满足地理不可区分性的偏好感知多对多任务分配算法

张朋飞[1],翟睿辰[1],程 祥[2,3],张治坤[4*],刘西蒙[5],王 杰[6]

(1. 安徽理工大学计算机科学与工程学院,安徽淮南 232001;2. 北京邮电大学计算机学院(国家示范性软件学院),北京 100876;
3. 网络与交换技术国家重点实验室(北京邮电大学),北京 100876;4. 浙江大学计算机科学与技术学院,浙江杭州 310058;
5. 福州大学计算机与大数据学院,福建福州 350108;6. 安徽理工大学安全科学与工程学院,安徽淮南 232001)

摘 要: 为空间众包中的工人分配任务是后续收集位置相关数据的重要前提. 为了应对可能的位置隐私泄露问题,研究者往往结合地理不可区分性进行保护. 现有满足地理不可区分性的任务分配方法通常针对一对一场景,其研究目标一般集中在最小化平均旅行距离,而不是最大化任务分配数量;同时,它们假设工人能分配去执行任意的任务. 此外,这些研究往往结合平面拉普拉斯机制实现地理不可区分性. 上述机制的随机性和无界性会导致工人上传的位置数据包含过量噪音,进而降低任务分配的效用,导致工人平均旅行距离较大或者任务无法完全分配. 为解决以上问题,本文提出满足地理不可区分性的偏好感知多对多任务分配算法 MONITOR(Many-to-many task allOcation under geo-iNdIsTinguishability for spatial crOwdsouRcing). 该算法主要思想是对工人的偏好任务进行分组加噪并上传工人真实位置到模糊偏好任务位置之间的距离以代替直接上传工人的模糊位置. 在 MONITOR 中,为了收集任务分配必需的工人到任务的距离信息,设计了基于分组的模糊距离收集方法 GroCol(Group-based obfuscated distance Collection);同时为了提高任务分配的效用,设计了参数无关的模糊距离比较方法 ParCom(Parameter-free obfuscated distance Comparison). 此外,本文进一步从理论上分析了 MONITOR 的隐私、效用和复杂度. 在 2 个真实数据集和 1 个模拟数据集上的实验结果表明 MONITOR 取得与非隐私任务分配类似的任务分配数量,且较基准方法的旅行距离降低了 20% 以上.

关键词: 空间众包;任务分配;隐私保护;地理不可区分性;平均旅行距离

基金项目: 安徽高校自然科学研究项目(No.2024AH050364)

中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0878-17

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240938

A Preference-aware Many-to-Many TAsk Allocation Algorithm Under Geo-Indistinguishability

ZHANG Peng-fei[1], ZHAI Rui-chen[1], CHENG Xiang[2,3], ZHANG Zhi-kun[4*], LIU Xi-meng[5], WANG Jie[6]

(1. *School of Computer Science and Engineering*, *Anhui University of Science and Technology*, *Huainan*, *Anhui* 232001, *China*;
2. *National Pilot Software Engineering School*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*;
3. *The State Key Laboratory of Networking and Switching Technology*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*;
4. *College of Computer Science and Technology*, *Zhejiang University*, *Hangzhou*, *Zhejiang* 310058, *China*;
5. *College of Computer and Data Science*, *Fuzhou University*, *Fuzhou*, *Fujian* 350108, *China*;
6. *School of Safety Science and Engineering*, *Anhui University of Science and Technology*, *Huainan*, *Anhui* 232001, *China*)

Abstract: In spatial crowdsourcing, task allocation is a crucial prerequisite for subsequent location-aware data collection. To tackle potential location privacy breaches, researchers often adopt geo-indistinguishability. Existing approaches that satisfy Geo-I are often designed for one-to-one scenarios, while implicitly assume that workers can perform any task, and they often focus on minimizing the average travel distance, rather than maximizing the number of task allocation. Furthermore, these studies often incorporate the planar laplacian mechanism to achieve Geo-I. However, due to the randomness and unbounded nature of PL, it can result in excessive noise in the location data uploaded by workers, significantly deteriorating

收稿日期:2024-10-17;修回日期:2025-01-13;责任编辑:覃怀银

*通讯作者:张治坤

the utility of task allocation. This can lead to either long distances or unassigned tasks. To address these problems, we propose MONITOR (Many-to-many task allOcation under geo-iNdIsTinguishability for spatial crOwdsouRcing), a new privacy-preserving task allocation approach for many-to-many scenario. The general idea of MONITOR is to upload the distances from each worker's true location to the obfuscated preferred tasks' locations instead of uploading each obfuscated worker's location. In MONITOR, to collect the distances for subsequent task allocation, we design an obfuscated distance collection method, called GroCol (Group-based obfuscated distance Collection). To improve the utility for task allocation, we develop a parameter independent obfuscated distance comparison method called ParCom (Parameter-free obfuscated distance Comparison). To illustrate the effectiveness of MONITOR, we first theoretically analyze its privacy guarantee, task utility, and computational complexity. We then empirically show on two real-world datasets and one synthetic dataset that MONITOR share similar results to that of non-private task allocation about the number of assigned tasks, and reduce the average travel distance by more than 20% compared to the baseline approaches.

Key words: spatial crowdsourcing; task allocation; privacy protection; geo-indistinguishability; average travel distance

Foundation Item(s): Natural Science Research Project of Anhui Educational Committee (No.2024AH050364)

1 引言

空间众包(Spatial Crowdsourcing,SC)通过汇聚大量普通群体完成传统上需要专业人员执行的空间任务,从而降低了总体开销,已成为现实世界中执行复杂任务的有效范式[1]. 在SC系统中,工人提交自身的位置信息到服务器进行任务分配;随后,工人根据任务要求前往指定地点,完成任务并将感知结果提交给服务器. 这种方式利用丰富的感知设备,能够以较低成本高效地完成大量数据的收集任务[2,3]. 通常,任务分配的效用通过平均旅行距离和成功完成任务的数量衡量[4]. 由于不同工人对同一任务的执行效果存在差异,因此任务分配应考虑将任务分配给多个工人,以便对不同的感知结果进行融合,从而提高最终的感知质量. 同时,囿于工人池的大小或SC发起者的激励机制等,每个工人倾向于承担更多的任务. 换言之,任务与工人之间不再是传统的一对一关系,而是多对多关系,即一个任务可能需要分配给多个工人,而一个工人也可以承担多个任务[5]. 此外,工人由于时间、空间的限制或个人兴趣等,通常只能接受有限的任务,因此,工人对任务存在地理位置上的偏好. 将工人不偏好的任务分配给他们,可能导致工人不愿意前往任务地点上传感知数据,进而降低SC的数据效用. 例如,某商家希望在不同区域推销商品,但不同人对不同区域的熟悉程度不同. 为了提高商品销量,显然可以将该问题建模为一个偏好感知的多对多任务分配问题[6]. 目前,偏好感知的多对多任务分配研究已广泛应用于智慧交通[7]、机器学习[8]和环境监测[9]等领域.

工人上传的数据中,显式或隐式地蕴含着个人敏感信息. 如果这些位置信息泄露,将导致工人对参与空间众包任务的意愿降低,导致候选工人不足等,从而影响任务分配的效用,甚至可能使任务分配无法顺利进行[10]. 为防止位置隐私泄露,地理不可区分性(Geo-Indistinguishability,Geo-I)[11]应运而生,其保护级别不依赖于攻击者的先验知识. 此外,Geo-I并不依赖于可信的服务器平台,也不假设系统中存在可信的中介方. 更重要的是,Geo-I可以通过平面拉普拉斯(Planar Laplacian,PL)机制简单且有效地实现. 在满足Geo-I的隐私保护方法中,工人的位置在发送给服务器之前会进行扰动. 由于精确位置从未离开工人的设备,这些方法能够有效防止工人受到来自服务器、其他工人或网络环境中黑客的攻击. 满足Geo-I要求的位置隐私保护方法已广泛应用于各种真实位置相关的情形,如LP-Guardian[12]、LP-Doctor[13]和"附近的人"的安全发现[14]等.

本文在满足Geo-I的条件下系统性地研究偏好感知的多对多任务分配问题. 平台根据收到的模糊位置信息给每个任务分配一些工人实现平均旅行距离最小化,同时保证任务的完全分配[15~18]. 尽管现有研究[19~26]已经对保护工人位置的任务分配问题展开了诸多有益探索,但它们假设工人能去执行任何任务,这在实践中可能并不现实. 此外,直接将现有满足Geo-I的方法扩展到多对多场景会导致任务无法完全分配或者带来较大的工人旅行距离. 其根本原因在于所设计的隐私保护机制无法直接套用到非隐私的多对多任务分配算法上,或者产生的模糊位置距离真实位置较远. 比如在文献[17]中,Zhang等人基于拉普拉斯机制设计了概率距离比较机制. 由于需要逐个对任务进行贪心的分配,在工人数量一定的前提下导致有些任务无法分配. 这是因为这些任务可能只受少数工人偏好,而由于贪心的对比,这些工人已经提前分配到任务. 在文献[22]中,Wang等人基于误差分析提出了基于模糊距离的任务分配方法,然而其隐私保护机制也仅适用于一对一场景.

在实际应用中,工人的偏好任务呈现一定的聚集

现象. 图1所示为文献[17]数据集中2个城市真实地图上任务的分布情况,不难发现任务存在聚集现象. 因此,如果一些位置在地理上彼此比较靠近,它们能用同一个位置代替. 比如,“滴滴出行”的司机和乘客,司机为了自身收益最大化,选择在大公司等客流量大的地方等待从而尽快载客,并在回程时路过类似地点避免出租车空载,而乘客在人流量多的地点也更容易打到车,这就不可避免地形成聚集现象,从而实现司机和乘客的双赢[3, 12, 13]. 基于此观察,本文提出一种基于分组加噪的偏好感知多对多任务分配方法,称为MONITOR(Many-to-many task allOcation under geo-iNdIsTinguishability for spatial crOwdsouRcing). 该方法的整体思路是每个工人上传其自身真实位置到偏好任务模糊位置的距离代替上传自身模糊位置. 通过使用一个位置代替一组地理上相互靠近的位置,该方法仅需对组内位置加一次噪音,从而有效减少噪音的注入次数. 此外,生成的模糊位置限制在该组内. 例如,偏好集内的多个位置若相同,那么只需要注入一次噪音即可,而不会产生额外的信息损失. 为了提高Geo-I保护下的多对多任务分配效用,本文通过注入噪音最小化,提出了一种最优分组方法. 尽管直接采用PL方法可以避免隐私预算分割,但它可能导致生成的模糊位置偏离真实位置,降低任务分配效用,而分组加噪机制则可以使模糊位置更接近真实位置,从而提高任务分配的效用.

