doi:10.19920/j.cnki.jmsc.2023.05.005

隐私规避的网络调查与间接估计方法①

吕　欣[1], 刘楚楚[1,2], 蔡梦思[1], 陈洒然[3]
(1. 国防科技大学系统工程学院, 长沙 410073; 2. 新加坡国立大学计算学院, 新加坡 117417;
3. 盲信号处理国家重点实验室, 成都 610041)

摘要: 在管理决策中,管理对象的真实状态往往因隐私、敏感等因素导致自我报告数据质量不高,样本数据存在较大偏差,进而难以掌握目标对象的真实情况. 针对这一问题,同时为了满足数字经济时代下的数据隐私保护需求,本文开发了一类基于社交网络间接报告的数据采集方法,并在网络抽样与统计推断理论的基础上,设计了基于间接报告样本数据的总体估计方法(ECM). 该方法操作简单,可对调查对象进行随机采样或实施普查,除采集样本的自我陈述数据外,同时采集每个样本关于其密切社交对象的报告数据,从而避免了自身因敏感原因等不愿提供数据或提供不真实数据的问题,提出的估计方法能在样本报告数据的基础上实现对总体的高精度估计,并能实现自报告数据和他报告数据的交互验证. 本文的研究方法在一个多达 556 627 名活跃用户的难接触人群在线社交网络上进行了充分验证,抽样实验表明 ECM 对全网平均好友数和总体特征的估计误差低于 3%. 进一步地,本文开展了实证研究,通过设计自报告和他报告问卷,对某企业职员的一般和隐私性问题进行了问卷调查,并通过间接估计方法实现了对目标的总体估计,展示了该方法的实用性和有效性.

关键词: 隐私; 数据真实性; 网络抽样; 间接估计; 统计推断
中图分类号: C811　　**文献标识码**: A　　**文章编号**: 1007 - 9807(2023)05 - 0103 - 18

0　引　言

全面准确地了解管理对象是有效制定管理方案、科学进行管理决策的重要基础. 抽样调查由于具有调查费用低、调查周期短、时效性强等特点,在公共卫生、科学研究、心理学、商业和社会管理等领域得到了广泛应用[1]. 然而,在大多数的社会调查实践中,管理决策经常涉及到一些隐私或敏感性问题,即具有私人机密性或大多数人认为不便在公开场合表态及陈述的问题[2],例如收入状况、职场关系、离职意向、价值取向、疾病状况、不当行为等[3,4]. 针对敏感性问题的调查能够为政府、企业、学校等提供关于社会政治、经济和生活等方面的重要管理决策信息,但由于牵涉到被调查者的隐私或者利益,在调查过程中往往会引起调查对象的难堪或抵触心理,进而不配合调查或者给出不真实的回答,产生无应答偏倚[5]或社会期望偏倚[6]等,影响调查结果的真实性和可靠性,这给通过样本来进行总体推断带来了极大挑战. 与此同时,在促进数字经济发展中,数据治理对于营造良好环境具有重要作用[7,8],这进而对数据的质量与完备性以及数据隐私保护提出了更高要求[9,10]. 因此,在保护个体隐私的前提下鼓励调查对象提供敏感性问题的真实回答,设计开发隐私规避的统计方法以提高调查对象的应答

① 收稿日期: 2021 - 12 - 31; 修订日期: 2022 - 12 - 26.
基金项目: 国家自然科学基金资助重大研究计划项目(91846301); 国家杰出青年科学基金资助项目(72025405); 国家社会科学基金资助重大项目(22ZDA102).
作者简介: 吕　欣(1984—), 男, 湖南常德人, 博士, 教授, 博士生导师. Email: xin_lyu@sina.com

率、降低和消除不真实回答的影响,具有重要理论和应用价值.

