软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(Suppl.(1)):18−26

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于时空关联和位置语义的个性化假位置生成方法*

周佳琪, 李燕君

(浙江工业大学 计算机科学与技术学院,浙江 杭州 310023)

通讯作者: 李燕君, E-mail: yjli@zjut.edu.cn

摘 要: 基于假位置的一类隐私保护方案在保护用户位置隐私的同时能够使用户获得准确查询信息,并无需依赖第三方和共享密钥.然而,当攻击者掌握一定的背景知识,例如道路时空可达信息、位置特征和用户的历史请求统计特性等,会导致假位置被识别的概率升高,降低隐私保护程度.针对上述问题,提出了基于时空关联和位置语义的个性化假位置生成算法.首先根据与前一次请求位置连续可达的条件产生假位置,然后通过建立语义树筛选出与真实位置语义相近的假位置,最后进一步筛选出与用户历史请求统计特性最接近的假位置.基于真实数据集将该算法与现有的算法进行比较,表明该算法在攻击者掌握相关背景知识的情况下,可以有效地降低位置隐私泄露的风险.

关键词: 假位置;时空关联;位置语义;历史信息;隐私保护

中文引用格式: 周佳琪,李燕君.基于时空关联和位置语义的个性化假位置生成方法.软件学报,2019,30(Suppl.(1)):18−26. http://www.jos.org.cn/1000-9825/19003.htm

英文引用格式: Zhou JQ, Li YJ. Personalized Dummy Generation Method Based on Spatiotemporal Correlations and Location Semantics. Ruan Jian Xue Bao/Journal of Software, 2019,30(Suppl.(1)):18−26 (in Chinese). http://www.jos.org.cn/1000-9825/19003.htm

Personalized Dummy Generation Method Based on Spatiotemporal Correlations and Location Semantics

ZHOU Jia-Qi, LI Yan-Jun

(School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Without the need for the third party and key sharing, the dummy-based privacy protection scheme enables the users to obtain precise query results while protecting their location privacy. However, when the adversary has certain background knowledge, e.g., the spatiotemporal reachability information, the location semantics, the users' historic query statistics, the probability of dummies being inferred will rise and the degree of privacy protection will be reduced. To solve this problem, a personalized dummy generation method based on spatiotemporal correlations and location semantics is proposed. Dummies are first generated based on the continuous reachability with previous request locations, and then filtered through the check of location semantic similarity and finally filtered by accessibility to user's historic query statistics. Experiments based on real datasets show that the proposed dummy generation method can effectively reduce the risk of privacy disclosure compared with current two dummy generation methods, especially when the adversary has related background knowledge.

Key words: dummy; spatiotemporal correlation; location semantics; historical information; privacy protection

随着无线通信和定位技术的发展以及智能移动终端的普及,基于位置的服务(locatioased services,简称 LBS)已成为最流行的移动终端服务之一.在 LBS 中,用户将指定的地理位置信息和服务请求发送给服务提供商(location-based service provider,简称 LSP),由 LSP 返回满足用户要求的查询结果.然而,用户在享受 LBS 带来便

* 基金项目: 国家自然科学基金(61772472, 61872322, 61472367); 浙江省自然科学基金(LY17F020020); 浙江省属高校基本科研业务费专项资金(RF-A2019002)

Foundation item: National Natural Science Foundation of China (61772472, 61872322, 61472367); Natural Science Foundation of Zhejiang Province (LY17F020020); Fundamental Research Funds for the Provincial Universities of Zhejiang Province (RF-A2019002)

收稿时间: 2019-09-15; 采用时间: 2019-10-24

利的同时,也面临着隐私信息被窃取滥用的风险.攻击者通过分析位置信息中包含的时空信息,结合已掌握的背景知识,可以挖掘出与用户查询位置相关的隐私信息,如健康状况、宗教信仰、家庭及工作地址等,给用户带来潜在威胁.

