大规模手机位置数据研究中的个体重识别风险及其与数据可用性的关系

尹 凌 胡金星 王 倩 汪 伟 蔡芷铃

（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 手机位置数据是一种新兴的轨迹数据源，在支持人类移动研究方面具有巨大的潜力。近期研究指出，基于手机用户独特的活动特征，许多用户能够被轻易地重识别。然而，隐私保护处理对原始数据的改变会导致数据可用性的损失。因此，使用详细位置数据进行活动分析的同时避免隐私风险成为一个挑战。本研究旨在揭示中国一个大型城市的手机用户重识别风险，以及将该数据用于人群移动分析时，用户重识别风险和数据可用性之间的量化关系。首先，以深圳市为例，评估全市某一主要运营商手机用户的重识别风险；然后，提出并实现一种空间泛化方法以保护用户隐私；最后，使用人群移动分析为例，评估隐私保护后数据可用性的损失。结果显示，深圳市的重识别风险不同于西方城市，证明了基于手机位置数据的重识别风险具有空间异质性。其次，发现了重识别风险(x)和数据可用性(y)之间的数学关系 $y=-ax^b+c$ ($a, b, c>0$; $0<x<1$)。该关系的发现，为数据发布者在权衡隐私风险和数据可用性之间的关系时提供了科学依据。本研究有助于更好地理解大规模轨迹数据中的个体重识别风险，以及隐私风险与数据可用性之间的权衡基准，有助于降低共享轨迹数据时的隐私风险。

关键词 手机数据；轨迹分析；重识别风险；数据可用性；移动分析

中图分类号 TP 39 **文献标志码** A

Re-Identification Risk Versus Data Utility for Aggregated Mobility Research Using Mobile Phone Location Data

YIN Ling HU Jinxing WANG Qian WANG Wei CAI Zhiling

(*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

Abstract Mobile phone location data is a newly emerging data source of great potential to support human mobility research. However, recent studies have indicated that many users can be easily re-identified based on their unique activity patterns. Privacy protection procedures will usually change the original data and cause a loss of data utility for analysis purposes. Therefore, the need for detailed data for activity analysis while

收稿日期：2015-11-11 **修回日期**：2015-11-17

基金项目：国家自然科学基金项目(41301440)；广东省自然科学基金(2014A030313684)；深圳市基础研究(JCYJ20140610151856728)

作者简介：尹凌(通讯作者)，博士，副研究员，研究方向为时空数据挖掘，E-mail：yinling@siat.ac.cn；胡金星，博士，正高级工程师，研究方向为地理信息系统；王倩，硕士研究生，研究方向为时空数据挖掘；汪伟，硕士研究生，研究方向为时空数据挖掘；蔡芷铃，硕士研究生，研究方向为时空数据挖掘。

avoiding potential privacy risks presents a challenge. The aim of this study is to reveal the re-identification risks from a Chinese city's mobile users and to examine the quantitative relationship between re-identification risk and data utility for an aggregated mobility analysis. The first step was to evaluate the re-identification risks in Shenzhen City, a metropolis in China. A spatial generalization approach to protecting privacy was then proposed and implemented, and spatially aggregated analysis was used to assess the loss of data utility after privacy protection. The results demonstrate that the re-identification risks in Shenzhen City are clearly different from those in regions reported in Western countries, which prove the spatial heterogeneity of re-identification risks in mobile phone location data. A uniform mathematical relationship has also been found between re-identification risk (x) and data utility (y) for both attack models: $y=-ax^b+c$ (a, b, $c>0$; $0<x<1$). The discovered mathematical relationship provides data publishers with useful guidance on choosing the right tradeoff between privacy and utility. Overall, this study contributes to a better understanding of re-identification risks and a privacy-utility tradeoff benchmark for improving privacy protection when sharing detailed trajectory data.

Keywords mobile phone data; trajectory analysis; re-identified risk; data utility; mobility analysis

1 引 言

出于账单记录和故障排查等目的，手机运营商每天都会收集用户的位置数据。这种自动远程获得大量个人轨迹数据的能力，为研究人类移动提供了空前的机会[1]。在过去的几年里，基于手机位置数据进行人类移动研究取得了快速进展，包括总结人类移动模式的普适规律[2-4]、预测人类移动[5,6]、估计人群移动起止（Original-Destination，OD）流[7-9]、模拟人类移动[10]、揭示人口动态和热点[11,12]、识别重要活动地点[13]以及挖掘日常活动结构[14]等。这些研究表明，手机位置数据具有成为人类移动研究主要数据源的潜力，可服务于交通、城市规划、传染病学和社会学等广泛的领域。