针对隐私或敏感性问题的统计调查方法主要从敏感人群抽样设计和敏感问题应答设计两方面开展. 一方面,对于天然具有隐私和敏感属性的一些社会群体,如被社会忽视或极少关注、被主流社会和文化所排斥的边缘人群等,他们具有难接触、隐匿性等特点,由于没有抽样框,传统的随机抽样、分层抽样等方法很难实施,因此主要通过非概率抽样方法对其进行研究[11]. 非概率抽样方法的最简单形式是便利抽样[12],是指招募在地点、时间和精力等方面最容易接近的人,包括采访朋友、商场拦截采访、访问最近的家庭样本、网站招募参与者等方式. 例如,基于设施的抽样被广泛应用于研究特定疾病相关的高危人群[13]. 目的抽样是一种“更严格”的非概率抽样过程[14],是指在选择受访者时要牢记研究目的并寻找满足要求的人,其变体包括异质性抽样、专家抽样、关键信息人抽样等,通常被用于研究极端或异常情况,例如杰出成就或重大失败、极端事件或危机等[15,16]. 配额抽样类似于分层抽样,不同的是研究人员可以自行设置配额(个体的入样概率)[17]. 滚雪球抽样是最著名的“链式”(chain-referral)抽样方法之一,通过招募若干个参与者作为种子,再要求参与者推荐他们所认识的人来进行进一步招募,从而形成一个“滚雪球”样本[18]. 此外,目标抽样、时间 – 地址抽样等方法及其变体也被广泛用于研究难接触人群[19,20].

另一方面,为调查研究对象的隐私或敏感信息,统计学家开发了一系列应答技术来规避受访者的隐私泄露问题. 其中,随机化回答技术(Randomized Response Technique, RRT)[21,22]是指通过采用特定的随机化装置(如掷骰子或抽卡片),使被调查者以一个预定的基础概率 p 从两个或两个以上的问题中选择一个问题进行回答,除被调查者以外的所有人(包括调查者)均不知道被调查者的回答是哪一个问题,最后根据概率论知识计算出敏感问题特征在人群中的真实分布情况. 例如,在最早的 Warner 模型中包含了两个互为对立面的陈述:具有敏感特征 A(例如,是病毒携带者)或不具有敏感特征 A(例如,不是病毒携带者)[22]. 在 Warner 模型的基础上,还发展出 Simmons 不相关问题模型[23]、Greenberg 双无关问题模型[24]、Mangat 模型[25]、混合随机化回答模型[26]、无关联问题随机化回答模型、两阶段随机化回答模型等,进一步提高了调查结果的保密性和准确性. 值得一提的是,随机化回答技术多次被荷兰政府用于社会福利欺诈行为调查研究. 非随机化回答技术(Non-Randomized Response Technique, NRRT)[27]则在敏感信息搜集过程中不使用任何随机化装置,使被试者能够更容易理解调查机制并给出真实答案,最终提高调查结果的真实性和有效性. 典型的非随机化回答技术包括十字交叉模型(Warner 模型的非随机化版本)[28]、三角模型[28]、对角模型[29]、平行模型[29]等,在特定行为相关的敏感信息收集中得到了验证与应用[30]. 负调查重构方法[31]是指利用负调查让参与者提供与自身类别不同的类别,再通过特定的统计学方法(称之为重构方法)从负调查结果中重构出所有参与者的敏感信息分布. 该方法可以在保护个人敏感信息隐私的前提下完成敏感信息收集,被用于用户位置信息、传感器网络数据等的收集任务[32]. 此外,非匹配计数技术(Unmatched Count Technique, UCT)[33,34]、条目计数技术(Item Count Technique, ICT)、三卡片法(Three-Card Method, TCM)[35]等也被用于敏感性问题调查.