近来,随着LBS中暴露出来的隐私问题越来越严重,位置隐私受到了国内外研究者的广泛重视,提出了许多对位置隐私进行保护的技术手段[1].其中基于假位置的隐私保护技术是较常采用的方案之一[2,3],它是指用户在发送服务请求时,产生多个假位置,与真实位置一起发送给 LSP,攻击者即使截获所有信息也不能分辨出用户的真实位置.相对于其他方案,如 *k*-匿名、差分隐私、密码学方案等,基于假位置的隐私保护技术有如下优势:1) 能够提供精确的查询结果;2) 无需依赖第三方;3) 用户和 LSP 之间无需共享密钥.在基于假位置的隐私保护方案中,假位置的质量决定了其隐私保护能力的强弱,如何避免假位置被攻击者识别是一个重要的研究问题.考虑到攻击者掌握一定的背景知识,例如道路时空可达信息、位置特征、用户的历史请求行为等,许多学者在随机产生假位置的基础上做出了改进.一些文献考虑在连续请求服务时,基于时间和空间的关联性判断假位置的连续可达性,以此来约束假位置集合;一些文献基于位置特征提出产生的假位置应与真实位置在语义上尽可能地区分,避免因为语义的重合而暴露用户的隐私.目前很少有文献综合考虑上述两个方面,我们通过基于真实数据集的实验发现,基于时空关联产生的假位置与真实位置在同一语义上的概率超过 29%,而仅考虑位置语义区分的假位置由于连续不可达被识别出的概率超过 45%.此外,当攻击者掌握用户的历史请求行为统计特性时,若产生的假位置为离群点,则很容易被识别.

针对上述问题,本文提出基于时空关联和位置语义的个性化假位置生成方法.首先根据与上一时刻位置连续可达的条件产生一定数量的假位置,然后通过建立语义树筛选出与真实位置语义相近的假位置,最后在此基础上进一步筛选出与用户历史请求统计特性最接近的假位置,本文的贡献在于:1) 综合考虑了时空关联性、位置语义差异性和用户的历史请求行为统计特性,提出了假位置生成算法;2) 基于真实的数据集将本文提出的算法与现有的算法进行比较,表明本文提出的算法有效降低了位置隐私泄露的风险.

1 相关工作

针对连续请求服务的场景,考虑攻击者掌握位置可达性信息,Liu 等人[4,5]通过时间可达、方向控制、出入度控制这 3 个约束来产生假位置点,降低假位置被识别的概率.李维皓等人[6]考虑了连续请求间的时空关联性,利用用户移动模型预测用户在下一个时间段可能出现的位置,以决定前一时刻的请求内容,提高假位置对攻击者的混淆程度.Takbiri 等人[7]通过跟踪分析用户真实位置的时间相关性,提出与之匹配的用户匿名和数据模糊尺度,以获得满足用户需求的隐私保护.上述工作仅在攻击者单纯掌握位置可达性背景知识时可提高位置隐私保护程度,但对于攻击者掌握其他背景知识时无法提供有效的保护.

考虑攻击者掌握位置特征,Chen 等人[8]通过构建语义树选择与真实位置在语义上相距较远的假位置,避免因为语义的重合而暴露用户的隐私.Li 等人[9]用分段时间访问量作为位置特征,通过判断分段时间访问量的相似程度来衡量两个位置在语义上的差距,作为选择假位置的指标.上述工作仅在攻击者单纯掌握位置特征背景知识时可提高位置隐私保护程度,但无法应对攻击者掌握其他背景知识的情况.

考虑攻击者掌握用户的真实历史请求数据,Niu 等人[10−12]提出筛选出与真实位置具有相同查询概率的位置作为假位置,并使生成的假位置分布尽可能地远.Hayashida 等人[13]提出生成与用户真实移动模式相近的假轨迹,使攻击者难以区分真人与假人.上述工作仅在攻击者单纯掌握用户历史请求信息时可提高位置隐私保护程度.

随着隐私攻击方式的日益多样化,攻击者可能同时掌握多种背景知识,在选择假位置时如果没有采用完善的选择策略,假位置提供的隐私保护能力相当脆弱.现有方案绝大多是针对攻击者拥有某一类背景知识进行假位置的筛选,在攻击者拥有复合背景知识的场景下,这些假位置有很大概率会被识别出来.针对上述问题,本文综合考虑了时空关联性、位置语义差异性和用户的历史请求行为统计特性,提出了假位置生成算法,可有效降低攻击者在拥有复合背景知识情况下假位置被识别的风险.

2 假位置生成方法

首先介绍了系统架构,然后针对随机产生的假位置分别基于时空关联性、位置语义和用户历史请求分布进行筛选,得到最终的假位置集合,总结假位置生成算法并分析了算法复杂度.

2.1 系统架构

由于地图语义位置信息庞大,移动端受到存储容量限制,不便直接存储.随着 5G 基础设施的建设与发展,“宏站+小站”的组网覆盖模式将日益普及,小基站与 WiFi 热点融合将成为日后的发展趋势.小基站具有一定的存储与计算能力,因此我们通过小基站来存储其附近的地图语义信息.本文的系统架构如图 1 所示.一个基站只用于维护其附近一定距离内的位置语义信息,并以语义树的形式进行存储维护.当用户需要使用 LBS 时,首先会向最近的基站请求获取当地的地图语义信息,基站在收到请求后会以位置语义树的形式向移动端发送其内部存储的地图语义信息.移动端收到位置语义树后,调用假位置生成方法获得假位置集合,将假位置和真实位置一起发送至 LSP 请求服务.LSP 收到服务请求后,进行处理并返回服务结果集合,移动端从结果集合中筛选出需要的服务信息.