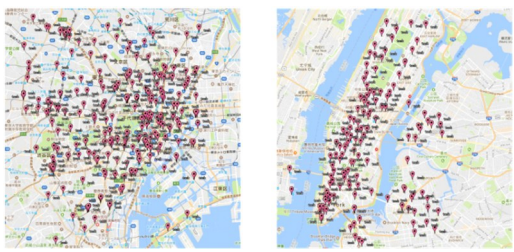


图1 任务聚集现象

根据MONITOR,本文设计了基于分组的模糊距离收集方法(Group-based obfuscated distance Collection method,GroCol),通过优化分组数和分组方式,平衡噪音和信息损失. 为了进一步提高任务分配的准确性,基于泰勒展开,本文设计了参数无关的模糊距离比较方法(Parameter-free obfuscated distance Comparison, ParCom),通过预估模糊距离的概率分布预测距离的真实值,从而有效解决了现有方法因模糊距离相同导致任务无法完全分配的问题.

综上,本文的主要贡献如下:

(1)针对保护工人位置的偏好感知多对多任务分配问题,本文提出对应的求解算法MONITOR,并进一步分析了MONITOR的隐私、效用和复杂度.

(2)为了收集任务分配所必需的工人到任务的距离信息,本文设计了GroCol方法,确定了最优分组数并设计了对应的分组方法. 本文进一步证明GroCol满足Geo-I,并给出了效用分析.

(3)本文基于模糊距离的概率分布进行距离比较,设计了ParCom方法,提高了任务分配效用,并避免了过度依赖参数调整.

2 相关工作

Hien等人[19]引入了一个基于Geo-I的私有框架,使工人参与而不损害他们的位置隐私;特别是提出了一个分析模型衡量任务完成的概率,在不确定工人位置的情况下找到合适的分区以确保高成功率的任务分配. 然而,上述方法和任务分配松耦合. 为进一步通过紧耦合提升效用,Wang等人[20]设计了混合整数非线性规划问题,既对工人的位置提供保护又使选定工人的预期平均旅行距离最小化. 为了对工人提供激励,Wang等人[21]在考虑工人不同保护需求的前提下,提供了一个个性化的概率优胜者选择机制,将每个任务分配给可能距离其最近的工人. 上述研究往往使平均旅行距离最小化,为了使工人对位置的期望覆盖率最大化,Wang等人[22]基于凸优化技术在保护工人位置的前提下形式化了任务的覆盖率最大化问题,并设计了求解方案. 为了应对在线任务分配场景,Hien等人[23]提出了一个三阶段框架量化工人和任务的可达性概率. Li等人[24]为了提升任务分配的效率,提出了隐私保护的批量任务分配方法. Tao等人[25]进一步缩短了文献[23]中方法的平均旅行距离,并基于竞争比提出了优化方案. Zhang等人[26]基于误差分析提出了分组的模糊位置生成方法和效用感知的模糊距离收集方法,并用于任务分配. 上述研究往往针对平面场景,为了应对更加实际的路网场景,Qiu等人[27]设计了新的隐私保护机制和隐私保护的任务分配算法. Tao等人[28]针对任务规划问题提出了满足Geo-I的新框架,量化分析了一个工人接受一个任务的概率.

针对空间众包中的非隐私任务分配问题,Lin等人[29]提出了一种基于聚合的双异构任务分配算法,研究了对偶异构对任务分配问题的影响,并寻求任务完成质量最大化和平均旅行距离最小化. Feng等人[30]提出了一种基于准实时求解框架的在线任务分配智能角色划分方法,在实时任务分配过程中有效解决了3种类型对象的匹配问题. 然而,上述研究方法没有考虑到任务分配过程中存在的隐私泄露问题. 为此,Gong等人[31]提出了一种新的双云服务器隐私保护方案,有效解决了不诚实的用户和云服务器勾结下的暴力攻击问题,可以保证有效任务分配和解决这种攻击,保护用户的隐私.

然而上述研究较少考虑工人的偏好,为此Miao等人[32]提出了一个联合偏好学习的任务分配框架,旨在从SC系统中通过考虑工人的偏好实现高效的任务分配,同时确保隐私数据的私密性. Lin等人[33]提出了一种基于马氏距离的距离评估方法,用于测量扰动位置中工人和任务之间的相关性;此外还设计了一种基于K-Means的分组算法,能够将共享任务的工人有效聚集在一起,通过任务交换提高了任务分配的成功率,同时保护了工人和任务的位置隐私. 此外,在多对多任务分配场景中,Rasoolabadi等人[34]提出了一种新的方法形式化处理群组任务分配问题,解决了群组多角色任务分配问题. 针对位置隐私保护分散计算中的任务分配问题,Hui等人[35]提出了一种基于不完全偏好表的任务分配模型,同时使用PL机制进行加噪保护节点的隐私信息.

表1对相关工作进行了总结对比. 由表1可知,尚无多对多场景下保护工人位置问题的探索,而且现有研究往往集中在平均旅行距离最小化,而不是分配任务数最大化,直接扩展现有研究到多对多场景可能导致任务无法完全分配或者带来较大的工人旅行距离. 具体原因如下:(1)由于PL机制加噪的随机性和无界性,加噪位置可能会显著偏离真实位置,且模糊位置中含有较多噪音,从而降低任务分配的效用,增大平均旅行距离;(2)无法分配的这些任务只在少数工人的偏好集内,而这些工人可能已经提前分配了其他任务.

表1　相关工作对比

相关工作	任务偏好	一对一	多对多	旅行距离最小化	分配率	在线
文献[19~21,27,28,31,33]		√		√		
文献[22]		√			√	√
文献[23,25]		√		√		√
文献[26,32,35]	√	√		√		
MONITOR	√		√	√	√	

3　预备知识

3.1　基于偏好感知的多对多任务分配

非隐私情境下的多对多任务分配问题有如下输入:N个工人$w_1,w_2,\cdots,w_N$和M个任务$t_1,t_2,\cdots,t_M$,每个工人共偏好ρ个任务,每个任务至少由g个工人承担,至多由h个工人承担. 根据文献[6],可以形式化为最大流最小割(Maximum Flow and Minimum Cut, MFMC)问题并采用匈牙利算法进行求解,具体建模方法如图2所示.

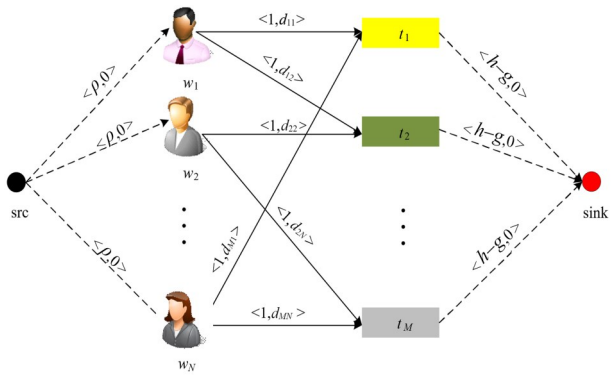


图2　MCMF建模方法

在介绍MCMF之前,首先需要指出MCMF边的属性为一个实数对,实数对中的前者表示该条路的"容量",后者表示经过该条路的"花费". 为了表示路径起始点,引入2个虚拟节点"src"和"sink". 在图2中,"src"到工人的"容量"表示流入某工人的流量最多为ρ,即每个工人至多做ρ个任务;"花费"表示经过该条路的花费为0;工人到任务之间的"容量"为1或者0,1表示该工人可以去做该任务,"花费"表示该工人去做该任务的旅行距离,为便于展示,图2对"容量"值为0者并未显示;任务到"sink"的"容量"表示从该任务流出至多$h-g$的单位,即该任务至多有$h-g$个人去做,"花费"为0,故只需求得从"src"到"sink"的最小费用最大流,便可保证所有任务完成的前提下工人的平均旅行距离最小化. 同时,为保证任务至少由g个人做,每个任务首先选择距离其最小的g个工人即可.

3.2　地理不可区分性

地理不可区分性形式化定义如下:

定义1　(ε-Geo-I)　对于一个随机算法$\mathcal{M}$,给定隐私预算ε和任意2个属于算法$\mathcal{M}$输入的位置点$x_1,x_2\in \mathrm{dom}(\mathcal{M})$以及任意属于算法$\mathcal{M}$输出的数据$z\in \mathrm{Range}(\mathcal{M})$,如果满足式(1)且$d(x_1,x_2)<r$,那么就称算法$\mathcal{M}$满足$\varepsilon$-Geo-I:

$$\Pr(\mathcal{M}(x_1)=z)\leqslant \mathrm{e}^{\varepsilon r}\times \Pr(\mathcal{M}(x_2)=z) \tag{1}$$

一种典型的实现Geo-I的方法是PL机制,其概率密度函数为$D(z)(x)=\frac{\varepsilon^2}{2\pi}\exp(-\varepsilon d(z,x))$. 为了便于表示,转换为式(2)的极坐标形式:

$$D(r,\theta)=\frac{\varepsilon^2}{2\pi}r\cdot\exp(-\varepsilon r) \tag{2}$$

其中,r和θ分别表示半径和角度. 分别对r和θ求边缘积分可得:

$$D_{\varepsilon,r}(r)=\int_0^{2\pi}D(r,\theta)\mathrm{d}\theta=\varepsilon^2 r\mathrm{e}^{-\varepsilon r}$$
$$D_{\varepsilon,\theta}(\theta)=\int_0^{\infty}D(r,\theta)\mathrm{d}r=\frac{1}{2\pi} \tag{3}$$

半径r的生成方式为

$$r=C_{\varepsilon}^{-1}(p)=-\frac{1}{\varepsilon}\left(W_{-1}\left(\frac{p-1}{\mathrm{e}}\right)+1\right) \tag{4}$$

其中，W_{-1}表示郎博万函数. 接着在$[0,2\pi)$随机生成角度，最后按照式(5)进行加噪：

$$z=x+(r\times\cos\theta,r\times\sin\theta) \tag{5}$$

针对Geo-I主要包含如下3种组合性质：

定理1 （串行机制） 假定$\mathcal{M}_1,\mathcal{M}_2,\cdots,\mathcal{M}_k$是工作在数据集$D$上的$k$个分别满足$\varepsilon_k$-Geo-I的随机算法，那么这些算法整体在$D$上满足$\sum\varepsilon_k$-Geo-I.

定理2 （并行机制） 假定$\mathcal{M}_1,\mathcal{M}_2,\cdots,\mathcal{M}_k$是分别工作在数据集$D_1,D_2,\cdots,D_k$上的$k$个分别满足$\varepsilon_k$-Geo-I的随机算法，那么这些算法整体在$D$上满足$\max(\varepsilon_k)$-Geo-I.

定理3 （后处理性质） 给定一个满足ε-Geo-I的随机算法$\mathcal{M}_1$，假定算法$\mathcal{M}_2$是基于$\mathcal{M}_1$输出的任意操作，那么复合算法$\mathcal{M}_1\circ\mathcal{M}_2$仍然满足$\varepsilon$-Geo-I.