然而，轨迹数据中包含的丰富时空间信息可能会导致隐私泄漏。近期两项研究指出，利用匿名的手机位置数据，结合少量的外部信息，个人轨迹的独特性会导致大量用户身份暴露（即个体重识别）。2011 年 MobiCom 会议上，美国 Sprint 手机运营商提出了一种基于前 N 个最频繁活动点的攻击模型，并使用美国主要城市的手机用户通话位置数据进行实验，发现使用用户轨迹中出现最频繁的两个或三个地点，分别能够唯一识别出 10%～50% 的用户[15]。2013 年 Scientific Report[16]发表的一项研究提出了一种基于随机时空点的攻击模型，使用欧洲某国家的手机用户通话数据进行实验，发现在用户轨迹中随机选取四个时空点就可以唯一识别 95% 的用户。

上述研究揭示了严重的轨迹数据个体重识别风险，然而基于手机位置数据等轨迹数据的研究正在日益增加，因此，亟需对使用这些数据时如何有效保护个人隐私进行更深入的研究和讨论[17]。一方面，研究者和政策制订者希望从详细的轨迹数据中获得对社会有益的信息；另一方面，数据发布者必须对原始数据进行隐私保护，保证这些共享数据不会导致隐私泄漏。研究者已经提出了一些方法来保护轨迹隐私[18-21]，例如添加

假数据[22]、抑制或泛化敏感信息[23-27]。然而，现存的研究已经表明，保护的隐私越多，损失的数据可用性越多[28-32]。总的来说，寻找此二者之间的最优平衡是一个挑战。从数据可用性方面来说，一些研究的目标在于分析特定个体的移动特征，例如分辨率在几百米之内的活动地点的识别[13,14]，这种情况下，引入隐私保护方法(例如位置扰动)，对个人轨迹进行改变，会对分析结果造成显著影响。另一方面，许多政策制定并不关注个人，而是关注人群移动的整体模式，例如空间分辨率为几千米的区域之间的人群移动 OD 流，这种应用情况则扩大了隐私保护和数据可用性之间相互妥协的空间。

为了加深对手机位置数据隐私风险的理解，并支持设计数据共享时有效的隐私保护方法，本研究使用中国一个大型城市的匿名手机位置数据，回答三个问题：第一，美国和欧洲国家手机位置数据中存在的高重识别风险是否同样存在于有着不同人口数量、文化和生活方式的中国城市？第二，在 Zang 等[15]和 de Mentjoye 等[16]提出的基于空间泛化的方法在降低重识别风险以后，如何影响对人类移动的分析，尤其是政策制定中重要的人群移动流分析？第三，重识别风险和基于人群移动分析的数据可用性之间是否存在一种统一的量化关系？回答这些问题有助于估计不同区域手机位置数据中的个体重识别风险、理解隐私风险和数据可用性之间的权衡、设计更好的隐私保护方法，以此促进手机位置数据等大规模轨迹数据的可持续使用。

2　数据和方法

2.1　数据集

由于 Zang 等[15]和 de Mentjoye 等[16]提出的重识别风险度量是基于手机通话详单(Call Detail Records，CDR)的，为了对结果进行严谨的比较，本研究亦使用 CDR 数据。CDR 的每条记录均包含用户主动呼叫和被动呼叫时的位置和时间，因此可以获得个人的轨迹信息。本研究使用的 CDR 由深圳交通部门提供，包含了 2011 年 4 570 000 手机用户(占总人口近 30%)的通话记录，时间跨度为 34 天，包括 25 个工作日和 9 个周末。数据集中的每个条目由 6 个字段组成：手机用户的匿名标识、通话事件发生的时间、通话时连接基站的经度和纬度坐标、打/接电话标识(1/0)以及基站所在城市的区号。CDR 自身融合了三种隐私保护方法：对手机用户的电话号码加密；用基站范围泛化手机用户的精确位置；使用通话事件的不规则稀疏采样抑制手机用户轨迹的暴露度。

2.2　重识别风险评估

本研究定义个人隐私风险为个体重识别概率。为了比较不同地区的隐私风险，我们使用 Zang 等[15]和 de Mentjoye 等[16]提出的方法评估数据集中的隐私风险，包括频繁活动点攻击模型和随机时空点攻击模型。

在频繁活动点攻击模型中，根据活动地点出现的频度进行由多至少排序，假定攻击者知道移动对象的前 N 个最频繁活动点。在现实生活中，移动对象的频繁活动点往往是个体活动的重要地点，例如家和工作地，或者是最常去的家附近的超市、医院和学校等。通常个体活动最频繁的前 N 个地点能够揭示个体的职业、经济状况、社会关系甚至健康状况等敏感信息。我们用 k-匿名值来量化个体重识别风险，k 代表具有相同频繁活动点的用户组成的匿名集的大小[24]。重识别风险随 k 变化。在 k 为 1 的极端情况下，重识别风险达到 100%。此外，频繁活动点模型只适用于通话频繁的用户。对于在观察期间很少打或接电话的用户，很难准确推断出他们的前 N 个频繁活动点。研究[15]定义平均每天至少打或接一个电话的人为通话频繁的用户。