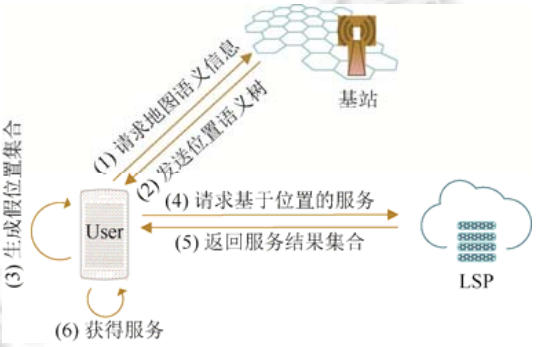


Fig.1 System architecture

图 1 系统架构

2.2 时空关联性筛选

在许多场景中,用户需要连续使用 LBS,此时相邻时间戳下的请求间往往存在时空关联性.有些假位置由于在请求时间间隔内无法到达,极易被排除,降低了位置隐私保护程度.如图 2 所示,A,B,C,D 为用户在 T_{i-1} 时刻请求服务时上传的位置集合,E,F,G,H 为 T_i 时刻上传的位置集合,其中 B,G 为真实位置.LSP 能够准确地知道相邻请求的时间间隔 $\Delta T=T_i-T_{i-1}$.由于地理情况的限制,T_{i-1} 时刻产生的位置点集合中不存在能够在 ΔT 时间内到达 E 的位置点,因此 E 能被轻易排除.

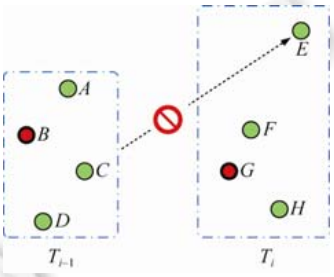


Fig.2 An example of time reachable attack against location dummies

图 2 一个通过时间可达性识别出假位置的例子

通过时空关联性筛选,可以确保上一时刻请求位置集合中至少存在 1 个位置点可以在请求时间间隔内到

达当前时刻的假位置,避免图 2 所示情况的发生.首先,用 $Loc_{i-1}=\{l_{i-1}^{(1)},l_{i-1}^{(2)},...,l_{i-1}^{(K_{i-1}-1)}\}\cup\{l_{i-1}^{(K_{i-1})}\}$ 表示第 i–1 次请求上传的位置集合,其中,$l_{i-1}^{(K_{i-1})}$ 表示第 i–1 次请求时的真实位置.用 $Loc_i^{cr}=\{l_i^{r(1)},l_i^{r(2)},...\}$ 表示第 i 次经过时空关联筛选后的位置集合,Loc_i^{cr} 满足式(1)描述的时空关联性筛选规则.

$$\begin{cases}\left|Loc_i^{cr}\right|\geqslant N(K_{i-1})\\ \forall l_i^{r(k)}\in Loc_i^{cr}(k=1,2,...,\left|Loc_i^{cr}\right|),\exists l_{i-1}^{(j)}\in Loc_{i-1}(j=1,2,...,K_{i-1}):\left|T(l_{i-1}^{(j)},l_i^{r(k)})-\Delta T\right|\leqslant\alpha\Delta T\end{cases}\tag{1}$$

其中,$|\bullet|$ 表示集合中元素的个数,K_i 表示第 i 次请求时的隐私保护需求,即在请求上传的位置集中至少包括 K_i–1 个假位置,N 表示候选假位置集合空间系数,即要求基于时空关联性筛选出的假位置候选集包含的位置数量为 $N(K_i-1)$,目的是给下一阶段位置语义筛选环节提供足够的筛选空间,α为用户自定义的可达时间误差系数,α 设置得越小越接近实际可达时间,$\Delta T=T_i-T_{i-1}$ 表示第 i 个与第 i–1 个请求之间的时间间隔,$T(l_{i-1}^{(j)},l_i^{r(k)})$ 表示 $l_{i-1}^{(j)}$ 和 $l_i^{r(k)}$ 两个位置点间的实际可达时间.

2.3 位置语义相似性筛选

随着数据挖掘技术的不断发展,LBS 中的隐私威胁不仅仅局限于真实位置的暴露,更多情况下是位置特征信息的暴露.即使产生的假位置与真实位置在地理距离上足够远,但如果假位置与真实位置的特征一致,就失去了隐私保护意义.这里,我们用位置语义来描述位置特征.如图 3 所示,A 为真实位置,B,C,D 为假位置,但它们的位置语义都与医疗相关,这样,攻击者即使无法知道用户的真实位置,也可以推断出用户可能生病的隐私信息.