尽管上述方法在提高受访者如实报告的意愿等方面更为有效[36,37],但在实施过程中仍存在很大的局限性. 例如,RRT 设计复杂、实施难度较高,不仅需要对调查者进行严格培训,而且容易使受访者难以理解调查机制;与此同时,有些方法会产生高方差,RRT 的方差在许多情况下至少是常规估计方差的四倍[38]. 由于难接触人群具有较强的敏感性和隐匿性,导致大部分传统的概率抽样方法无法实施,便利抽样、滚雪球抽样等非概率抽样方法虽然一定程度上解决了难接触的问题,但针对样本的结论仍然无法推广到总体,不能实现对总体的无偏估计,这也是非概率抽样的最大局限性. 此外,大部分方法在本质上难以解决错误回答或回答不完整等现实问题.

随着社交网络抽样与统计推断问题的提出,同伴驱动抽样(Respondent Driven Sampling, RDS)成为一类极具前景的难接触人群抽样调查方法,在针对特殊人群的调查研究中得到了广

泛应用,其实证研究遍布 120 多个国家[39,40]. RDS 在总体构成的社交网络上执行链式抽样,并依据每个节点的入样概率去校正总体估计的误差,具有完备的样本获取和统计推断理论,不仅操作简单便利,而且能设计基于样本数据的总体无偏估计量,从而提高群体特征估计的精度[41]. 例如,在估计目标人群中具备某个特征的人员组成时,研究人员可以招募若干该人群样本(入样节点)参加调查,并让入样节点推荐其认识的其他此人群成员来参与调查,通过直接获取当前入样节点的相关属性信息用于统计推断,进而无偏估计目标人群中具备被关心特征的人群比例.

然而,在包含敏感性问题的调查中,上述抽样策略访问的样本均不可避免的会出现样本属性信息"缺失"或"不可信"的问题;甚至有可能会出现入样节点属性都"不可信"或"未知"的极端情况. 针对该问题,本文借鉴基于社交网络的抽样推断方法思路,探索一种基于中心网络的抽样推断方法,并创新性地把网络抽样中获取样本节点信息扩展到获取样本节点的邻居节点信息,从而有效解决敏感人群自我报告数据的不完整或不真实的问题. 研究表明,人们更愿意谈论朋友的敏感特征,而不是自己的敏感特征,并且他们可以对这些朋友的具体数量提供可靠的回答[38,42,43]. 因此,在敏感人群的朋友信息相对容易获取的前提下,去收集邻居节点的相关信息,并利用潜在的网络节点去进行统计推断可能有效实现对总体的无偏估计.

1　模型方法

在敏感信息抽样调查中,目标是估计总体中具备某类敏感特征(属性)A 的人群比例 $P(A)$. 在概率抽样条件下,以完全随机抽样为例,可根据抽样个体(受访者)提供的自身属性信息(A 或非 A,非 A 可用 B 表示),对该总体参数 $P(A)$ 进行估计,即 $\hat{P}(A) = S_A/S$. 其中 S_A 为样本中自报告属性为 A 的个体数量, S 为样本规模. 但是,在抽样实践过程中,让受访者提供自身敏感属性信息往往存在较大难度,同时,受访者为了保护自身隐私,提供的属性信息往往存在"不可靠"的问题.

为解决上述困难,本文提出中心网络抽样统计推断方法(Ego-Centric sampling Method, ECM),依托调查对象的社交网络,不直接收集样本自身的属性信息(self-reported information),转而收集样本的中心网络信息(ego-centric information),然后利用网络潜在结构去得到总体参数 $P(A)$ (如特定人群比例、特定行为发生率等)的估计 $\hat{P}(A)$. ECM 方法主要包括两个步骤:1)隐私规避的中心网络抽样,即抽取一定数量的样本节点,获取样本节点的度和其邻居的目标属性;2)总体特征间接估计,即利用样本节点报告的邻居属性信息对该属性总体值进行估计. 此外,本文提出 3)基于 bootstrap 的置信区间估计方法,实现对总体估计结果的可信度评估.