在随机时空点攻击模型中，每个移动对象 U 的轨迹 T_u 由多个时空点组成，表示为 $T_u=\{(x_1, y_1, t_1), (x_2, y_2, t_2),\cdots, (x_n, y_n, t_n)\}$。其中，$(x_i, y_i, t_i)$ $(1\leqslant i\leqslant n)$ 表示一个时空点，记录了移动对象在 t_i 时刻的位置 (x_i, y_i)；n 表示移动对象 U 的轨迹中时空点的个数。该模型假设攻击者知道目标对象移动轨迹中的某些时空点。该模型下的隐私风险同样用 k 匿名值度量，此时 k 代表由有相同时空点的用户组成的匿名集的大小。同理，当 k 为 1 时，匿名集中个体的重识别风险为 100%。随机时空点模型适用于所有手机用户。

能被唯一识别的人的比率 x($x=k$ 为 1 的人数/总人数)能有效表示人口的重识别风险[16]。本研究将这一比率用于评估隐私保护方法降低隐私风险的效果，以及探索隐私风险和数据可用性之间的量化关系。

2.3 隐私保护过程

基于降低移动对象空间位置精度的思想，本研究提出了一种空间泛化的隐私保护方法。如图 1(a)所示，首先将一个基站的服务区域用一个泰森多边形(Voronoi Polygon)近似表示。然后，创建一个特定空间尺度的网格(图 1(b))，将重心点位于同一个特网格中的基站服务范围聚合为一个更大的区域(图 1(c))，聚合后的区域作为数据发布的新空间单元(图 1(d))。如图 1，基站 2(lat_2, $long_2$, t)处的原始 CDR 记录将会变成({(lat_1, $long_1$),(lat_2, $long_2$),(lat_3, $long_3$)}, t)。以更大的空间范围作为分析单元，一个用户则更可能和别的用户共享同一分析单元，由此可降低用户轨迹的独特性。

2.4 数据可用性分析

由于人群移动 OD 流是交通分析及城市计算研究领域内至关重要的输入数据，因此本研究选择 OD 流作为衡量基站聚合后数据可用性的指标。

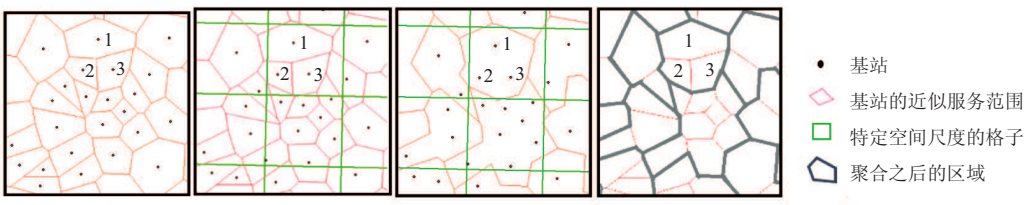
为了评估数据可用性的损失 E，需要对比使用隐私保护方法前后 OD 流的变化。本研究使用 T 时段中 OD 流的绝对误差总量占总 OD 流量的百分比表示 E，如公式(1)所示。

$$E=\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|OD_{ij}-OD'_{ij}\right|}{\sum_{i=1}^{N}\sum_{j=1}^{N}OD_{ij}} \tag{1}$$

其中，OD_{ij} 表示根据原始数据统计得到的 T 时段内从区域 i 到区域 j 的 OD 流量；OD'_{ij} 表示由隐私保护后的数据统计得到的相应的值；N 表示区域的总个数。

公式(1)中的区域单元使用由城市交通规划单位划分的交通分析小区(Traffic Analysis Zone，TAZ)。为了比较不同空间尺度对数据可用性的影响，本研究使用了两个 TAZ 集，一个包含 1 112 个 TAZ，平均半径为 636 m；另一个包含 491 个 TAZ，平均半径为 994 m。其中，TAZ 的半径由面积与之相等的圆的半径确定。为了展示对整个数据集的影响，时间段 T 设置为整个 CDR 数据集的时间跨度(即 34 天)。

在进行隐私保护之前，需要统计每个用户从一个基站到另一个基站的移动情况。具体



(a)原始基站 (b)网格划分 (c)基站聚合 (d)聚合结果

图 1 基于基站聚合的手机用户隐私保护方法

Fig. 1 Privacy protection method for mobile phone users based on aggregating base stations

方法是将每个移动对象所有的通话记录按照时间先后排序，找到连续的通话记录中的位置移动，将一次移动中时间点在前的位置视作出发点(Origin，O)，时间点在后的位置视作目的地(Destination，D)。由于通话事件已经被证明不是随时间均匀分布的[33]，并且 CDR 数据集只包含用户通话事件发生时的地理位置，这种稀疏的采样方法极大地影响 OD 流估计的准确性。例如，一个人早上在位置 A 打了一个电话，随后几次移动都没有打电话，晚上在位置 B 打了一个电话。那么，使用该方法提取的从 A 到 B 的 OD 流是不合理的。此外，相邻基站之间的信号跳转也会产生不合理的短距离移动信息[14]。为了减少此类错误，本研究对每条 OD 记录的时间跨度进行了限制。基于深圳市的出行调查数据(调研时间比手机运营商提供的数据集早半年)，98% 的出行时间在 5～100 分钟。因此，本研究将有效的 OD 出行时间阈值设定为 5～100 分钟，排除不满足该时间限制的 OD 出行。然后将 OD 流矩阵的空间单元由基站转换成 TAZ，使用公式(2)计算从 TAZ_i 到 TAZ_j 的流量 OD_{ij}^{TAZ}：

$$OD_{ij}^{\mathrm{TAZ}}=\sum_{k=1}^{M}\sum_{l=1}^{M}\frac{A_i^k}{A^k}\cdot\frac{A_j^l}{A^l}\cdot OD_{kl}^{\mathrm{BS}} \tag{2}$$