为了准确地描述位置点间的语义差别,通过建立位置语义树来衡量两个位置间的语义距离.其中,叶子结点表示真实地理位置,非叶子结点表示语义分类.两个叶子结点间经过的路径数表示这两个位置的语义距离.图 4 为一棵 4 层语义树的一部分,第 1 层为根结点,第 2 层为语义主分类,第 3 层为语义细分类,其中,A 酒店到 B 学校经过了酒店、餐饮、根、教育、学校这 5 个结点,路径数为 6,所以在这棵位置语义树上,A 酒店和 B 学校的语义距离为 6.

Fig.3 An example of privacy exposure based on semantic similarity

图 3 一个基于语义相似性暴露隐私的例子

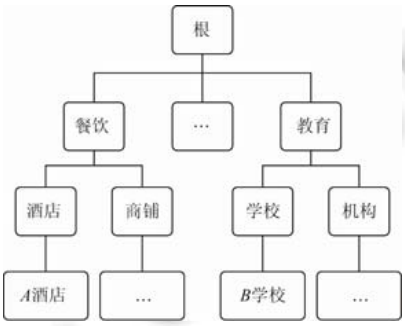


Fig.4 A part of the location semantic tree

图 4 位置语义树的一部分

位置语义筛选要确保位置间的语义是分散的.用 $Loc_i^{cs}=\{l_i^{s(1)},l_i^{s(2)},...\}$ 表示 Loc_i^{cr} 进一步经位置语义筛选后的位置集合,Loc_i^{cs} 满足式(2)描述的位置语义筛选规则.

$$\begin{cases}\left|Loc_i^{cs}\right|\geqslant K_i-1\\ \forall l_i^{s(k)}\in Loc_i^{cs}(k=1,2,...,\left|Loc_i^{cs}\right|):SD(l_i^{s(n)},l_i^{(K_i)})\geqslant L\end{cases}\tag{2}$$

其中,$SD(l_i^{s(n)},l_i^{(K_i)})$ 表示第 i 次请求的候选假位置 $l_i^{s(n)}$ 与真实位置 $l_i^{(K_i)}$ 之间的语义距离,L 为用户自定义的语义阈值,取决于构建的语义树的深度.Loc_i^{cs} 中的假位置个数必须达到 K_i–1 个,否则达不到用户给定的隐私保护需

求,需增加额外的候选位置加入 Loc_i^{cr}.

2.4 历史请求概率分布筛选

为了使产生的假位置更符合用户的历史请求分布,在位置语义筛选后我们选取更符合用户历史请求概率分布的假位置.用 n 表示用户历史请求记录的个数,用语义树中与根结点直接相连的结点作为历史请求的语义类别,假设共有 s 个语义类别,用 m_k 表示用户的历史请求归为第 k 个语义类别的频次,则用户的历史请求归为第 k 个语义类别的概率可表示为

$$P_k = \frac{m_k}{n} \tag{3}$$

用平均历史请求概率来描述用户的历史请求行为,表示为

$$\sigma = \sum_{k=1}^{s}(P_k m_k)/n \tag{4}$$

考虑到 $Loc_i^{cs}=\{l_i^{s(1)},l_i^{s(2)},...\}$ 为位置语义筛选后的位置集合,我们需要从 Loc_i^{cs} 中元素的 $t=C_{\left|Loc_i^{cs}\right|}^{K_i-1}$ 种组合情况中挑选出最接近用户历史请求行为的 K_i–1 个假位置,即 $Loc_i^{ch}=\{l_i^{h(1)},l_i^{h(2)},...,l_i^{h(K_i-1)}\}$,假设 $A_i^{(j)}$ 为从集合 Loc_i^{cs} 中挑选出 K_i–1 个假位置的第 j 个组合,用 $HD(A_i^{(j)})$ 表示集合 $A_i^{(j)}$ 中的假位置在 s 个语义类别上的平均分布概率与平均历史请求概率的差异,其计算方法为

$$HD(A_i^{(j)}) = \left|\frac{1}{K_i-1}\sum_{k=1}^{s}(P_k m'_{i,k}) - \sigma\right| \Big/ \sigma \tag{5}$$

其中,$m'_{i,k}$ 表示从 Loc_i^{cs} 中从挑选出的 K_i–1 个假位置归为第 k 个语义分类的频次.我们认为,HD 越小,产生的假位置集就越接近平均历史请求概率,由此 Loc_i^{ch} 需满足:

$$Loc_i^{ch} = \underset{A_i^{(j)}=A_i^{(1)},...,A_i^{(t)}}{\arg\min} HD(A_i^{(j)}) \tag{6}$$

经时空关联性,位置语义和历史请求概率分布筛选后,在第 i 次请求服务时上传的位置集合为 $Loc_i = Loc_i^{ch} \cup l_i^{(K_i)}$.

