraging information element group filtering and distance matrix computation. Experimental results in a public facility environment demonstrate a device clustering accuracy of 97.87% and an OD matching accuracy of 81.8%, validating the method's effectiveness. Compared to existing techniques, our approach achieves high precision and low computational complexity while preserving user privacy, offering novel insights for public safety, transportation planning, and emergency management.

Keywords: pedestrian OD identification; Wi-Fi probe requests; randomized MAC addresses; clustering algorithms; privacy preservation