其中，基站 k 覆盖面积为 A^k，与 TAZ_i 重叠的面积为 A_i^k；基站 l 覆盖面积为 A^l，与 TAZ_j 重叠的面积为 A_j^l；OD_{kl}^{BS} 是从基站 k 到基站 l 的流量；M 是基站总数。

实施隐私保护后，假设有 S 个聚合基站。第一步是使用公式 (3) 计算聚合基站 p 到 q 的流量 OD_{pq}^{AgrBS}。然后使用公式 (4) 计算隐私保护后基于 TAZ 的流量 OD'^{TAZ}_{ij}。最后使用公式 (1) 计算数据可用性损失：

$$OD_{pq}^{\mathrm{AgrBS}}=\sum_{u=1}^{U}OD_{uv}^{\mathrm{BS}} \tag{3}$$

其中，聚合基站 p 包含 U 个基站；聚合基站 q 包含 V 个基站；OD_{uv}^{BS} 表示基站 u 到基站 v 的流量。

$$OD'^{\mathrm{TAZ}}_{ij}=\sum_{p=1}^{S}\sum_{q=1}^{S}\frac{A_i^p}{A^p}\cdot\frac{A_j^q}{A^q}\cdot OD_{pq}^{\mathrm{AgrBS}} \tag{4}$$

其中，聚合基站 p 面积为 A^p，与 TAZ_i 重叠的面积为 A_i^p；聚合基站 q 面积为 A^q，与 TAZ_j 重叠的面积为 A_j^q。

3　结果和讨论

3.1　个体重识别风险

对于频繁活动点攻击模型，本研究计算了每个频繁通话用户的匿名集规模，并且画出当 $K\leqslant k$ 时匿名集人数的百分比。如图 2 所示，匿名集规模随频繁活动点数量 N 的增大而减小，说明增大已知频繁活动点数目 N 会增加重识别风险。如图 2 和表 1 所示，当排序第一的最频繁活动点被暴露时，几乎没有用户能够被唯一识别；然而，当前 2 个和前 3 个最频繁活动点被暴露时，能唯一识别的人数分别增加至 17% 和 49%。

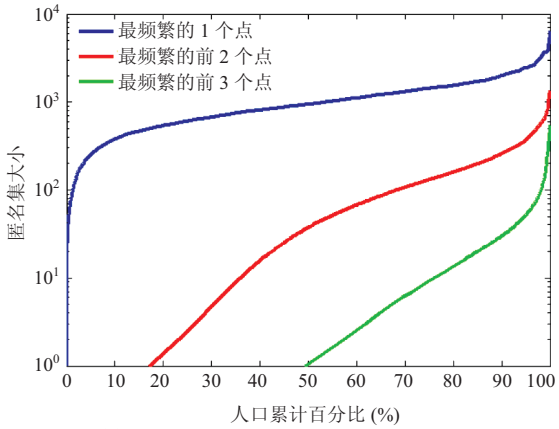


图 2　频繁活动点攻击模型下的手机用户匿名集规模

Fig. 2　Size of anonymity set under the attack model of top *N* locations

如表 1 所示，对于频繁通话用户的重识别风险，深圳与美国的整体情况极其相似。然而，对于所有手机用户，在已知前 2 个频繁活动点的情况下，在深圳能唯一识别的人数仅是美国的一半；在已知前 3 个频繁活动点的情况下，被唯一

表 1 深圳和美国匿名值为 1 的人口百分比

Table 1 Percentage of population with *k*-anonymity value＝1 in Shenzhen City and the United States

用户组	地区	前 1 个频繁活动点（%）	前 2 个频繁活动点（%）	前 3 个频繁活动点（%）
频繁通话用户	深圳	0	17	49
	美国	0	15	50
所有用户	深圳	0	6	18
	美国	0	12	40

识别人数低于美国的一半。究其原因，深圳的手机用户中只有 37% 的频繁通话用户，而美国则有 80% 的频繁通话用户。深圳非活跃手机用户的比例如此之高，可能和如下几种情况相关，例如深圳的手机用户比美国的用户更少打电话，或者深圳许多用户使用多个手机号码导致有的手机号码没有频繁使用。上述的情况有可能是导致深圳用户隐私风险低于美国用户的原因。