2.5 假位置生成算法

算法 1. 基于时空关联和位置语义的个性化假位置生成算法.

输入:

1. 用户的真实位置与请求时间 $l_i^{(K_i)}, T_i$
2. 前一时刻请求时上传的位置集合与请求时间 Loc_{i-1}, T_{i-1}
3. 位置语义树及用户历史请求概率 $P=(P_1,P_2,...,P_s)$
4. 隐私保护需求 K_i、语义距离阈值 SD_{th}、可达时间误差系数 α、候选假位置集合空间系数 N

输出:第 i 次请求服务时上传的位置集合 Loc_i

1. $Loc_i^{cr}=\varnothing, Loc_i^{cs}=\varnothing$;
2. **while** $\left|Loc_i^{cs}\right| < K_i - 1$ **do**
3. 生成包含 $\left|K_i - 1 - \left|Loc_i^{cs}\right|\right| \cdot N$ 个假位置的集合 L, $\forall l_i \in L$ 满足:
   $\exists l_{i-1}^{(j)} \in Loc_{i-1}(j=1,2,...,K_{i-1}): \left|T(l_{i-1}^{(j)}, l_i) - \Delta T\right| \leqslant \alpha \Delta T$;
4. $Loc_i^{cr} = Loc_i^{cr} \cup L$;
5. **for** each $l_i \in L$ **do**
6. **if** $SD(l_i, l_i^{(K_i)}) \geqslant SD_{th}$ **then**
7. $Loc_i^{cs} = Loc_i^{cs} \cup \{l_i\}$;
8. **end if**

9. **end for**
10. **end while**
11. **for** j=1:$C_{|Loc_i^{cs}|}^{K_i-1}$ **do**
12. 从集合 Loc_i^{cs} 中挑选出 K_i–1 个假位置,生成第 j 个组合 $A_i^{(j)}$;
13. **end for**
14. $Loc_i^{ch}=\underset{A_i^{(j)}=A_i^{(1)},\ldots,A_i^{(t)}}{\arg\min} HD(A_i^{(j)})$;
15. **output** $Loc_i=Loc_i^{ch}\bigcup l_i^{(K_i)}$;

2.6 算法复杂度分析

时空关联性筛选对应算法 1 的 2 行~4 行,这一阶段要求至少产生满足时间可达性的 $N(K_i-1)$个候选假位置供下一阶段基于位置语义进一步筛选,由于经位置语义筛选后生成的假位置可能不足 K_i–1 个,就需要返回第 1 阶段补充满足时间可达性的候选假位置,因此,在最差情况下将生成 $aN(K_i-1)$个候选位置,当 N 和 K_i–1 确定时,a 为确定的常数,而判断每个候选位置的时间可达性最多进行 K_{i-1} 次运算,因此,时空关联性筛选的时间复杂度为 $O(NK_iK_{i-1})$;位置语义筛选对应算法 1 的 5 行~9 行,对于每个候选位置都需要计算它与真实位置间的语义距离,由于候选位置最多为 $aN(K_i-1)$个,因此,位置语义筛选的时间复杂度为 $O(NK_i)$;历史请求概率分布筛选对应算法 1 的 11 行~14 行,从集合 Loc_i^{cs} 中挑选出 K_i–1 个假位置共有 $C_{|Loc_i^{cs}|}^{K_i-1}$ 种组合情况,而 Loc_i^{cs} 中最多有 $N(K_i-1)$个元素,因此,历史请求概分布筛选的时间复杂度为 $O(C_{N(K_i-1)}^{K_i-1})$.综上,算法 1 的时间复杂度为 $O(NK_iK_{i-1}+C_{N(K_i-1)}^{K_i-1})$.在一般情况下,$K$ 与 N 的取值均为个位数,算法 1 的运行时间可以接受.

3 算法性能评估

3.1 实验设置

采用 Python3.5 来实现文中算法,实验环境是 Win10 系统,AMD-A6 2.40GHz CPU,4GB 内存.实验数据集来自 Weeplaces 网站[14,15],该网站的功能主要是可视化用户的签到信息.该数据集包含 7 658 368 条签到记录,每条记录包含用户名、签到时间、签到位置名、经度、纬度、城市和位置语义类别.数据集一共包含了 15 799 个用户生成的 971 309 个位置,我们从中选取了纽约市区的 21 469 个位置点构建了一棵 5 层语义树,所有的位置在语义上被分成 9 大类:餐饮、教育、出行、商业、社会服务、娱乐、公共设施、住宅及其他.