1.1　隐私规避的中心网络抽样方法

基于隐私规避的中心网络抽样方法的基本思想是:在由目标总体的朋友/熟人等关系构成的社交网络上,通过受访者(网络中心节点)对其邻居节点关于敏感属性 A 的间接报告,根据网络结构的数学性质构建总体参数 $P(A)$ 的估计量. 其具体实施过程如图 1 左侧所示,首先通过随机采样或链式采样的方式获取样本节点(敏感属性信息未知),然后获取这些样本节点的中心网络中邻居节点(敏感属性信息已知)的数量. 由此,可以得到若干个"中心网络"(ego network),属性未知的样本节点便是这个中心网络的中心(ego),收集到属性信息的节点便是这个中心的邻居(neighbor). 在没有抽样框的条件下,可以采用链式抽样的方式,通过随机游走抽样、同伴驱动抽样等方法采集链路上每个节点的邻居信息,这在复杂的难接触人群抽样、限制措施较多的网络爬虫数据获取等场景中都更灵活适用[44].

需要注意的是,该抽样方法建立在"邻居信息"已知的前提下. 实际上,在很多场景下,相比于直接获取研究对象自身信息,获取研究对象邻居(或朋友)的信息会更加容易. 例如,在社会学中,对于隐私性强、敏感性高的个体,难以直接获取其自身的特定信息,但可以利用社交网络来获取其敏感性较低的"朋友"的相关信息;在网络数据爬取中,虽然无法直接获取设置了隐私保护的

用户敏感信息,但可以爬取此类用户的朋友(如粉丝)的这些信息. 相比之下,这种不泄露被调查者自身信息的方式会更加容易收集到有价值的信息. 许多研究表明,收集社交网络中中心网络信息的方式在实际操作中具有较高的可行性[45,46],这是有信心去进行中心网络抽样的重要依据. 在收集中心网络信息时,通过询问受访者两个问题予以实现,即"在受访群体中你的好友(或者熟人)的数量?"(获取受访个体在总体社交网络中的度)和"好友(或熟人)中拥有 A 类属性的数量?"(获取受访个体中心网络中 A 类属性邻居的数量);注意,两个问题都是"匿名"问题、且只关乎数量信息,不需要受访者提供具体的"邻居"名单.

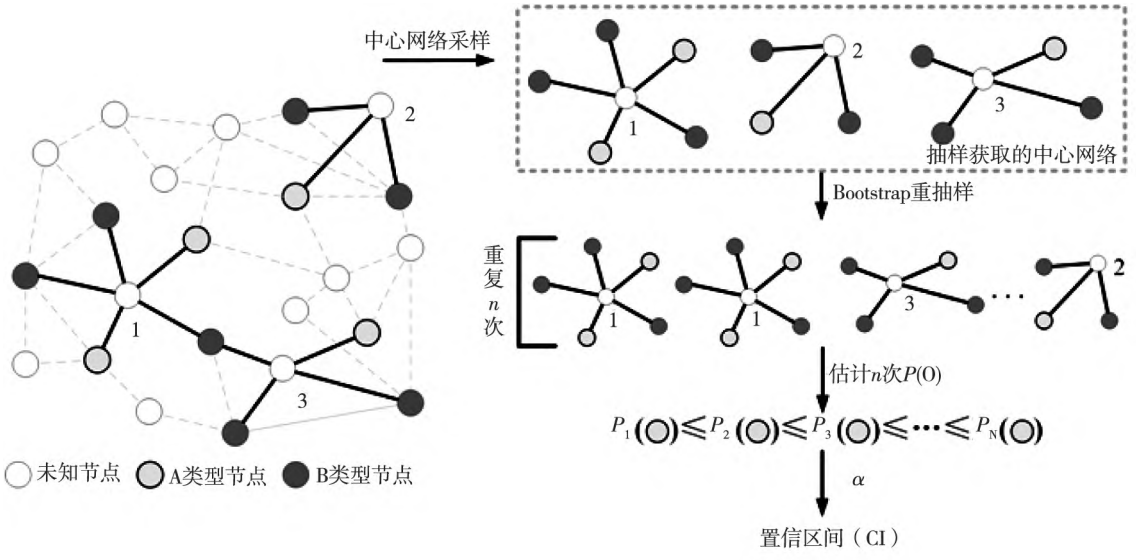

图 1　中心网络抽样方法示意图

Fig. 1 Ego-centric sampling method (ECM)