3.3 问题陈述

图3为本文研究问题的示意图. 给定SC系统中的N个工人，表示为$W=\{w_1,w_2,\cdots,w_N\}$，共有M个任务需要完成，表示为$T=\{t_1,t_2,\cdots,t_M\}$，其中$N>M$. 一方面，第i个工人的偏好集表示为S_i，其中有ρ个任务；另一方面，考虑到完成任务的质量和发起者的经济消耗等，每个任务至少由g个工人承担，至多由h个工人承担(约束2)；同时，每个工人最多做ρ个任务(约束3). 在保护工人位置隐私前提下，保证任务尽量完成. 特别地，第i个工人选择去做第j个任务的前提是t_j在S_i中(约束1和约束4). 形式化定义如下：

$$\arg\min_{I}\sum_{w_i\in W}\sum_{t_j\in T}I(w_i,t_j)d(w_i,t_j)$$
$$\text{s.t. } I(w_i,t_j)\in\{0,1\} \quad \text{约束1}$$
$$g\leqslant\sum_{i=1}^{N}I(w_i,t_j)\leqslant h \quad \text{约束2}$$
$$0\leqslant\sum_{j=1}^{M}I(w_i,t_j)\leqslant\rho \quad \text{约束3}$$
$$S(i,j)-I(w_i,t_j)\in\{0,1\} \quad \text{约束4.}$$

综上，本文研究问题是识别出一系列元组$\langle w,t\rangle$，同时满足t在w的偏好集内. 服务器需要在平均旅行距离最小化的前提下保证所有任务完成的同时使工人位置满足Geo-I保护.

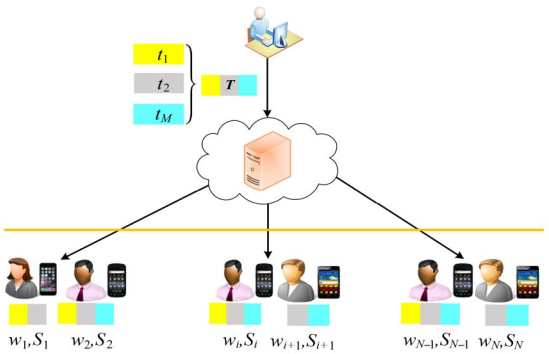


图3 本文研究问题示意图

表2给出了本文常用符号.

表2 本文常用符号

符号	符号定义
W	工人集合
T	任务集合
N	工人数
M	任务数
g	每项任务至少由多少人做
h	每项任务至多由多少人做
w_i	第i个工人
t_j	第j项任务
S_i	第i个工人的偏好集
ρ	偏好集长度
μ	一个组的代表性位置
μ'	一个组加噪后的位置

4 MONITOR算法

4.1 算法概述

由于工人的时间或者空间限制，每个工人偏好集内的任务呈现一定的聚集现象. 根据此聚集现象，考虑到进行任务分配所需要的是工人到任务的距离，提出分组加噪的思想.

图4左图表示对工人位置进行加噪，右图表示对工人偏好任务进行加噪. 如图4左图所示，对工人位置进行PL加噪得到的模糊位置距离偏好任务的距离可能发生显著变化，进而导致后续任务分配的平均旅行距离较大. 与之不同的是，如果按照图4右图的方式对偏好任务进行加噪，由于偏好任务的聚集现象，那么一组内的任务可以用一个任务表示. 同时，由于加噪的位置限制在组内，从而可以有一定的保序效果，进而能得到较好的效用.

为此，本文提出基于分组加噪的MONITOR算法，其核心思想是对工人的偏好任务进行加噪并上传工人真实位置到偏好任务的距离. 图5所示是MONITOR模型的

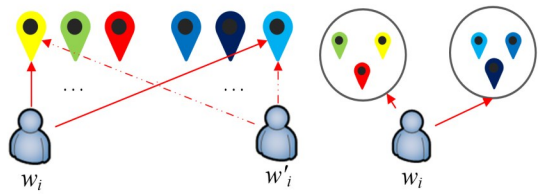


图4　工人位置加噪与分组加噪

概述图. 由图5可知,MONITOR模型包含如下5个阶段.

阶段1:每个工人在本地根据自身位置、偏好任务和隐私预算调用GroCol计算最优分组数.

阶段2:每个工人在得到自身的最优分组数后,调用GroCol对自身偏好任务进行分组.

阶段3:在得到分组结果后,每个工人计算并上传任务分配所需要的模糊距离等相关信息.

阶段4:服务器构建MCMF模型[6]并调用匈牙利算法和ParCom进行任务分配.

阶段5:服务器通知所选择的工人.

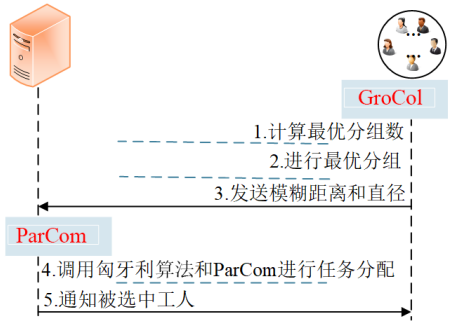


图5　MONITOR流程图

4.2　基于分组加噪的模糊距离收集方法

根据MONITOR,为了收集任务分配所必需的工人到任务的距离信息,基于分组加噪思想,一种直观的方法是对每个组内的代表性位置采用PL机制加噪. 然而,由于PL分布的特性,以及对分组加噪涉及注入噪音量和用一个位置代表组内位置导致的信息损失的权衡,该直观方案会面临以下2方面的挑战.

挑战1:由于PL分布的随机性和无界性,直接采用PL加噪会导致注入噪音量过大,生成的模糊位置距离真实位置较远.

挑战2:较大的分组数会带来较大的注入噪音,而较小的分组数会带来较大的信息损失. 为了权衡注入噪音和信息损失,涉及分组方式的设计以及分组数的确定.

为此,基于工人的偏好任务呈现聚集现象的观察,设计了基于分组的模糊距离收集方法GroCol,同时证明了GroCol满足地理不可区分性并给出了效用分析. 在GroCol中,通过优化分组数和分组方式平衡噪音和信息损失,并证明了GroCol满足Geo-I的要求. 此外,结合真实位置产生模糊位置,使生成的模糊位置仍局限在每个组内,能保证较高的隐私保护效用.

由于分组数和分组方式是基于加噪方式设计的,因此首先给出加噪方式并给出隐私和效用分析,然后给出最优分组数和分组方式,最后给出算法流程.

(1)加噪机制设计. 将每个组内的平均位置作为代表性位置,然后组内的其他位置按照到该位置的距离长短得到其被选为模糊位置的概率. 为此,得到式(6)的模糊位置生成概率:

$$p(z|x)=\frac{e^{[\mu-d(x,z)]}}{\sum_{v\in C}e^{[\mu-d(x,v)]}} \tag{6}$$

其中,C表示某个分组,μ表示组内的代表性位置,x表示真实位置,z表示选中的模糊位置(来自组内的某个位置).

接下来,给出加噪机制的隐私分析.

定理4　根据式(6)生成的模糊位置满足Geo-I.

证明　为了证明式(6)满足Geo-I,需要证明$\frac{p(z_1|x)}{p(z_2|x)}\leqslant e^{\varepsilon d(z_1,z_2)}$:

$$\frac{p(z_1|x)}{p(z_2|x)}=\left(\frac{e^{\varepsilon[\mu-d(x,z_1)]}}{\sum_{v\in C}e^{\varepsilon[\mu-d(x,v)]}}\right)\Bigg/\left(\frac{e^{\varepsilon[\mu-d(x,z_2)]}}{\sum_{v\in C}e^{\varepsilon[\mu-d(x,v)]}}\right)$$

$$=\frac{e^{\varepsilon[\mu-d(x,z_1)]}}{e^{\varepsilon[\mu-d(x,z_2)]}}=\frac{e^{\varepsilon\mu}e^{-\varepsilon d(x,z_1)}}{e^{\varepsilon\mu}e^{-\varepsilon d(x,z_1)}}$$

$$\leqslant e^{\varepsilon d(z_1,z_2)}.$$

证毕.

最后,给出加噪机制的效用分析. 首先给出如下引理.

引理1　给定$\eta\sim U(0,1)$,d_0为一个组内任务位置到代表性位置的最远距离,那么$e^{\varepsilon\eta}$的期望为$\frac{e^{\varepsilon d_0}-1}{\varepsilon d_0}$.

证明　假设有辅助变量$y=e^{\varepsilon\eta}$,那么有

$$P(Y\leqslant y)=P(e^{\varepsilon H}\leqslant y)=P\left(0\leqslant H\leqslant\frac{\ln y}{\varepsilon d_0}\right)$$

$$=\int_0^{\frac{\ln y}{\varepsilon d_0}}1\cdot dH=\frac{\ln y}{\varepsilon d_0}.$$

由于$P'(Y\leqslant y)=\frac{1}{\varepsilon d_0\eta}$,

那么$pdf(y)=\begin{cases}\frac{1}{\varepsilon d_0\eta}, & 1\leqslant\eta\leqslant e^{\varepsilon d_0}\\ 0, & \text{otherwise}\end{cases}$,

因此$E(y)=\frac{e^{\varepsilon d_0}-1}{\varepsilon d_0}$.

证毕.

在给出上述引理的基础上,给出如下定理.

定理 5 加噪机制的平均误差和最大误差均小于PL.

证明 假设用E表示加噪机制的平均误差,其表示每个可能的模糊位置到真实位置的加权距离,权重是根据真实位置汇报该模糊位置的概率. 假定一个组内有n个位置,那么有

$$E\left[d\left(\mu,t\right)\right]=\sum_{i=1}^{n}p\left(t_i|\mu\right)\cdot d\left(\mu,t_i\right)=\sum_{i=1}^{n}\frac{\mathrm{e}^{\varepsilon\left[\mu-d\left(\mu,t_i\right)\right]}}{\sum_{j=1}^{n}\mathrm{e}^{\varepsilon\left[\mu-d\left(\mu,t_j\right)\right]}}\cdot d\left(\mu,t_i\right).$$

证毕.

为了计算便利,使用常量d_0和一个辅助变量o_i表示$d\left(\mu,t_i\right)$,其中d_0表示μ和t_i之间的最大距离,o_i表示t_i对μ的贡献. 而且,由于o_i越大,$d\left(\mu,t_i\right)$越小,为此,设置$d\left(\mu,t_i\right)=\left(1-o_i\right)\cdot d_0$.