美国通话频繁用户的数据来自 Zang 等[15]的研究。该研究没有列出美国所有用户的数据，这部分数据由本文作者根据美国通话频繁用户匿名值为 1 的比例和频繁通话用户占总人口的比例计算得来。

本研究还与 Zang 等[15]研究中不同城市匿名集大小为 1 的频繁通话用户的比例进行了比较，如表 2 所示。深圳匿名集大小为 1 的频繁通话用户的百分比介于萨克拉门托和芝加哥之间。Zang 等[15]认为重识别风险可能与城市和乡村生活方式的不同有关。更城市化的生活导致更多样化的活动，因此产生规模更小的匿名集。根据这一解释，深圳城市化生活程度可能介于萨克拉门托和芝加哥之间。

表 2 不同城市匿名值为 1 的人口百分比(*N*＝2)

Table 2 Percentage of frequent users with *k*-anonymity value＝1 in different cities (*N*＝2)

城市	匿名值为 1 的人口百分比（%）*	人口密度（人/m^2）
深圳	17	20 205
堪萨斯城	10	1 168
萨克拉门托	15	4 660
芝加哥	25	11 868
洛杉矶	25	12 451
旧金山	30	17 246

*美国不同城市匿名值为 1 的人口百分比来自 Zang 等[15]研究。美国不同城市人口密度的数据是基于 2010 年美国人口普查的结果估计得来，对城市边界的划分不同可能会得到不同结果

如图 3 所示，在随机时空点攻击模型中，随着随机时空点个数逐渐增加，被唯一重识别的人口百分比急剧增长。4 个随机点就可以唯一重识别深圳 75% 左右的手机用户。当时空点的个数增加到 4 个之后，隐私风险的增长幅度变缓，逐渐趋于收敛，唯一重识别人口百分比的最大值在 80% 左右。

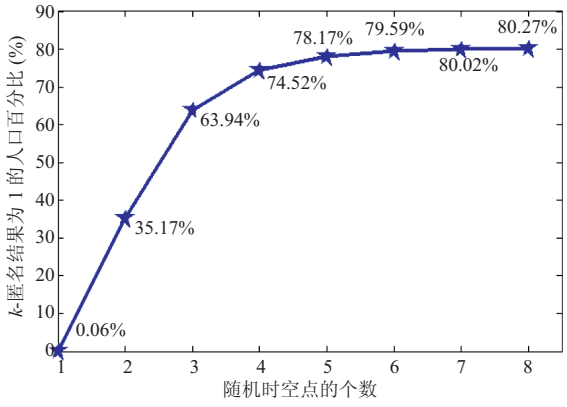


图 3 随机时空点攻击模型下的手机用户匿名集规模

Fig. 3 Size of anonymity set under the attack model of random spatiotemporal points

De Mentjoye 等[16]指出，4 个随机时空点就足以唯一识别欧洲某一小型国家 95% 的手机用户。该研究认为，他们的研究结果可推广到人口密度更高的区域。该推论基于以下两方面考虑：

一方面，更高的人口密度可能会增大匿名集；另一方面，人口密度大的地方会建设更多基站，因此会增加轨迹的空间精度，这将抵消高人口密度带来的影响。然而，本研究的结果显示，基于随机时空点模型，深圳的手机用户隐私风险明显比欧洲区域低，而作为中国人口密度最高的城市，深圳的人口密度应当比小型欧洲国家高。根据本研究的结果，高人口密度致使更多基站被设立，但这不能完全抵消高人口密度带来的影响。此外，如表 2 所示，在美国，不同城市的个体重识别风险随人口密度增大而增加。因此，综合本研究和 Zang 等[15]、de Mentjoye 等[16]研究的结果，阐释重识别风险的空间异质性，除了人口密度，还需要考虑其他因素，如城市化生活程度。

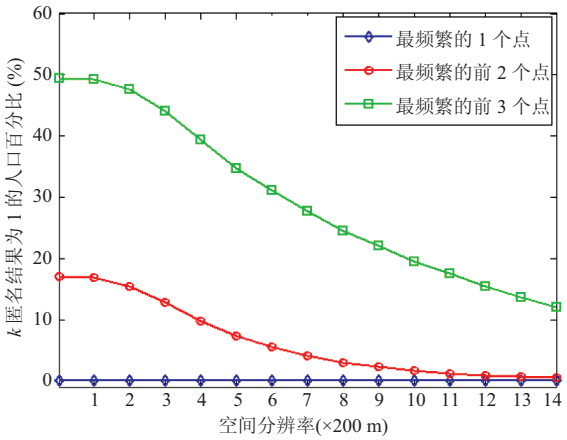
3.2　隐私保护措施对重识别风险的影响

实验结果表明，本研究提出的空间泛化方法能有效地降低隐私风险。如图 4(a)所示，对于频繁活动点攻击模型，随着空间聚合范围增大，N=1 时隐私风险的变化不明显；当 N=2 时，唯一重识别人口百分比明显下降；N=3 时下降更为显著。当空间分辨率降低到 2 800 m 时，N=2 时唯一重识别人口比例从最开始的 17% 下降到 1%；N=3 时则从 49% 下降到 12%。