1.2　总体特征间接估计

由于入样概率无法提前设计,一般条件下非概率抽样方法无法根据所得样本对总体情况进行无偏的统计推断. 在基于中心网络的隐私规避抽样方法中,巧妙地利用了网络中边相等的原则来建立估计量方程,得到对总体参数 $P(\mathrm{A})$ 的间接估计 $\hat{P}(\mathrm{A})$.

设抽样总体的社交网络 $G = \{V,E,Z\}$ 为无向连通图,其中 $V = \{v_1,v_2,\cdots,v_N\}$ 表示网络中的节点集合, $E = \{e_{ij}\} \subseteq V \times V$ 表示 N 个节点间的边集合,并有 $e_{ij} = 1$ 表示 v_i 与 v_j 之间存在边(如朋友/熟人关系),否则 $e_{ij} = 0$. 令 $Z = \{a_1,a_2,\cdots,a_N\}$ 表示节点属性 A(非 A 的属性记为 B)的集合,若节点 i 的属性为 A,则 $a_i = 1$, 否则(即属性为 B) $a_i = 0$.

在此基础上,利用中心网络抽样得到的受访者 i(即中心网络的中心节点)的邻居信息可以表示为

1)受访者在网络中的邻居数量,即节点 i 的度 $k_i = \sum_j e_{ij}$;

2)受访者在网络中的 A 类属性邻居数量,即 $k_i^{\mathrm{A}} = \sum_{e_{ij}=1} a_j$.

由于 G 是无向网络, $\forall\ 1 \leqslant i,j \leqslant N$,有 $e_{ij} = e_{ji}$,进而网络中由 A 类属性节点(简称"A 类节点")出发指向 B 类属性节点(简称"B 类节点")的边数量 E_{AB} 应该与由 B 类节点指向 A 类节点的边数量 E_{BA} 相等,即

$$E_{\mathrm{AB}} = E_{\mathrm{BA}} \tag{1}$$

令 $P_{\mathrm{A}}(k_i)$ 、$P_{\mathrm{B}}(k_i)$ 分别为度为 k_i 的节点是 A 类、B 类节点的概率, k_i^{A} 、k_i^{B} 分别为度为 k_i 的节点的邻居中 A 类属性节点和 B 类属性节点的数量. 则 E_{AB} 、E_{BA} 的期望值分别为

$$\mathrm{E}(E_{\mathrm{AB}}) = \sum_{i=1}^{N} P_{\mathrm{A}}(k_i) k_i^{\mathrm{B}} \tag{2}$$

$$\mathrm{E}(E_{\mathrm{BA}}) = \sum_{i=1}^{N} P_{\mathrm{B}}(k_i) k_i^{\mathrm{A}} \tag{3}$$

由式(1)、$P_{\mathrm{A}}(k_i) + P_{\mathrm{B}}(k_i) = 1$ 和 $k_i^{\mathrm{A}} + k_i^{\mathrm{B}} = k_i$,可得

$$\sum_{i=1}^{N} P_{\mathrm{A}}(k_i)(k_i - k_i^{\mathrm{A}}) = \sum_{i=1}^{N} (1 - P_{\mathrm{A}}(k_i))k_i^{\mathrm{A}} \quad (4)$$

由式(4)化简得到

$$\sum_{i=1}^{N} P_{\mathrm{A}}(k_i)k_i = \sum_{i=1}^{N} k_i^{\mathrm{A}} \quad (5)$$

由此,对于所有度为 k 的节点有

$$n_k P_{\mathrm{A}}(k)k = \sum_{k_i=k} k_i^{\mathrm{A}} \quad (6)$$

其中 $P_{\mathrm{A}}(k)$ 为度为 k 的节点为 A 类节点的概率,n_k 为 G 中度为 k 的节点数量.