因此,有

$$E\left[d\left(\mu,t\right)\right]=d_0\sum_{i=1}^{n}\frac{\left(1-\eta_i\right)\mathrm{e}^{\varepsilon\left[\mu-\left(1-\eta_i\right)d_0\right]\left(1-\eta_i\right)}}{\sum_{j=1}^{n}\mathrm{e}^{\varepsilon\left[\mu-\left(1-\eta_j\right)d_0\right]d_0}}=d_0\frac{\sum_{i=1}^{n}\left(1-\eta_i\right)\mathrm{e}^{\varepsilon d_0\eta_i}}{\sum_{j=1}^{n}\mathrm{e}^{\varepsilon d_0\eta_j}}=d_0\left(1-\frac{\eta_1\mathrm{e}^{\varepsilon d_0\eta_1}+\eta_2\mathrm{e}^{\varepsilon d_0\eta_2}+\cdots+\eta_n\mathrm{e}^{\varepsilon d_0\eta_n}}{\mathrm{e}^{\varepsilon d_0\eta_1}+\mathrm{e}^{\varepsilon d_0\eta_2}+\cdots+\mathrm{e}^{\varepsilon d_0\eta_n}}\right).$$

进而有

$$E\left[d\left(\mu,t\right)\right]=d_0\frac{\sum_{i=1}^{n}\left(1-\eta_i\right)\mathrm{e}^{\varepsilon d_0\eta_i}}{\sum_{j=1}^{n}\mathrm{e}^{\varepsilon d_0\eta_j}}\geqslant\varepsilon d_0\sum_{i=1}^{n}\eta_i^2+\sum_{i=1}^{n}\eta_i\geqslant n\left(\frac{\varepsilon d_0}{3}+\frac{1}{2}\right).$$

根据泰勒展开,有$\mathrm{e}^x\geqslant x+1$. 因此,

$$\sum_{i=1}^{n}\eta_i\mathrm{e}^{\varepsilon d_0 o_i}\geqslant\sum_{i=1}^{n}\left(\varepsilon d_0\eta_i+1\right)\eta_i=\varepsilon d_0\sum_{i=1}^{n}\eta_i^2+\sum_{i=1}^{n}\eta_i.$$

根据大数定理,有$\sum_{i=1}^{n}x_i=n\cdot E\left(x\right)$和$E\left(x^2\right)=E^2\left(x\right)+D\left(x\right)$. 而且,由于没有先验知识,假设$o_i$是从均匀分布中采样得到,因此其均值和方差为$\frac{1}{2}$和$\frac{1}{12}$. 因此,有$\sum_{i=1}^{n}\eta_i\mathrm{e}^{\varepsilon d_0\eta_i}\geqslant n\left(\frac{\varepsilon d_0}{3}+\frac{1}{2}\right)$.

为了得到$\sum_{i=1}^{n}\mathrm{e}^{\varepsilon d_0\eta_i}$,根据引理1,可以得到:

$$E\left[d\left(\mu,t\right)\right]\leqslant d_0\left(1-\frac{n\left(\frac{\varepsilon}{3}d_0+\frac{1}{2}\right)}{n\cdot E\left(y\right)}\right)=d_0\left(1-\frac{\varepsilon d_o\left(\frac{\varepsilon}{3}d_0+\frac{1}{2}\right)}{\mathrm{e}^{\varepsilon d_0}-1}\right)\approx d_0\left(\frac{1}{2}-\frac{\varepsilon}{3}d_0\right).$$

由于PL的期望误差是

$$E_{\mathrm{PL}}=\int_0^{+\infty}\int_0^{2\pi}rD\left(r,\theta\right)\mathrm{d}r\mathrm{d}\theta=\int_0^{+\infty}\int_0^{2\pi}r\frac{\varepsilon^2}{2\pi}r\mathrm{e}^{-\varepsilon r}\mathrm{d}r\mathrm{d}\theta=\frac{2}{\varepsilon},$$

根据$E\left[d\left(\mu,t\right)\right]$的最大值为$\frac{3}{16\varepsilon}$,有$E\left[d\left(\mu,t\right)\right]\ll E_{\mathrm{PL}}$.

至于最大误差,PL的最大误差是正无穷,而所设计加噪机制的最大误差是d_0.

(2)最优分组数确定. 为了求得最优分组数,使所设计方案的误差小于直观方案的误差,从而得到分组数的约束条件,首先计算得到所设计方案的平均旅行距离.

如图6所示,假设一个工人做了f项任务,那么其旅行的真实距离为

$$d_1=d\left(w,t_1\right)+d\left(w,t_2\right)+\cdots+d\left(w,t_f\right)\tag{7}$$

其旅行的模糊距离为

$$d_2=d\left(w,r_1\right)+d(w,r_2),\cdots+d\left(w,r_f\right)\tag{8}$$

二者的差值为

$$\begin{aligned}\Delta&=d_2-d_1\\&=d\left(w,r_1\right)-d\left(w,t_1\right)+\cdots+d\left(w,r_f\right)-d\left(w,t_f\right)\\&\leqslant d\left(t_1,r_1\right)+\cdots+d\left(t_f,r_f\right)\end{aligned}\tag{9}$$

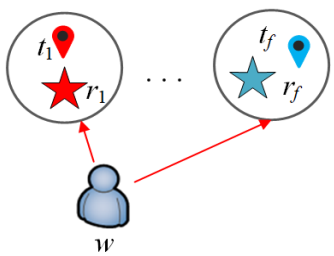


图6 加噪机制平均旅行距离分析

根据定理5,得到:

$$|C|<\frac{32}{3}\tag{10}$$

再者,使设计方案的噪音小于对位置加噪的噪音,

从而求得分组数的第2个约束. 已知定理5,同时根据加噪机制,距离向量中至少有一个距离是准确的,那么,所设计方案的总误差为

$$E_1=\left(\rho-|C|\right)d_0\left(\frac{1}{2}-\frac{\varepsilon_1 d_0}{3}\right) \tag{11}$$

其中,$\varepsilon_1=\dfrac{\varepsilon}{|C|}$,$|C|$为一个工人的分组数,$\max\left(E_1\right)=\dfrac{\rho}{2}+\dfrac{\varepsilon d_0}{3}-2\sqrt{\dfrac{\rho}{2}\cdot\dfrac{\varepsilon d_0}{3}}$. 同时,对工人位置加噪的总误差为$E_2=\dfrac{2}{\varepsilon_2}|C|$.

由$\max\left(E_1\right)<E_2$,可得到:

$$|C|\geqslant\frac{\varepsilon\left(\dfrac{\rho}{2}+\dfrac{\varepsilon d_0}{3}-2\sqrt{\dfrac{\rho}{2}\cdot\dfrac{\varepsilon d_0}{3}}\right)}{2} \tag{12}$$

为此,根据式(11)和式(12),可以设置分组数为

$$|C|=\frac{64+\varepsilon\left(\dfrac{\rho}{2}+\dfrac{\varepsilon d_0}{3}-2\sqrt{\dfrac{\rho}{2}\cdot\dfrac{\varepsilon d_0}{3}}\right)}{12} \tag{13}$$

(3)分组方法. 由于所设计方案的距离误差为$\Delta=d_0\left(\dfrac{1}{2}-\dfrac{\varepsilon d_0}{3|C|}\right)f$,直观方案的误差为$\Psi=\dfrac{2}{\varepsilon}f$,使得$\Delta<\Psi$,可以得到:

$$3|C|\left(\frac{d_0}{2}-\frac{2}{\varepsilon}\right)<\varepsilon d_0^2 \tag{14}$$

为了使式(14)恒成立,需要$\dfrac{d_0}{2}-\dfrac{2}{\varepsilon}<0$,那么可得:

$$d_0\leqslant\frac{4}{\varepsilon} \tag{15}$$

根据式(15)得到每个组需要满足的约束条件. 为此,可以在每个组内首先生成一个点,然后以此点为圆心,以式(15)的约束为直径,只要到圆心的距离小于式(15)对应值的一半,那么可以归为一个组.

$$\begin{aligned}
&E_1=\left(\rho-|C|\right)d_0\left(\frac{1}{2}-\frac{\varepsilon_1 d_0}{3}\right)\\
&\Rightarrow\frac{\left(\rho-|C|\right)d_0}{2}-\frac{2|C|}{\varepsilon_1}<\frac{\varepsilon_1 d_0^2\left(\rho-|C|\right)}{3}\\
&\Leftrightarrow\frac{\left(\rho-|C|\right)d_0}{2}-\frac{2|C|}{\varepsilon_1}<0\\
&\Leftrightarrow d_0<\frac{4|C|^2}{\varepsilon\left(\rho-|C|\right)}\\
&\Leftrightarrow d_0<\min\left\{\frac{4|C|^2}{\varepsilon\left(\rho-|C|\right)},\frac{4}{\varepsilon}\right\}.
\end{aligned}$$

在此基础上给出算法的整体流程.

(4)GroCol算法流程如算法1所示.

算法1　GroCol算法

输入:工人数N, 偏好集S, 工人真实位置x

输出:每个工人真实位置到偏好集内模糊位置的距离

1. 根据式(13)计算分组数
2. REPEAT
3. FOR i=1 TO N:
4. 根据$|C|$初始化位置点
5. 根据式(17)更新每一个簇
6. END FOR
7. FOR j=1 TO $\left|C_j\right|$:
8. 根据式(6)得到每个点被选中概率
9. 根据生成(0,1)随机变量ζ选择位置
10. END FOR
11. 每个工人计算自身位置到模糊位置的距离并上传

在算法1中给出了GroCol的算法流程. 特别地,首先计算出最优分组数;然后在初始化每个组初始点的基础上,根据式(17)更新每个簇;接着,根据式(6)得到每个组内位置点被选中的概率,进而依据随机变量的大小确定选中的位置;最后计算工人真实位置到模糊位置的距离并上传给服务器.

4.3　参数无关的模糊距离比较方法

在得到工人到任务的模糊距离并构建MCMF模型后,服务器求解调用匈牙利算法以得到任务分配的结果. 特别地,在匈牙利算法中需要求解2点之间的最短路径. 然而,工人到任务的这些模糊最短路径中可能存在较多相同值,根据这些相同距离随意地配对工人和任务会导致平均旅行距离增大或者任务无法完全分配. 为了提升MCMF的效用,需要区分这些相同的距离.