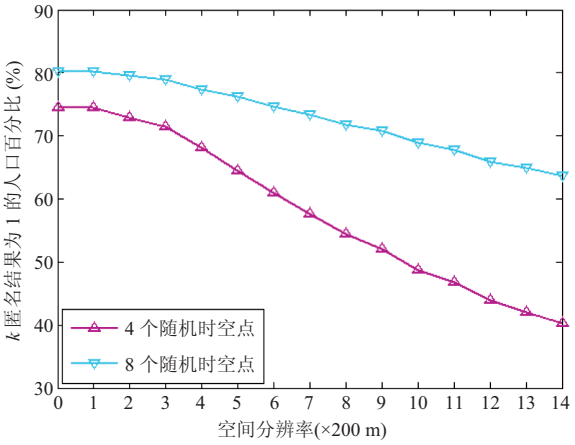
对于随机时空点攻击模型，由于时空点个数为 4 和 8 时是图 3 曲线上两个重要的转折点，因此本研究选择这两种情况进行探讨。当空间分辨率降低到 2 800 m 时，对于唯一重识别的人口百分比，在 4 个点时从 75% 下降到 40%，8 个点时从 80% 下降到 64%。8 个时空点时人口比率下降缓慢，说明当攻击者知道更多轨迹点时，空间泛化方法降低隐私风险的难度增加。

3.3　隐私保护措施对数据可用性的影响

如图 5 所示，当空间分辨率从 200 m 降低到 2 800 m 时，数据可用性损失急剧增大。此外，数据可用性在较大的空间单元(491 个 TAZ，平



(a) 频繁活动点攻击模型



(b) 随机时空点攻击模型

图 4　实施隐私保护后重识别风险的下降曲线

Fig. 4　Privacy risk reduction after protecting privacy

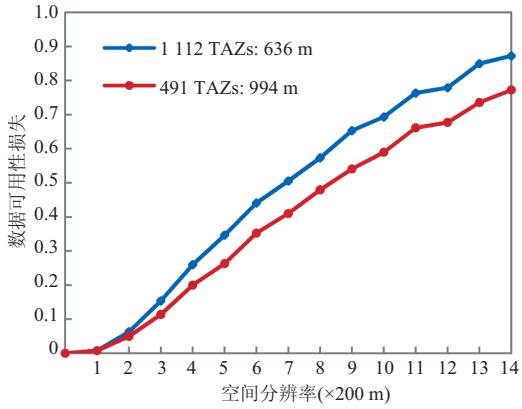


图 5　实施隐私保护后数据可用性损失的上升曲线

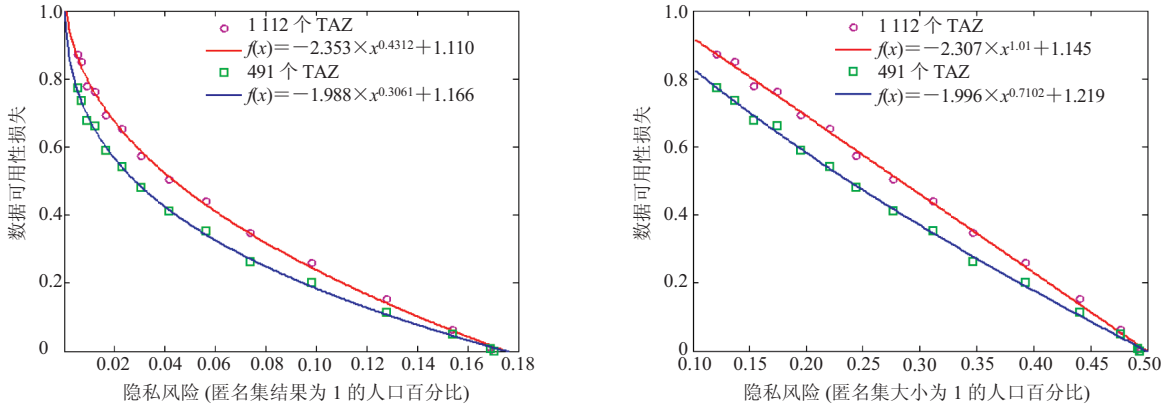
Fig. 5　Data utility loss after protecting privacy

均半径 994 m)上的损失也明显小于较小的空间单元(1 112 个 TAZ，平均半径 636 m)。

3.4 重识别风险和数据可用性之间的关系

为了揭示隐私风险降低与数据可用性损失之间的权衡，尝试量化两者之间的关系。如图 6 和图 7 所示，我们发现两种不同攻击模型下的隐私风险(唯一重识别的人口百分比，用 x 表示)和数据可用性损失(用 y 表示)之间的关系可以用函数 $y=-ax^b+c(a, b, c>0; 0<x<1)$ 表示。用上述函数拟合结果的显著性很高($P<0.001$)，即该函数对两者关系的拟合效果很好。