而 A 类属性节点数 N_{A} 的期望值可表示为

$$\mathrm{E}(N_{\mathrm{A}}) = \sum_{k=\mathrm{K}_{\min}}^{\mathrm{K}_{\max}} P_{\mathrm{A}}(k)n_k \quad (7)$$

其中 $K_{\min}$ 及 $K_{\max}$ 为网络中节点度的最小值和最大值. 则 G 中 A 类节点的比例 $P(\mathrm{A})$ 可表示为

$$P(\mathrm{A}) = \frac{\sum_{k=K_{\min}}^{K_{\max}} \sum_{k_i=k} k_i^{\mathrm{A}}/k}{N} \quad (8)$$

当中心网络节点为随机抽样样本时,样本总量 S 中 A 类节点的比例 $\hat{P}(\mathrm{A})$ 即为 $P(\mathrm{A})$ 的无偏估计,即

$$\hat{P}(\mathrm{A}) = \frac{\sum_{k=k_{\min}}^{k_{\max}} \sum_{k_i=k} k_i^{\mathrm{A}}/k}{S} \quad (9)$$

利用上式,完成了利用中心网络信息对总体参数 $P(\mathrm{A})$ 的估计. 在抽样调查过程中,只需收集中心节点的度信息以及中心节点各属性的邻居数量这两个简单的信息,就能够根据上述估计量有效计算和推断出总体参数 $P(\mathrm{A})$ 的真实值.

实际操作过程中,由于获取中心网络的全部邻居信息在一些场景下仍然是较难完成的,例如在线社交网络中,用户的好友(或关注)数量可能非常多,对应中心网络的规模(即中心节点的度)也很大,但多数网络平台对爬虫的访问频率都有严格的限制,因此很多情况下通过爬虫往往只能得到中心网络的部分邻居信息. 针对这种情况,本文采用选取中心网络中部分邻居的策略进行统计推断. 对于抽取到的中心节点 i,若从它的全部邻居中随机选择 k_r 个邻居(或仅 k_r 个邻居信息已知),可将式(9)转化为

$$\hat{P}_R(\mathrm{A}) = \frac{\sum_{r=1}^{S} k_r^{\mathrm{A}}/k_r}{S} \quad (10)$$

其中 k_r^A 为选择的 k_r 个邻居中 A 类节点的数量. 通过式(10),能够通过只选取部分邻居信息或在只有部分邻居信息已知的场景下,实现对总体参数的有效估计.

1.3　置信区间估计的 bootstrap 方法

由于网络抽样方法通常受网络度分布、节点异质性等许多因素的影响,其方差估计的求解通常是不可行的. 为此,本文采用非概率抽样方法中置信区间估计较为通用的 bootstrap 方法[47,48],构造中心网络抽样估计值 $\hat{P}(\mathrm{A})$ 的置信区间(Confidence Interval, CI),主要步骤如下(图 1 右侧).

1)对抽样得到的中心网络(包括中心和邻居)进行有放回的重抽样,得到一个重抽样中心网络集合 B_1;

2)根据 B_1 获取的中心网络信息,利用本文提出的总体特征间接估计方法计算总体参数的估计值 $\hat{P}_1(\mathrm{A})$;

3)重复步骤 1)、步骤 2)ℓ次,分别得到 ℓ 个总体参数的估计值 $\hat{P}_1(\mathrm{A}),\hat{P}_2(\mathrm{A}),\cdots,\hat{P}_{\ell}(\mathrm{A})$;