一种可能的做法是结合文献[17]设计的概率比较机制,假设$\tilde{d}_1$和$\tilde{d}_2$表示2个冲突距离,d_1和d_2分别表示2个冲突的距离,不失一般性,有$\tilde{d}_1=d_1+g_1$和$\tilde{d}_2=d_2+g_2$,而$g_1,g_2\sim N\left(\mu,\sigma^2\right)$,那么可以得到:

$$\begin{aligned}
P\left(d_1\leqslant d_2\right)&=P\left(\tilde{d}_2-\tilde{d}_1\leqslant g_1-g_2\right)\\
&=\iint_D f\left(g_1,g_2\right)\mathrm{d}g_1\mathrm{d}g_2
\end{aligned} \tag{16}$$

其中,

$$F=\left\{\left(g_1,g_2\right)\middle|\,g_1-g_2\geqslant\tilde{d}_2-\tilde{d}_1\right\},$$

$$f\left(g_i\right)=\frac{1}{\sqrt{2\pi}\,\sigma_i}\mathrm{e}^{-\frac{\left(g_i-\mu_i\right)^2}{2\sigma_i^2}},$$

$$f\left(g_1,g_2\right)=f\left(g_1\right)f\left(g_2\right).$$

如果$P\left(d_1\leqslant d_2\right)\geqslant\theta$,那么就认为$d_1\leqslant d_2$. 然而比较概率$\theta$的设置是一个难点,过大或过小的$\theta$均可能导致计算得到的结果不准确. 同时,$\theta$并非越大越好,这是由概率比较的不确定性导致的,过大的θ会导致所

有模糊距离之间的比较都无法通过,而过小的θ导致所有模糊距离之间的比较都能通过. 所以,在实践中可能难以预先确定一个恰当的比较概率. 再者,因为θ的设置需要接触原始数据,为了得到最优的θ,不得不多次消耗隐私预算,进而带来大量噪音. 因此,基于泰勒展开,本文设计了参数无关的模糊距离比较方法ParCom. 该方法通过预估模糊距离的概率分布预测距离的真实值,从而有效解决了现有方法因模糊距离相同导致任务无法完全分配的问题. 根据ParCom,只需要得到每个距离的均值和方差,在不需要任何超参数的情况下即可进行距离比较. 该方法的整体思路是根据模糊距离的概率分布预估2个相同距离发生微小改变情况下的变化量,根据变化量预测真实距离的大小. 如果发生微小改变后$\tilde{d}_1+\Delta<\tilde{d}_2+\Delta$,那么可以认为$d_1<d_2$.

具体地讲,假设μ_1、σ_1、μ_2和σ_2表示2个冲突距离对应的均值和偏差,根据泰勒展开,有:

$$p\left(d_i\leqslant X\leqslant d_i+\Delta d\right)\approx f\left(d_i\right)\cdot\Delta d \tag{17}$$

其中,$f\left(d_i\right)=\dfrac{1}{\sqrt{2\pi}\,\sigma_i}\mathrm{e}^{-\frac{\left(d_i-\mu_i\right)^2}{2\sigma_i^2}}$. 如果$p_1>p_2$,即$f\left(d_1\right)>f\left(d_2\right)$,那么:

$$\begin{aligned}&\frac{1}{\sqrt{2\pi}\,\sigma_1}\mathrm{e}^{-\frac{\left(d_1-\mu_1\right)^2}{2\sigma_1^2}}>\frac{1}{\sqrt{2\pi}\,\sigma_2}\mathrm{e}^{-\frac{\left(d_2-\mu_2\right)^2}{2\sigma_2^2}}\\&\Leftrightarrow\frac{1}{\sqrt{2\pi}\,\sigma_1}\mathrm{e}^{-\frac{\left(d_1-\mu_1\right)^2}{2\sigma_1^2}}\Big/\frac{1}{\sqrt{2\pi}\,\sigma_2}\mathrm{e}^{-\frac{\left(d_2-\mu_2\right)^2}{2\sigma_2^2}}>1\\&\Leftrightarrow\ln\left[\frac{1}{\sqrt{2\pi}\,\sigma_1}\mathrm{e}^{-\frac{\left(d_1-\mu_1\right)^2}{2\sigma_1^2}}\Big/\frac{1}{\sqrt{2\pi}\,\sigma_2}\mathrm{e}^{-\frac{\left(d_2-\mu_2\right)^2}{2\sigma_2^2}}\right]>0\\&\Leftrightarrow\frac{\sigma_2}{\sigma_1}+\mathrm{e}^{-\frac{\left(d_1-\mu_1\right)^2}{2\sigma_1^2}}\mathrm{e}^{-\frac{\left(d_2-\mu_2\right)^2}{2\sigma_2^2}}>0\end{aligned}\tag{18}$$

即如果式(18)成立,那么$d_1>d_2$.

在明确比较方法后,需要确定相同距离对应的均值和标准差. 根据文献[11],如果一个机制满足Geo-I,那么模糊位置到真实位置的距离服从$G\left(2,1/\varepsilon\right)$. 其中$G$表示伽马分布,那么模糊距离的方差为$\dfrac{2}{\varepsilon^2}$. 至于均值,根据定理5,可表示为$d_0\left(\dfrac{1}{2}-\dfrac{\varepsilon}{3}d_0\right)$,其中$d_0$表示该任务所在组的2项任务的最远距离.

5 MONITOR分析

5.1 隐私分析

定理6 MONITOR算法满足ε-Geo-I.

证明 根据MONITOR,只有阶段3计算偏好任务的模糊位置时服务器需要接触工人的真实位置. 根据定理4,假设分组数为$|C|$,对每个组生成模糊位置的过程满足$\dfrac{\varepsilon}{|C|}$-Geo-I;根据定理1中差分隐私的串行组合性质,将所有隐私预算相加,每个工人的所有模糊偏好位置满足ε-Geo-I;根据定理2中差分隐私的并行组合性质,每个工人的位置信息均满足ε-Geo-I;在MONITOR的后续操作中,服务器只需要接触带噪信息,根据定理3中差分隐私的后处理性质,MONITOR整体满足ε-Geo-I.

5.2 效用分析

定理7 MONITOR带来的平均旅行距离和最大旅行距离均小于PL机制.

证明 根据式(11),一个工人选择去做一项任务的总误差上界为$d_0\left(\dfrac{1}{2}-\dfrac{\varepsilon_1 d_0}{3}\right)$,其中$\varepsilon_1=\dfrac{\varepsilon}{|C|}$,那么在多对多场景下,MONITOR的总误差上界为$\Delta=d_0\left(\dfrac{1}{2}-\dfrac{\varepsilon d_0}{3|C|}\right)f$,其中$f$为一个工人选中去做的任务数. 特别地,MONITOR的总误差可能小于Δ,因为上界的计算方式是假设工人分配的任务全为模糊位置,然而如果直接选中的任务为模糊位置,则Δ变为$d_0\left(\dfrac{1}{2}-\dfrac{\varepsilon d_0}{3|C|}\right)(f-1)$. 所以采用MONITOR算法的平均旅行距离的误差为

$$0<\Delta\leqslant d_0\left(\frac{1}{2}-\frac{\varepsilon d_0}{3|C|}\right)f \tag{19}$$

而由于实现Geo-I的PL机制的无界性,PL的最大误差是正无穷. 因此,类似定理5的证明过程,MONITOR算法平均旅行距离和最大旅行距离均小于PL.

5.3 复杂度分析

(1)通信复杂度. 第1次通信发生在工人上传模糊距离和每个组内的最大距离d_0. 由于每个工人需要上传ρ个距离和$|C|$个d_0,那么共传送$N\cdot O\left(\rho+|C|\right)$. 第2次通信发生在服务器通知选中的工人执行任务,那么需要$N\cdot O\left(\rho\right)$.

综上,共需交换$N\cdot O\left(2\rho+|C|\right)$. 该复杂度仅仅和工人数量$N$线性相关,实践中可以接受.

(2)计算复杂度. 从算法总体流程可以看出,运算主要时间消耗在于分组、计算并上传模糊距离以及调用匈牙利算法进行任务分配.

对于分组方法,在确定分组数量后,对所有偏好任务进行一次遍历,那么需要$O\left(\rho\cdot|C|\right)$,进而总共需要$N\cdot O\left(\rho\cdot|C|\right)$. 对于计算并上传模糊距离,需要$2\cdot O\left(\rho\right)$.

至于任务分配,由于非隐私的匈牙利算法需要

$O(nm^2)$,其中n和m分别为点和边的数量. 但实际相对于任务分配,模糊距离比较所需要的时间可忽略,那么总体需要$O\left((N+M+2)\left(N+M+N\rho\right)^2\right)$. 该复杂度和非隐私任务分配的复杂度相当,实践中可以接受.

6 实验评估

6.1 数据集

本文利用从Foursquare[17]收集的2个公开的数据集和1个生成的模拟数据集进行实验.

(1)NYC:该数据集包含来自纽约的227 428条签到记录,共有142个地铁站和492个办公室地址. 随机选择50个地铁站和300个办公室地址分别作为任务位置和工人位置进行任务分配.

(2)TKY:该数据集收集于东京的地铁和办公室情况,包含325个地铁站和503个办公室地址,其中共有573 708条签到记录. 随机选择100个地铁站作为任务位置,选择500个办公室作为工人位置.

(3)SYN:为了验证算法在不同场景的鲁棒性,本文合成了该模拟数据集. 不失一般性[20,21],本文改变工人任务的位置分布、数量和所处区域的半径等,默认在1 km × 1 km的区域内随机生成100个工人和50个任务位置.

6.2 评价指标和实验环境

本文采用广泛使用的平均旅行距离(Average Travel Distance, ATD)和未分配任务数(Not Assigned Tasks, NAT)判断最终得到的噪音任务分配结果的效用:

$$D_{\mathrm{AT}}=\left(\sum_{(w,t)}d(w,t)\right)\Big/Q \tag{20}$$

$$T_{\mathrm{NA}}=M-Q \tag{21}$$

其中,D_{AT}表示成功分配的工人任务旅行距离除以总的成功分配任务数,T_{NA}表示未分配的任务数,Q表示总的成功分配任务数.

此外,为了验证所提模块算法GroCol和ParCom的有效性,本文采用均方根误差(Root Mean Square Error, RMSE)E_{RMS}来衡量真实位置距离直观方法生成的模糊位置和GroCol生成的模糊位置的差距.

$$E_{\mathrm{RMS}}=\sqrt{\frac{\sum_{i=1}^{D}\left(d_i-\tilde{d}_i\right)^2}{D}} \tag{22}$$

其中,D表示模糊位置的总数,d_i和$\tilde{d}_i$分别对应模糊距离和真实距离.

再者,为了验证所提模块算法ParCom的有效性,本文采用准确率A(Accuracy)模糊距离比较的正确次数.

$$A=\frac{E}{F} \tag{23}$$

其中,E和F分别表示在距离比较中,比较正确的个数和总的比较次数.

最后,为验证数据集中的聚集现象,本文采用广泛使用的F-Measure和Entropy进行分组效果评估. F-Measure可以由式(24)表示:

$$F(c,i)=\frac{2\cdot P(c,i)R(c,i)}{P(c,i)+R(c,i)} \tag{24}$$

其中,$P(c,i)$和$R(c,i)$分别表示分组结果类c与标准类i的准确率和召回率,可以由式(25)和式(26)得到:

$$P(c,i)=\frac{N_{ci}}{N_c} \tag{25}$$

$$R(c,i)=\frac{N_{ci}}{N_i} \tag{26}$$

其中,N_{ci}表示结果类c与标准类i交集的对象数量,N_c表示结果类c的对象数量,N_i表示标准类i的对象数量. 标准类来自位置聚类算法的结果. 整个分组算法的F-Measure是每个标准类i的加权平均数,可以表示成式(27):

$$F=\frac{\sum_i N_i\times F(c,i)}{\sum_i N_i} \tag{27}$$

Entropy可以表示为

$$e_c=-\sum_{i=1}^{C}P(c,i)\log P(c,i) \tag{28}$$

整个分组算法的Entropy是所有标准类加权Entropy的和,如式(29)所示:

$$e=\sum_c^C\frac{N_c}{N}e_c \tag{29}$$

本文的所有算法均在一台配置为Inter i7-7700HQ、8 GB内存的4核8线程笔记本电脑上进行,采用Python 3.7编写所有算法.