我们将本文提出的算法与文献[8]提出的 SimpMaxMinDisDS 和文献[4]提出的 SpaCorDS(spatioemporal correlation-aware dummy-based privacy protection scheme,简写为 SpaCorDS)两种算法进行比较.其中,SimpMaxMinDisDS 侧重于解决假位置与真实位置间语义重合的问题,SpaCorDS 侧重于处理连续请求下如何合理产生假位置.实验参数设置见表 1.

Table 1 Settings of the experimental parameters

表 1 实验参数设置

参数名	取值范围	默认值
可达时间误差系数α	[0,1]	0.1
候选假位置集合空间系数 N	[1,∞)	2
语义阈值 L	[2,8]	6
隐私保护需求 K_{i-1}	[2,10]	5
隐私保护需求 K_i	[2,10]	–

3.2 性能比较

实验基于本文提出的算法,SimpMaxMinDisDS 算法和 SpaCorDS 算法生成假位置,针对每个给定的隐私保

护需求 K_i,每个算法都独立运行 100 次,产生 100(K_i–1)个假位置,然后分别基于时空关联性,位置语义的差异性和历史请求概率分布对产生的假位置进行检验,判断其实际满足用户隐私保护需求的程度.

首先,基于时空关联性对所有算法生成的假位置进行检验,统计假位置能够通过连续可达性筛选规则的比例,结果如图 5 所示.本文提出的算法和 SpaCorDS 算法由于在生成假位置时考虑了时空关联性,因此产生的假位置均通过了连续可达性筛选,而 SimpMaxMinDisDS 算法未考虑时空关联性,导致产生的假位置中通过筛选的平均比例只有 38.1%.在 K_i=3 时通过筛选的比率最高,但也有 45%的假位置被识别出来,在 K_i=6 时,通过筛选的比例最低,有 67.8%的假位置被识别.由此可见,在设计假位置生成方法时考虑时空关联性是极为重要的.

其次,基于位置语义相似性对所有算法产生的假位置进行检验,设定语义距离阈值为 6,统计假位置通过语义相似性检验的比例,结果如图 6 所示.本文提出的算法和 SimpMaxMinDisDS 算法由于在生成假位置时考虑了位置语义相似性,因此产生的假位置均通过了语义相似性检验,而 SpaCorDS 算法没有考虑考虑假位置与真实位置间的语义关系,产生的假位置中通过筛选的平均比例为 67.3%.在 K_i=8 时通过筛选的比例最高,为 70.8%,即有 29.2%的假位置与真实位置的语义相同或相近.当用户请求服务位置的语义被推测出来时也会泄露用户隐私,因此,在假位置生成方法中考虑位置语义的差异性是很有必要的.

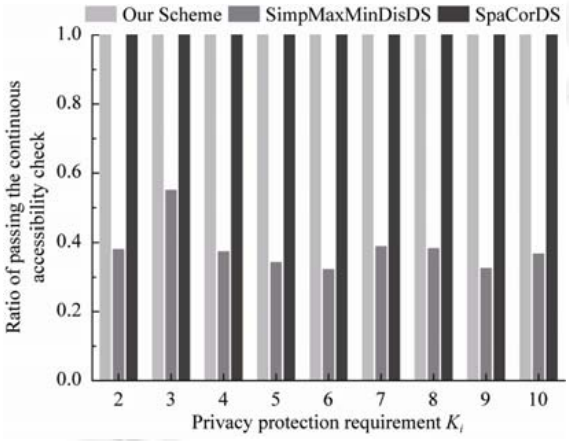


Fig.5 Ratio of passing the continuous accessibility check

图 5 通过连续可达性筛选的假位置比例

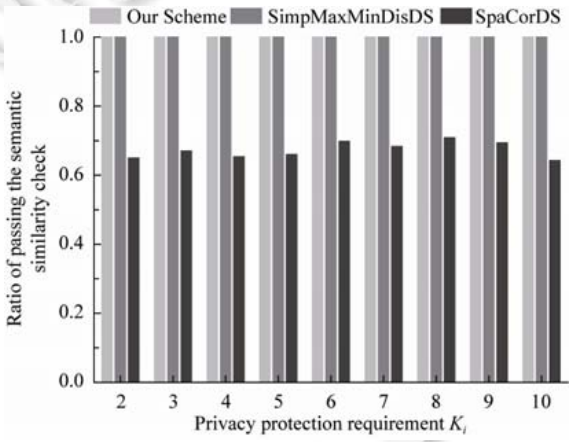


Fig.6 Ratio of passing the semantic similarity check

图 6 通过语义相似性筛选的假位置的比率

接着,基于历史请求概率分布对所有算法产生的假位置进行检验.针对给定的隐私保护程度需求 K_i,用集合 $Loc_{K_i}^{test}$ 表示包含 $100K_i$ 个位置的数据集, $m_k^{t(K_i)}$ 表示 $Loc_{K_i}^{test}$ 中归为第 k 个语义类别的位置个数,共有 s 个语义类别,则该数据集的平均分布概率为