4)将 $\hat{P}_1(\mathrm{A}),\hat{P}_2(\mathrm{A}),\cdots,\hat{P}_{\ell}(\mathrm{A})$ 由小到大排列得到 $\hat{P}_1^{in}(\mathrm{A}),\hat{P}_2^{in}(\mathrm{A}),\cdots,\hat{P}_{\ell}^{in}(\mathrm{A})$;

5)若置信水平为 $1-\alpha$,则最终由百分位数法得到置信区间 $\left[\hat{P}^{in}_{\frac{\alpha}{2}\ell}(\mathrm{A}),\ \hat{P}^{in}_{(1-\frac{\alpha}{2})\ell}(\mathrm{A})\right]$.

2　大规模社交网络试验

2.1　网络数据

为了验证提出方法的有效性,分别在仿真网络和真实网络上对本文提出的中心网络抽样方法进行检验. 首先,通过设置不同总体比例的仿真网络,分别对隐私规避的中心网络抽样方法的总体特征间接推断方法、基于 bootstrap 的置信区间估计方法的真实性能进行检测. 此外,在某大规模难接触人群的真实在线社交网络上,通过对其中多种类型的个体比例进行估计,与真实数据进行对比以验证 ECM 在真实社交网络上的可行性与可靠性.

在仿真实验部分,主要通过生成三种不同的模型网络,分别为 ER 网络[49]、BA 无标度网络[50,51]

以及 KOSKK 网络[52],来验证本文方法的性能. ER 网络模型是经典的随机网络模型,由于节点之间随机连边,所以网络中大部分节点度都与平均度接近. BA 网络模型则是通过优先连接的生长过程生成网络,即拥有更多边的节点更易与新节点产生连接. 虽然 ER 和 BA 这两个网络模型与真实的社会网络结构仍然存在较大差异,但常用于测试实验模型或算法在网络上的性能. 作为模型网络的重要补充,本文将另一个广泛应用的 KOSKK 网络也应用于实验测试中. KOSKK 网络能够生成与真实世界社会网络相似的网络结构,被认为是可以模拟真实社交关系的最好模型之一. 在本文的仿真实验中,生成的 ER 网络、BA 网络和 KOSKK 网络都包含 10 000 个节点和 100 000条边,并在各网络中分别随机选取比例为 10%、20%、30%和40%的节点,将其属性设置为 A. 即对于每一种网络模型, $P(\mathrm{A})$ 分别为 0.1、0.2、0.3 和0.4.

除仿真网络以外,本文还在匿名的真实社交网络上验证所提方法的性能. 该社交网络由某难接触人群社交平台上用户之间的"相互关注"关系形成[53,54],共包含 556 627 个节点,以及16 963 498条无向边. 其中每个节点具有三种特征,分别为 i)年龄(是否小于 30 岁);ii)所在国家(是否在中国);iii)所在洲(是否在亚洲). 统计得到这三种特征在总体中的比例分别为 0.477、0.614 和 0.793,且该网络的度分布是极度不均匀的,网络节点平均度为 30.5,但其中 80%节点的度小于 28,即 80%以上的用户只有少于 28 个相互关注的好友.

2.2 试验设计

在每次实验中,随机选取实验网络中 10%的节点,并收集这些节点相应的中心网络信息,包括节点的度信息(邻居数量)以及其邻居中具有 A 属性的节点数量. 然后应用上文中的估计方法,通过收集到的中心网络信息对总体比例 $P(\mathrm{A})$ 进行估计. 对于每个网络,将重复仿真实验 500 次以消除随机误差的影响. 本文通过计算中心网络推断方法的估计值与真实比例间的差值来衡量该抽样方法的性能,即

$$Bias = |\hat{P}(\mathrm{A}) - P(\mathrm{A})| \tag{11}$$

$Bias$ 越小,则表明估计值与真实值间的偏差越小,该方法的推断效果越好. 对于 M 次重复实验的估计结果 $\hat{P}_m(\mathrm{A})$ ($m = 1,2,\cdots,M$),则用每次偏差的平均值来衡量最终的估计效果,即

$$\overline{Bias} = \frac{\sum_{m=1}^{M} |\hat{P}_m(\mathrm{A}) - P(\mathrm{A})|}{M} \tag{12}$$