6.3 对比算法

根据对相关工作的分析,发现现有方案无法直接用来解决本文的问题. 为了验证所提方案的有效性,将MONITOR算法与以下设计的方案进行对比.

(1)NoPriv:在知道工人真实位置的前提下直接求解[6].

(2)IntWL:采用文献[11]中的离散PL对工人位置进行模糊.

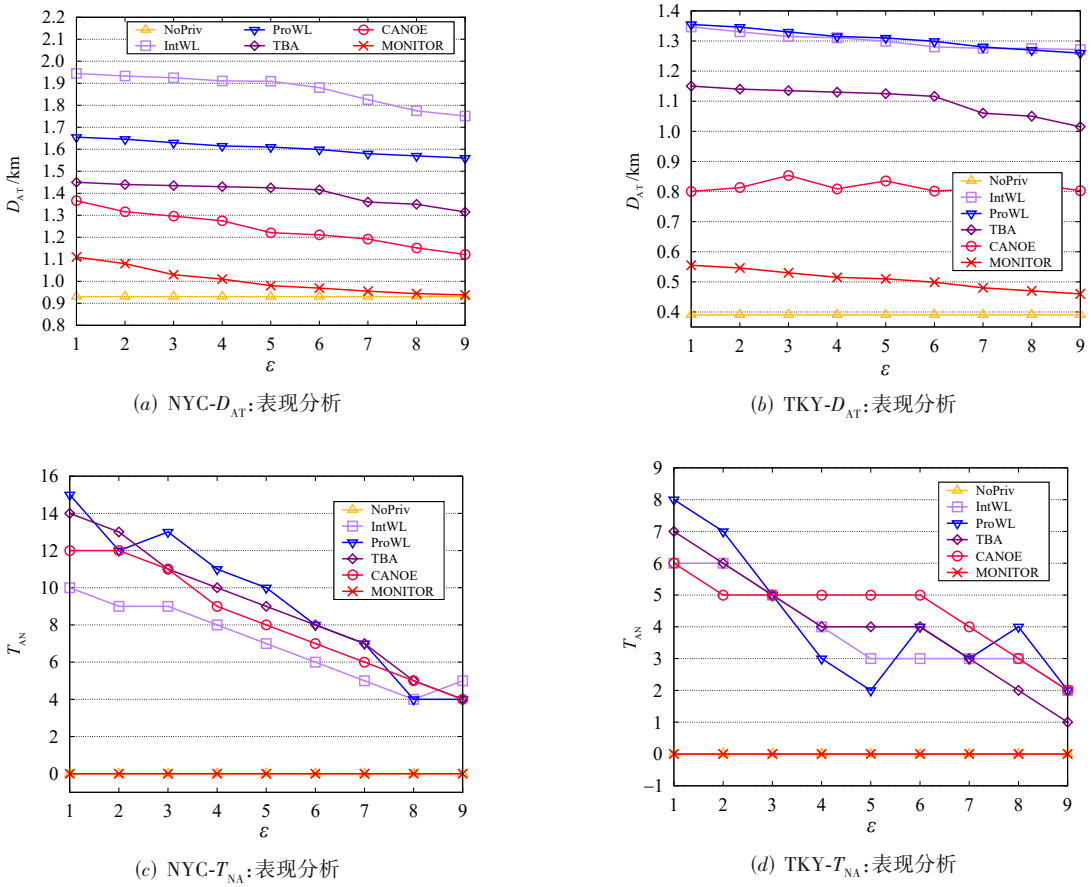
(3)ProWL:将文献[17]中的一对一概率任务分配机制扩展到本文的多对多场景. 特别地,服务器调用文献[17]中的算法计算距离任务最近的工人. 如果一个模糊距离大于另一个模糊距离的概率大于0.5,那么就认为前者较大. 其中0.5为默认的比较概率,在接下来的实验中,需要改变该值,以验证所设计的参数无关方法ParCom的有效性.

(4)TBA:在该方法中,服务器调用文献[21]中的算法求得任务分配结果. 特别地,对每项任务选择距离最近的g个工人.

(5)CANOE:在该方法中,服务器调用文献[22]中的算法求得任务分配结果. 特别地,对每项任务选择距离最近的g个工人.

6.4 实验结果与实验分析

(1)ε的影响. 图7所示是ε的变化对实验结果的影响,图7(a)和图7(c)对应于NYC数据集,默认设置$\rho=50$. 由图7可知,随着ε的增加,所有算法的效果整体向好,即D_{AT}变小同时T_{NA}也变小,这是因为随着ε的增加,所有算法包含的总体噪音减小. 其次,MONITOR算法效果最好,这是因为通过联合同组内任务的位置限制生成的模糊位置范围,带来了较高的概率,使模糊位置靠近真实位置,总体噪音较小,所以D_{AT}较小. 同时,采用MCMF进行任务分配,只要在非隐私偏好任务情况下即可保证任务完全分配,MONITOR算法也能保证任务完全分配. 然而,对于对比算法,由于往往从根本上采用PL进行加噪,因此生成的模糊位置距离真实位置较远,导致D_{AT}较大. 再者,对于对比算法,任务优先分配给距离其最近的工人,这些工人可能并不偏好该任务,因此T_{NA}较大. 特别地,对于ProWL,其满足的是弱化的Geo-I,即便这样的隐私放松情况,且有最好的超参数设定,MONITOR的效果仍然远远优于它. 这是因为ProWL是针对一对一场景所设计的,其自身的隐私机制无法较好地适用于多对多场景. 此外,TBA基于先验位置范围生成模糊位置,不同的先验范围可能带来不同的结果,且可能无法提前得到位置范围. 进一步地,ProWL和TBA皆针对平均旅行距离最小化场景所设计,而本文不仅需要最小化平均旅行距离,也需要保证任务完全分配.



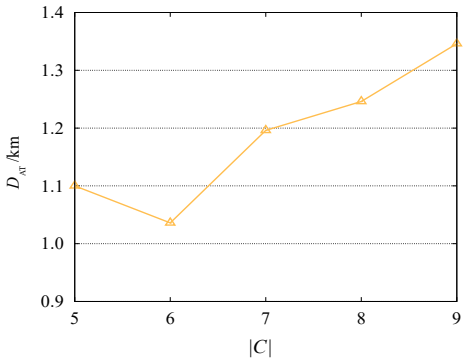
(a) NYC-D_{AT}:表现分析 (b) TKY-D_{AT}:表现分析

(c) NYC-T_{NA}:表现分析 (d) TKY-T_{NA}:表现分析

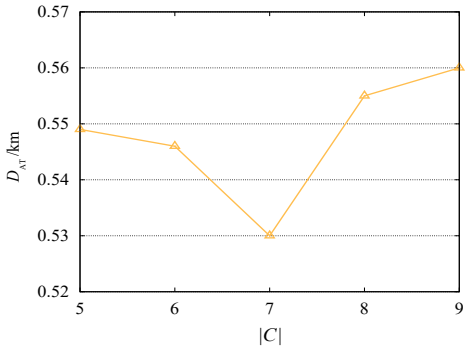
图7 ε的影响

(2)分组数$|C|$的影响与聚集现象评估. 图8展示了由式(16)得到的分组数对实验结果的影响. 可以观察到,随着分组数的增加,D_{AT}的值先降后增,在NYC数据集中近似最优分组数为6,在TKY数据集中最优分组数为7. 这与根据公式计算得到的结果近似一致,证明了设计分组数求解的正确性. 随着分组数的变化,D_{AT}变化显著,也从侧面说明方案容易训练.

表3为本文聚集现象的评估结果. 由表3数据可知,F-Measure分组效果的精度在75%以上,Entropy分组效果的精度在37%以上. 即便在隐私预算极小情况



(a) NYC:分组数



(b) TKY:分组数

图8　分组数的影响

下($\varepsilon=0.1$),分组效果的精度也分别在75%和37%附近.因此,实际任务的确呈现分组现象,也从侧面证明本文的分组效果较好.

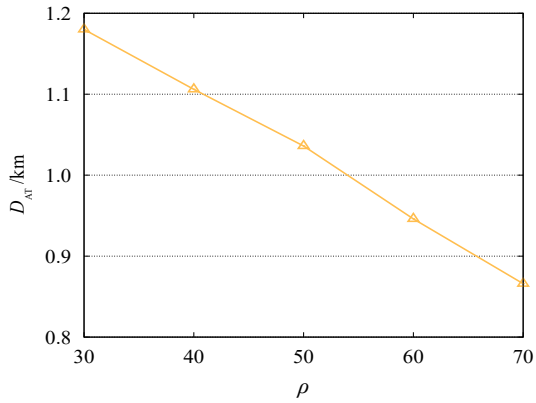
(3)偏好任务数ρ的影响.图9展示了偏好任务数量对实验结果的影响.可以观察到,随着ρ的增加,所有方法的效果趋于优化,这是因为在固定任务数量时,更大的ρ带来更大的概率,使每个任务分配给更好的工人.

表3　聚集现象评估

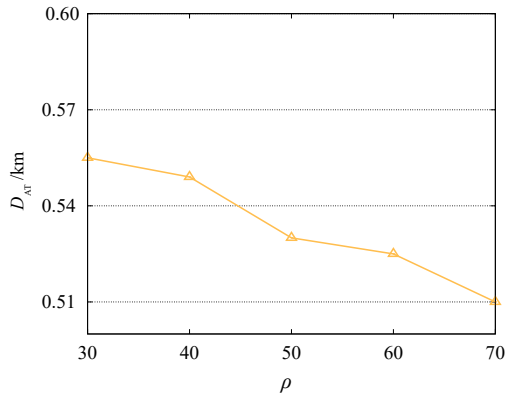
ε	NYC		TKY	
	F-Measure	Entropy	F-Measure	Entropy
0.1	0.807	0.207	0.758	0.376
1	0.848	0.168	0.784	0.321
3	0.868	0.147	0.807	0.297
5	0.895	0.126	0.824	0.267

(4)GroCol影响.在该部分实验中,采用D_{AT}和E_{RMS}评价GroCol的有效性.

就E_{RMS}而言,将其与PL相比,实验结果如图10(c)和图10(d)所示,可以看出,由于采用分组加噪的思想,通过MONITOR计算得到的模糊位置和真实位置之间的距离要小于直接基于PL计算得到的距离,侧面验证了GroCol的有效性.