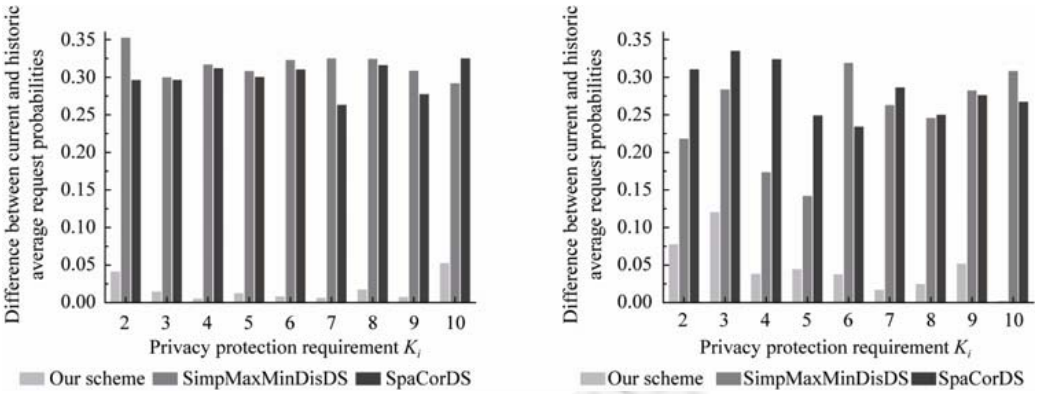
$$\sigma'_{K_i}=\sum_{k=1}^{s}(P_k m_k^{t(K_i)})/100\cdot K_i \tag{6}$$

用 $HD(Loc_{K_i}^{test})$ 表示在 K_i 下, σ'_{K_i} 与历史平均请求概率 σ 间的差异度,该值越小表示 σ'_{K_i} 越接近平均历史请求概率 σ,认为该方法产生的假位置更接近用户历史请求概率分布,计算方法为

$$HD(Loc_{K_i}^{test})=\left|\sigma'_{K_i}-\sigma\right|/\sigma \tag{7}$$

我们选取了用户 A 和用户 B 前 100 个历史记录,由公式(4)计算出其平均历史请求概率分别为 0.167 和 0.196 8.图 7 显示了两个用户在给定隐私保护程度 K_i 下,通过 3 种算法得到的上传位置数据的平均分布概率与历史平均请求概率间的差异度.可以看出,对于用户 A 和用户 B,在所有的 K_i 下,本文提出算法的平均差异度始终最小.上传数据的平均分布概率与历史请求概率间的差异度会受到位置语义的分布情况,用户行为差异等因素

的影响,但是,我们的算法在假位置生成中加入了历史请求概率分布筛选,能使得产生的假位置更符合用户的历史请求行为,进一步提高了个性化隐私保护程度.



(a) 用户 A 的结果 (b) 用户 B 的结果

Fig.7 Difference between current and historic average request probabilities

图 7 当前平均请求概率与历史平均请求概率间的差异度

最后,我们对算法的通信量和响应时间做进一步分析.由于本文所提算法的主要通信量在于从基站获取位置语义树,因此我们集中分析这一部分的通信量和响应时间.本实验采用的位置语义信息经编码后仅为 237KB,而 5G 网络的传输速率实际可达 1Gbps,相当于每秒可以传输约 125MB 的数据,且具有较高的稳定性.因此在 5G 网络下传输 237KB 的数据仅需 1.185ms,该量级的响应时间是适用于 LBS 场景的.

4 结 论

针对攻击者掌握一定的背景知识,例如道路时空可达信息、位置特征、用户的历史请求统计特性等,会导致假位置被识别的概率升高,降低隐私保护程度的问题,提出了基于时空关联和位置语义的个性化假位置生成算法,首先根据与前一次请求位置连续可达的条件产生假位置,然后通过建立语义树筛选出与真实位置语义相近的假位置,最后进一步筛选出与用户历史请求统计特性最接近的假位置.基于真实数据集将本文提出的算法与现有的算法进行了比较,表明本文提出的算法在攻击者掌握相关背景知识时,可有效降低假位置被识别的风险.