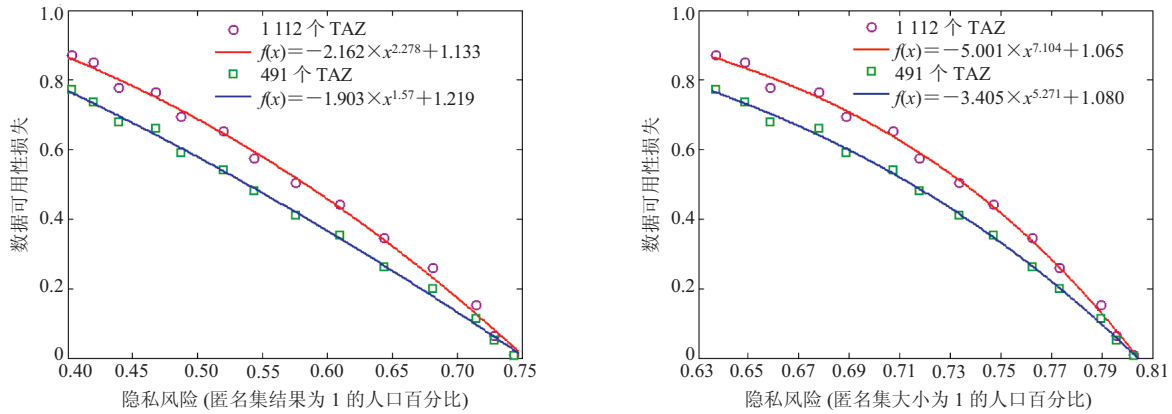
拟合函数中的指数 b 影响拟合曲线的形状。当 $0<b<1$ 时，如图 6(a)所示，若想降低隐私风险(可以从图的 x 轴从右往左看过去)，由此带来的数据可用性损失的代价在开始阶段相对较小，随着风险逐渐减小，数据可用性的损失速度逐渐升高。当 b 接近 1 时，如图 6(b)中红色曲线所示，隐私风险和数据可用性损失之间的关系接近于线性的。当 $b>1$ 时，如图 7 所示，数据可用性损失的增长幅度一开始是剧烈的，随着聚合程



(a)前 2 个频繁点 (b)前 3 个频繁点

图 6 频繁活动点攻击模型下隐私风险与数据可用性的关系

Fig. 6 Relationships between privacy risk and data utility for top *N* locations



(a)4 个随机点 (b)8 个随机点

图 7 随机时空点攻击模型下隐私风险与数据可用性的关系

Fig. 7 Relationships between privacy risk and data utility for random spatiotemporal points

度提高，数据可用性损失的增长幅度变缓。总体而言，较小的指数 b 在空间聚合的初级阶段可以在数据可用性方面提供较好的折衷。

从前 2 个频繁点、前 3 个频繁点、4 个随机点到 8 个随机点攻击模型，指数 b 的值逐渐增大，这与攻击者背景知识增多的趋势是一致的，同时表明相应的隐私保护难度也逐渐升高。

4　贡献和展望

本研究的贡献包括以下两个方面。第一，该研究证明空间异质性不仅出现在频繁活动点攻击模型中，对随机时空点攻击模型同样有效。该发现进一步证明不同地区通话详单中的重识别风险是不同的。因此，在共享个人轨迹数据的时候，需要根据当地情况考虑保护隐私的方法和体系。第二，不管是频繁活动点模型还是随机时空点模型，本研究发现了一种二者皆适用的重识别风险和数据可用性之间的关系：$y=-ax^b+c(a, b, c>0; 0<x<1)$，其中指数 b 随着攻击者背景知识的增加而增大，b 值的增大意味着降低重识别风险的开始阶段越困难。这一关系给其他基于手机位置数据的分析提供了权衡的参考，同时也为数据发布者提供了平衡隐私风险和数据可用性两者关系的依据。

未来研究中尚有几个需要解决的问题：(1) 基于本研究和 Zang 等[15]、de Mentjoye 等[16]的发现，手机位置数据中的重识别风险的空间异质性可能是由多种原因导致的，例如手机使用习惯、人口密度和生活方式等，为了更深入理解影响个体重识别风险的因素，需要深入探索手机位置数据的隐私风险和可能的影响因素之间的量化关系；(2) 本研究提出的空间泛化方法能够降低手机位置数据的重识别风险，然而实施该方法处理以后，人群移动分析的数据可用性急剧下降，因此，需要研究更加复杂和精妙的隐私保护方法；(3) 除了人群移动分析，隐私保护对个体级别分析的影响需要进一步的研究；(4) 如果数据可获得，与其他地区的用户隐私风险进行比较的研究也是有意义的。

参 考 文 献

[1] Jiang S, Fiore GA, Yang Y, et al. A review of urban computing for mobile phone traces: current methods, challenges and opportunities [C] // Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013: 2.

[2] Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns [J]. Nature, 2008, 453(7196): 779-782.

[3] Phithakkitnukoon S, Smoreda Z, Olivier P. Socio-geography of human mobility: a study using longitudinal mobile phone data [J]. PloS One, 2012, 7(6): e39253.