(a) NYC-D_{AT}:偏好任务数量



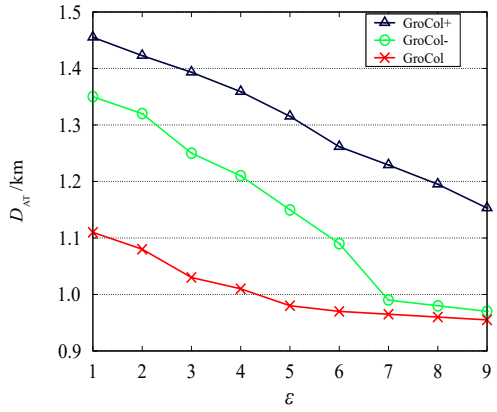
(b) TKY-D_{AT}:偏好任务数量

图9　偏好任务数的影响

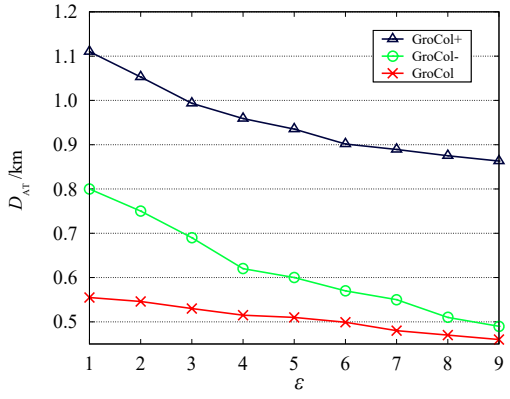
就D_{AT}而言,GroCol主要包含加噪机制(式(6))和分组机制(式(17)),为此设计GroCol+和GroCol-验证GroCol的有效性.其中,GroCol+在分组基础上采用PL对每个组加噪,GroCol-不分组只对工人位置采用式(6)加噪.由图11(a)和图11(b)可知GroCol-效果优于GroCol+,且GroCol效果最好.这是因为对工人位置采用式(6)加噪的总体噪音小于对偏好任务分组采用PL加噪的噪音.另外,GroCol同时进行分组和式(6)加噪,使总体噪音最小.

(5)ParCom影响.在该部分实验中,采用D_{AT}和A来评价ParCom的有效性.

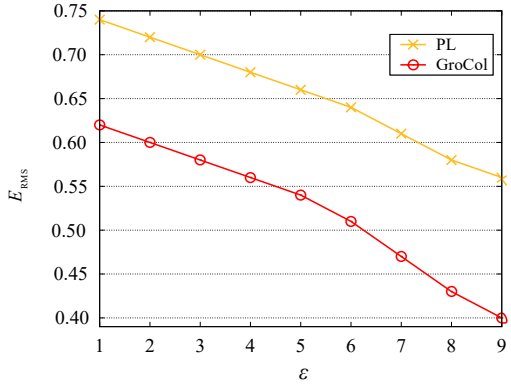
就D_{AT}而言,ParCom包括比较模式(式(20))和参数确定机制(均值和方差).为验证效果,ParCom+只使用



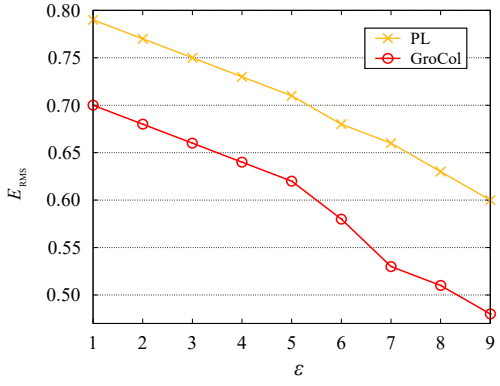
(*a*) NYC-D_{AT}:GroCol有效性



(*b*) TKY-D_{AT}:GroCol有效性

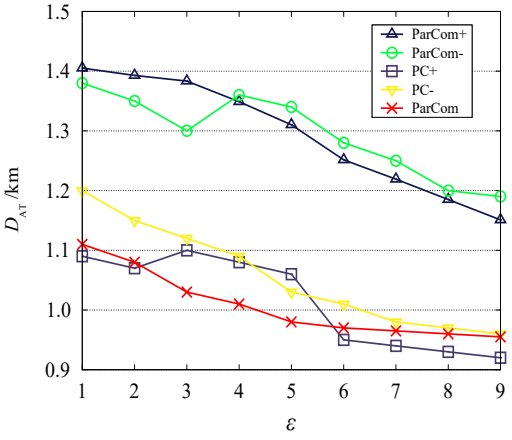


(*c*) NYC-RMSE:GroCol有效性

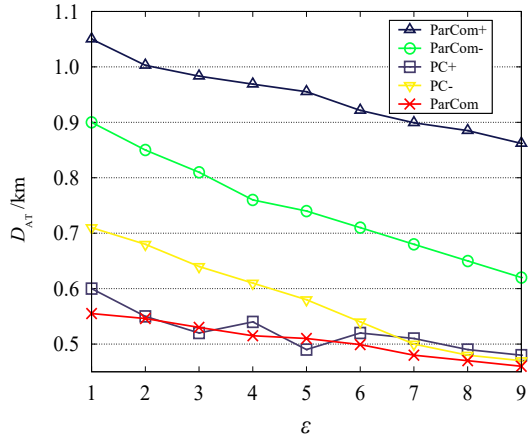


(*d*) TKY-RMSE:GroCol有效性

图10 GroCol的有效性



(*a*) NYC:ParCom有效性



(*b*) TKY:ParCom有效性

图11 ParCom的有效性

设计的方差,ParCom-只使用设计的均值;PC+和PC-对应于本文6.3节带超参数θ的直观方法,特别地,分别设置θ为0.5和0.6.

实验结果如图11所示,结论如下.(*a*)ParCom-的效果好于ParCom+. 这是因为在2个模糊距离比较过程中,均值起到决定作用,而根据定理5所求得的均值较为准确.(*b*)PC+和PC-好于ParCom+和ParCom-. 这是因为参数的确定至关重要,而ParCom+和ParCom-只使用

了部分“正确”的参数,因此效果提升并不显著.(*c*)PC+和PC-的效果相当.这是因为在直观方法中,超参数的设置是关键,而直观方法无法自动确定超参数.(*d*)ParCom效果最好,且偶尔和PC+相当.这是因为ParCom通过良好的比较模式和参数设置,使总体噪音最小;而对于PC+,由于无法自动确定最优参数,使结果并不稳定.

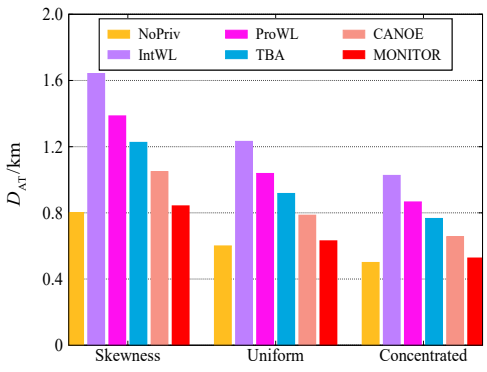
就A而言,需衡量比较正确模糊距离的对数占整个模糊距离对数的比例,实验结果如表4所示.表4中,ProWL+和ProWL-的比较概率分别为0.4和0.6,ProWL的比较概率默认为0.5.设置此对比实验以验证所涉及的参数无关方案的鲁棒性.

随着ε变大,所有算法的效果变好,即A变大.ProWL等对比算法对比较概率较为敏感.实验发现,并非比较概率越大越好,这是由概率比较的不确定性导致的:过大的比较概率会导致所有模糊距离之间的比较都无法通过,而过小的比较概率导致所有模糊距离之间的比较都能通过.所以,在实践中可能难以预先确定一个恰当的比较概率,故此效果不稳定.本文提出的MONITOR算法效果一直最好,唯一的异常是在超大ε的时候ParCom的效果和ProWL(ProWL+)相当.由于ParCom不需要超参数,其适用范围更加广阔.

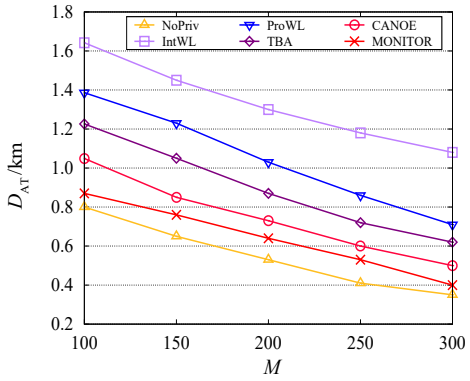
表4　ParCom算法和ProWL的A对比

数据集	ε	对比算法			
		ParCom	ProWL-	ProWL	ProWL+
NYC	0.1	**0.724 1**	0.721 6	0.725 6	0.722 3
	1	**0.784 5**	0.762 2	0.771 4	0.771 1
	3	**0.805 6**	0.795 8	0.795 3	0.794 1
	5	**0.841 2**	0.836 5	0.838 9	0.831 2
TKY	0.1	**0.715 1**	0.713 2	0.715 4	0.711 8
	1	**0.786 6**	0.781 1	0.785 5	0.781 1
	3	**0.814 7**	0.807 9	0.814 4	0.804 4
	5	**0.861 5**	0.831 4	0.851 4	0.850 1

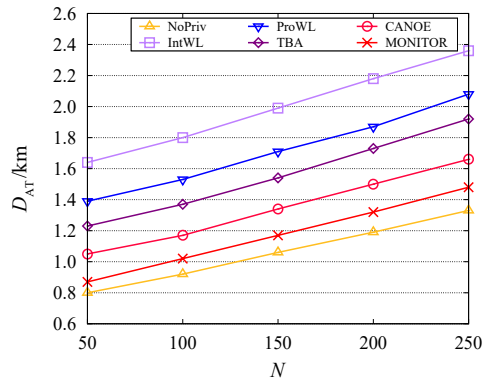
(6)MONITOR的鲁棒性.图12展示了在模拟数据集SYN上的实验结果,以验证MONITOR的鲁棒性.图12(*a*)展示了位置分布的影响,其中Skewness、Uniform和Concentrated分别表示工人和任务的位置在感知范围内呈现偏态分布、均匀分布和集中式分布.由图12(*a*)可知:随着位置分布的集中化程度越来越高,所有算法的D_{AT}逐渐变小,这是因为工人任务对之间的距离本身变



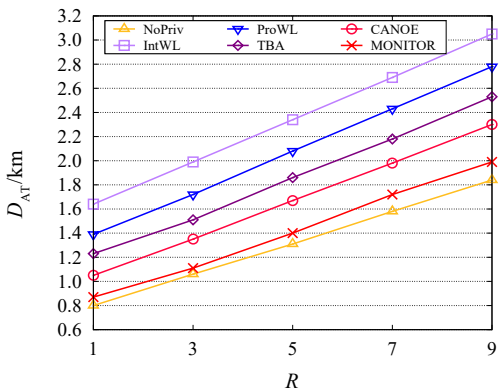
(*a*) SYN:位置分布的影响



(*b*) SYN:工人数量的影响



(*c*) SYN:任务数量的影响



(*d*) SYN:感知范围半径的影响

图12　MONITOR的鲁棒性

小. 图12(*b*)和图12(*c*)分别展示了工人数量M和任务数量N的影响,可知随着M和N的增加,D_{AT}分别逐渐变小和变大,这是因为更多的工人提供了任务更多的可能性、更多的任务不可避免分配到更远的工人. 图12(*d*)展示了感知范围半径R的影响,可知随着R的增加,D_{AT}逐渐变大. 这是因为在位置分布不变的情况下,更大的范围避免导致工人任务对之间的距离增加. 然而,在上述所有场景中,MONITOR的效果仍然最好,这是因为所设计的模糊距离收集方法GroCol在生成模糊位置过程中充分融合了临近的位置信息,使生成的模糊位置仍然以较大概率距离真实位置较近;同时在生成过程中,本文量化了满足Geo-I的任务分配中的总噪音并使噪音总量最小化,使生成的模糊位置对噪音注入更具鲁棒性. 此外,设计了模糊距离下的冲突消解方法ParCom,使任务分配的效用近似于非隐私下的效果.