References:

[1] Huguenin K, Bilogrevic I, Machado JS, *et al*. A predictive model for user motivation and utility implications of privacy-protection mechanisms in location check-ins. IEEE Trans. on Mobile Computing, 2018,17(4):760−774. [doi: 10.1109/TMC.2017.2741958]

[2] Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location-based services. In: Proc. of the IEEE Int'l Conf. on Pervasive Services. Santorini: IEEE Computer Society, 2005. 88−97. [doi: 10.1109/PERSER.2005.1506394]

[3] Kido H, Yanagisawa Y, Satoh T. Protection of location privacy using dummies for location-based services. In: Proc. of the IEEE Int'l Conf. on Data Engineering Workshops. Tokyo: IEEE, 2005. 1248−1248. [doi: 10.1109/ICDE.2005.269]

[4] Liu H, Li X, Li H, *et al*. Spatiotemporal correlation-aware dummy-based privacy protection scheme for location-based services. In: Proc. of the IEEE Conf. on Computer Communications. Atlanta: IEEE, 2017. 1−9. [doi: 10.1109/INFOCOM.2017.8056978]

[5] Liu H, Li XH, Wang EM, Ma JF. Privacy enhancing method for dummy-based privacy protection with continuous location-based service queries. Journal on Communications, 2016,37(7):140−150 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x.2016142]

[6] Li WH, Ding S, Meng JJ, Li H. Spatio-temporal aware privacy-preserving scheme in LBS. Journal on Communications, 2018,39(5): 134−142 (in Chinese with English abstract). [doi: CNKI:SUN:TXXB.0.2018-05-013]

[7] Takbiri N, Houmansadr A, Goeckel DL, *et al*. Matching anonymized and obfuscated time series to users' profiles. IEEE Trans. on Information Theory, 2019,65(2):724−741. [doi: 10.1109/TIT.2018.2873134]

[8] Chen S, Shen H. Semantic-aware dummy selection for location privacy preservation. In: Proc. of IEEE Trustcom/BigDataSE/ISPA. Tianjin: IEEE, 2016. 752−759. [doi: 10.1109/TrustCom.2016.0135]

[9] Li Y, Cao X, Yuan Y, *et al*. PrivSem: Protecting location privacy using semantic and differential privacy. World Wide Web, 2019. 1−30. [doi: 10.1007/s11280-019-00682-0]

[10] Niu B, Li Q, Zhu X, *et al*. Achieving *k*-anonymity in privacy-aware location-based services. In: Proc. of the IEEE Conf. on Computer Communications. Toronto: IEEE, 2014. 754−762. [doi: 10.1109/INFOCOM.2014.6848002]

[11] Niu B, Zhang Z, Li X, *et al*. Privacy-area aware dummy generation algorithms for location-based services. In: Proc. of the IEEE Int'l Conf. on Communications. Sydney: IEEE, 2014. 957−962. [doi: 10.1109/ICC.2014.6883443]

[12] Niu B, Li Q, Zhu X, *et al*. Enhancing privacy through caching in location-based services. In: Proc. of the IEEE Conf. on Computer Communications. Kowloon: IEEE, 2015. 1017−1025. [doi: 10.1109/INFOCOM.2015.7218474]

[13] Hayashida S, Amagata D, Hara T, *et al*. Dummy generation based on user-movement estimation for location privacy protection. IEEE Access, 2018,6:22958−22969. [doi: 10.1109/ACCESS.2018.2829898]

[14] Liu Y, Wei W, Sun A, *et al*. Exploiting geographical neighborhood characteristics for location recommendation. In: Proc. of the ACM Int'l Conf. on Conf. on Information and Knowledge Management. Shanghai: ACM, 2014. 739−748. [doi: 10.1145/2661829.2662002]

[15] Liu X, Liu Y, Aberer K, *et al*. Personalized point-of-interest recommendation by mining users' preference transition. In: Proc. of the ACM Int'l Conf. on Conf. on Information and Knowledge Management. San Francisco: ACM, 2013. 733−738. [doi: 10.1145/2505515.2505639]

附中文参考文献:

[5] 刘海,李兴华,王二蒙,马建峰.连续服务请求下基于假位置的用户隐私增强方法.通信学报,2016,37(7):140−150. [doi: 10.11959/j.issn.1000-436x.2016142]

[6] 李维皓,丁晟,孟佳洁,李晖.基于位置服务中时空关联的隐私保护方案.通信学报,2018,39(5):134−142. [doi: CNKI:SUN:TXXB.0.2018-05-013]

周佳琪(1995－),男,浙江宁波人,硕士生,CCF 学生会员,主要研究领域为隐私保护.

李燕君(1982－),女,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为物联网.