[4] Kang C, Ma X, Tong D, et al. Intra-urban human mobility patterns: an urban morphology perspective [J]. Physica A: Statistical Mechanics and Its Applications, 2012, 391(4): 1702-1717.

[5] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility [J]. Science, 2010, 327(5968): 1018-1021.

[6] Qin SM, Verkasalo H, Mohtaschemi M, et al. Patterns, entropy, and predictability of human mobility and life [J]. PloS One, 2012, 7(12): e51353.

[7] Wang MH, Schrock SD, Vander Broek N, et al. Estimating dynamic origin-destination data and travel demand using cell phone network data [J]. International Journal of Intelligent Transportation Systems Research, 2013, 11(2): 76-86.

[8] Iqbal MS, Choudhury CF, Wang P, et al. Development of origin-destination matrices using mobile phone call data [J]. Transportation Research Part C: Emerging Technologies, 2014, 40: 63-74.

[9] Friedrich M, Immisch K, Jehlicka P, et al. Generating origin-destination matrices from mobile phone trajectories [J]. Transportation Research Record: Journal of the Transportation Research Board, 2010 (2196): 93-101.

[10] Buckee CO, Wesolowski A, Eagle NN, et al. Mobile phones and malaria: modeling human and parasite travel [J]. Travel Medicine and Infectious Disease, 2013, 11(1): 15-22.

[11] Ahas R, Aasa A, Silm S, et al. Daily rhythms of suburban commuters'movements in the Tallinn metropolitan area: case study with mobile positioning data [J]. Transportation Research Part C: Emerging Technologies, 2010, 18(1): 45-54.
[12] Kang C, Liu Y, Ma X, et al. Towards estimating urban population distributions from mobile call data [J]. Journal of Urban Technology, 2012, 19(4): 3-21.
[13] Isaacman S, Becker R, Cáceres R, et al. Identifying important places in people's lives from cellular network data [M] // Pervasive Computing. Springer Berlin Heidelberg, 2011: 133-151.
[14] Schneider CM, Belik V, Couronné T, et al. Unravelling daily human mobility motifs [J]. Journal of the Royal Society Interface, 2013, 10(84): 20130246.
[15] 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, 2011: 145-156.
[16] de Montjoye YA, Hidalgo CA, Verleysen M, et al. Unique in the crowd: the privacy bounds of human mobility [J]. Scientific Reports, 2013, 3: 1376.
[17] Butler D. Data sharing threatens privacy [J]. Nature News, 2007, 449(7163): 644-645.
[18] Wernke M, Skvortsov P, Dürr F, et al. A classification of location privacy attacks and approaches [J]. Personal and Ubiquitous Computing, 2014, 18(1): 163-175.
[19] Giannotti F, Pedreschi D. Mobility, Data Mining and Privacy: Geographic Knowledge Discovery [M]. Springer Science & Business Media, 2008.
[20] Fung B, Wang K, Chen R, et al. Privacy-preserving data publishing: a survey of recent developments [J]. ACM Computing Surveys (CSUR), 2010, 42(4): 14.
[21] Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories [C] // The 9th International Conference on Mobile Data Management, 2008: 65-72.
[22] Hoh B, Gruteser M. Protecting location privacy through path confusion [C] // First International Conference on Security and Privacy for Emerging Areas in Communications Networks, 2005: 194-205.
[23] Chen R, Fung BCM, Mohammed N, et al. Privacy-preserving trajectory data publishing by local suppression [J]. Information Sciences, 2013, 231: 83-97.
[24] Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 571-588.
[25] Nergiz ME, Atzori M, Saygin Y. Towards trajectory anonymization: a generalization-based approach [C] // Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS, 2008: 52-61.
[26] Mohammed N, Fung B, Debbabi M. Walking in the crowd: anonymizing trajectory data for pattern analysis [C] // Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009: 1441-1444.
[27] Domingo-Ferrer J, Sramka M, Trujillo-Rasúa R. Privacy-preserving publication of trajectories using microaggregation [C] // Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, 2010: 26-33.
[28] Rastogi V, Suciu D, Hong S. The boundary between privacy and utility in data publishing [C] // Proceedings of the 33rd International Conference on Very Large Data Bases, 2007: 531-542.
[29] Brickell J, Shmatikov V. The cost of privacy: destruction of data-mining utility in anonymized data publishing [C] // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008: 70-78.
[30] Loukides G, Shao J. Data utility and privacy protection trade-off in *k*-anonymisation [C] // Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, 2008: 36-45.
[31] Li T, Li N. On the tradeoff between privacy and utility in data publishing [C] // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009: 517-526.
[32] Guo S, Chen K. Mining privacy settings to find optimal privacy-utility tradeoffs for social network services [C] // 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), and 2012 International Confernece on Social Computing (SocialCom), 2012: 656-665.
[33] Barabasi AL. The origin of bursts and heavy tails in human dynamics [J]. Nature, 2005, 435(7039): 207-211.