7 总结和展望

针对隐私保护的多对多任务分配问题,本文提出了一种满足地理不可区分性的偏好感知多对多任务分配算法MONITOR,并给出了算法的隐私、效用和复杂度保证. 根据该算法,为降低任务分配过程中的注入噪音,提出一种基于分组加噪的模糊距离收集方法和一种参数无关的模糊距离比较方法,在2个公开数据集上的实验结果验证了方案的有效性. 在未来的研究中,拟结合激励机制研究工人偏好任务具敏感性的情况.

参考文献

[1] 王健, 刘嘉欣, 赵国生, 等. 移动群智感知中基于协同排序的任务推荐方法[J]. 电子学报, 2021, 49(10): 2012-2019.
WANG J, LIU J X, ZHAO G S, et al. Task recommendation method based on collaborative ranking in mobile crowd sensing[J]. Acta Electronica Sinica, 2021, 49(10): 2012-2019. (in Chinese)

[2] 蒋伟进, 张婉清, 陈萍萍, 等. 基于IWOA群智感知中数量敏感的任务分配方法[J]. 电子学报, 2022, 50(10): 2489-2502.
JIANG W J, ZHANG W Q, CHEN P P, et al. Quantity sensitive task allocation method based on IWOA in group intelligence perception[J]. Acta Electronica Sinica, 2022, 50(10): 2489-2502. (in Chinese)

[3] 宋天舒, 童咏昕, 王立斌, 等. 空间众包环境下的3类对象在线任务分配[J]. 软件学报, 2017, 28(3): 611-630.
SONG T S, TONG Y X, WANG L B, et al. Online task assignment for three types of objects under spatial crowdsourcing environment[J]. Journal of Software, 2017, 28(3): 611-630. (in Chinese)

[4] 范泽军, 沈立炜, 彭鑫, 等. 基于约束的空间众包多阶段任务分配[J]. 计算机学报, 2019, 42(12): 2722-2741.
FAN Z J, SHEN L W, PENG X, et al. Multi stage task allocation on constrained spatial crowdsourcing[J]. Chinese Journal of Computers, 2019, 42(12): 2722-2741. (in Chinese)

[5] LI Y C, ZHAO Y, ZHENG K. Preference-aware group task assignment in spatial crowdsourcing: A mutual information-based approach[C]//2021 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE, 2021: 350-359.

[6] LIU Y, GUO B, WANG Y, et al. TaskMe: Multi-task allocation in mobile crowd sensing[C]//Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. New York: ACM, 2016: 403-414.

[7] YU Y T, XUE X P, MA J X, et al. Efficient privacy-preserving task allocation with secret sharing for vehicular crowdsensing[J]. IEEE Internet of Things Journal, 2024, 11(6): 9473-9486.

[8] FAN Y, LIU L, ZHANG X X, et al. MAPP: An efficient multi-location task allocation framework with personalized location privacy-protecting in spatial crowdsourcing[J]. Information Sciences, 2023, 619: 654-678.

[9] WANG X D, PENG M Y, LIN H, et al. A privacy-enhanced multiarea task allocation strategy for healthcare 4.0[J]. IEEE Transactions on Industrial Informatics, 2023, 19(3): 2740-2748.

[10] DUGUÉPÉROUX J, ALLARD T. From task tuning to task assignment in privacy-preserving crowdsourcing platforms[M]//Transactions on Large-Scale Data- and Knowledge-Centered Systems: XLIV. Berlin: Springer, 2020: 67-107.

[11] ANDRÉS M E, BORDENABE N E, CHATZIKOKOLAKIS K, et al. Geo-indistinguishability: Differential privacy for location-based systems[C]//Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security - CCS 2013. New York: ACM, 2013: 901-914.

[12] WANG Z H, GUO C Q, LIU J H, et al. Accurate and privacy-preserving task allocation for edge computing assisted mobile crowdsensing[J]. IEEE Transactions on Computational Social Systems, 2022, 9(1): 120-133.

[13] WANG S P, LI J, WU G J, et al. Joint optimization of task offloading and resource allocation based on differen-

tial privacy in vehicular edge computing[J]. IEEE Transactions on Computational Social Systems, 2022, 9(1): 109-119.

[14] YANG M C, ZHU J H, XI H R, et al. Privacy-aware task allocation based on deep reinforcement learning for mobile crowdsensing[M]//Wireless Algorithms, Systems, and Applications. Cham: Springer Nature Switzerland, 2022: 191-201.

[15] JIANG Y L, ZHANG K, QIAN Y, et al. Preserving location privacy and accurate task allocation in edge-assisted mobile crowdsensing[C]//2022 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway: IEEE, 2022: 704-709.

[16] CHEN Z P, XU M M, SU C X. Online quality-based privacy-preserving task allocation in mobile crowdsensing[J]. Computer Networks, 2024, 251: 110613.

[17] ZHANG Q, WANG T C, TAO Y, et al. Location privacy protection method based on differential privacy in crowdsensing task allocation[J]. Ad Hoc Networks, 2024, 158: 103464.

[18] GUO B, LIU Y, WANG L Y, et al. Task allocation in spatial crowdsourcing: Current state and future directions[J]. IEEE Internet of Things Journal, 2018, 5(3): 1749-1764.

[19] HIEN T, GHINITA G, FAN L Y, et al. Differentially private location protection for worker datasets in spatial crowdsourcing[J]. IEEE Transactions on Mobile Computing, 2017, 16(4): 934-949.

[20] WANG L Y, YANG D Q, HAN X, et al. Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation[C]//Proceedings of the 26th International Conference on World Wide Web. Geneva: International World Wide Web Conferences Steering Committee, 2017: 627-636.

[21] WANG Z B, HU J H, LV R Z, et al. Personalized privacy-preserving task allocation for mobile crowdsensing[J]. IEEE Transactions on Mobile Computing, 2019, 18(6): 1330-1341.

[22] WANG L Y, QIN G H, YANG D Q, et al. Geographic differential privacy for mobile crowd coverage maximization[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2018: 200-207.

[23] HIEN T, SHAHABI C, XIONG L. Privacy-preserving online task assignment in spatial crowdsourcing with untrusted server[C]//2018 IEEE 34th International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2018: 833-844.

[24] LI M C, WANG J C, ZHENG L B, et al. Privacy-preserving batch-based task assignment in spatial crowdsourcing with untrusted server[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM, 2021: 947-956.

[25] TAO Q, TONG Y X, ZHOU Z M, et al. Differentially private online task assignment in spatial crowdsourcing: A tree-based approach[C]//2020 IEEE 36th International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2020: 517-528.

[26] ZHANG P F, CHENG X, SU S, et al. Task allocation under geo-indistinguishability via group-based noise addition[J]. IEEE Transactions on Big Data, 2023, 9(3): 860-877.

[27] QIU C X, SQUICCIARINI A, PANG C, et al. Location privacy protection in vehicle-based spatial crowdsourcing via geo-indistinguishability[J]. IEEE Transactions on Mobile Computing, 2022, 21(7): 2436-2450.

[28] TAO Q, TONG Y X, LI S Y, et al. A differentially private task planning framework for spatial crowdsourcing[C]//2021 22nd IEEE International Conference on Mobile Data Management (MDM). Piscataway: IEEE, 2021: 9-18.

[29] LIN X C, WEI K M, LI Z T, et al. Aggregation-based dual heterogeneous task allocation in spatial crowdsourcing[J]. Frontiers of Computer Science, 2023, 18(6): 186605.

[30] FENG Z H, XIAO R B. Three-dimensional task allocation for smart transportation in spatial crowdsourcing: An intelligent role division approach[J]. Advanced Engineering Informatics, 2024, 62: 102736.

[31] GONG Z M, LI J Y, LIN Y P, et al. A novel dual cloud server privacy-preserving scheme in spatial crowdsourcing[J]. Computers & Security, 2024, 138: 103659.

[32] MIAO H, ZHONG X L, LIU J X, et al. Task assignment with efficient federated preference learning in spatial crowdsourcing[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(4): 1800-1814.

[33] LIN Y M, JIANG Y J, LI Y, et al. Privacy-preserving batch-based task assignment over spatial crowdsourcing platforms[J]. Computer Networks, 2024, 241: 110196.

[34] RASOOLABADI M N, ZHU H B, WANG C. Solving the many to many grouped task allocation problem via E-CARGO[C]//2023 International Conference on Networking,

Sensing and Control (ICNSC). Piscataway: IEEE, 2023: 1-6.
[35] HUI H W, LIN F H, MENG L, et al. Many-to-many matching based task allocation for dispersed computing[J]. Computing, 2023, 105(7): 1497-1522.

作者简介

张朋飞 男,1992年1月出生于河南省鄢陵县. 现为安徽理工大学计算机科学与工程学院讲师、研究生导师. 主要研究方向为数据隐私保护与可信人工智能.
E-mail: zpf.bupt@bupt.cn

翟睿辰 男,2001年3月出生于安徽省马鞍山市. 现为安徽理工大学计算机科学与工程学院硕士研究生. 主要研究方向为数据隐私保护.
E-mail: ruichenzhai@aust.edu.cn

程 祥 男,1984年10月出生于北京市. 现为北京邮电大学计算机学院教授、博士生导师. 主要研究方向为数据隐私保护与可信人工智能.
E-mail: chengxiang@bupt.edu.cn

张治坤 男,1993年8月出生于浙江省杭州市. 现为浙江大学计算机科学与技术学院副教授、博士生导师. 主要研究方向为隐私计算、数据隐私保护、机器学习隐私与安全.
E-mail: zhikun@zju.edu.cn

刘西蒙 男,1988年10月出生于福建省福州市. 现为福州大学计算机与大数据学院教授、博士生导师. 主要研究方向为数据隐私保护与数据安全.
E-mail: nbnix@qq.com

王 杰 男,1964年1月出生于重庆市. 现为安徽理工大学安全科学与工程学院教授、博士生导师. 主要研究方向为数据隐私保护、智能检测与智能仪表、粉尘防治技术. 中国电子学会会员编号:E190103366M.
E-mail: 2024095@aust.edu.cn