此外,为了对上文提出的基于 bootstrap 的置信区间估计方法进行验证,还将利用具有不同总体比例配置的三个模型网络(ER、BA 和 KOSKK)进行测试实验. 对于具有不同比例属性 A 节点 $P(\mathrm{A})$ 的网络,首先抽取其中 10%的节点,采集他们的中心网络作为样本. 按照上文提出的置信区间估计方法计算出 95%置信度水平的置信区间:即将抽取到的中心网络作为总体,进行重抽样,每次重抽样获取 1 000 个中心网络,然后根据这 1 000 个中心网络的信息估计得到置信水平为 95%的置信区间. 重复该过程 1 000 次,得到 1 000 个置信区间,然后计算总体变量的真实值(即真实比例 $P(\mathrm{A})$)落在构建的置信区间内的比率 Ω,也称为覆盖度(coverage probability).

除了考虑不同网络类型(BA、ER、KOSKK)及不同 A 类属性节点比例(0.1、0.2、0.3、0.4)以外,本文主要从以下四个方面进一步探索影响 ECM 方法性能的因素.

1)网络密度(density)

网络密度用于刻画网络中节点连接的紧密程度,其定义为

$$\Psi = 2E/[N(N-1)] \tag{13}$$

其中 E 为网络中实际存在的连边数, N 为网络中的节点个数.

为了评估网络密度对 ECM 效果的影响,根据已生成的 BA 网络($\Psi = 0.002$),采取随机加边策略,得到 $0.004 \leqslant \Psi \leqslant 0.012$ 的 10 个 BA 网络. 具体地,在已生成的 BA 网络上,随机选择两个节点 v_i 和 v_j ,若 v_i 和 v_j 之前没有边相连,则将 v_i 和 v_j 相连接;重复上述过程,直至达到设定的网络密度为止.

2)聚集系数(clustering coefficient)

聚集系数用于刻画网络中各个节点结集成团的程度，即聚集性. 某个节点 v_i 的聚集系数 C_i 可以定义为

$$C_i = \frac{\text{包含 } v_i \text{ 的三角形的数目}}{\text{以 } v_i \text{ 为中心的连通三元组的数目}} \quad (14)$$

对于整个网络，可用平均聚集系数 C_{avg} 来刻画网络的聚集性

$$C_{avg} = \sum_{i=1}^{N} C_i/N \quad (15)$$

为了评估网络聚集性对 ECM 效果的影响，根据已生成的 BA 网络（$C_{avg} = 0.007$），采取随机加边策略，得到 $0.008 \leqslant C_{avg} \leqslant 0.017$ 的 10 个 BA 网络. 具体地，在已生成的 BA 网络上，随机选择任意节点 v_i，得到其邻居节点集合；从 v_i 的邻居节点集合中，随机选择邻居节点 v_j 和 v_k，若 v_j 和 v_k 之间没有边相连，则将 v_j 和 v_k 相连接；重复上述过程，直至达到设定的平均聚集系数为止.

3）同质性（homophily）

同质性用于刻画网络中的节点与其属性相同节点相连的特性，若该属性为 A，则同质性 h_A 可定义为

$$h_A = 1 - s_{AB}/(1 - P(A)) \quad (16)$$

其中 $s_{AB} = E_{AB}/(E_{AB} + E_{AA})$ 为网络中 A 属性节点指向 B 属性节点的连边占所有 A 属性节点连边的比例.

为了评估同质性对 ECM 效果的影响，根据已生成的 BA 网络（$h_A \approx 0.0$），采取断边重连的策略，得到 $0.02 \leqslant h_A \leqslant -0.20$ 的 10 个 BA 网路. 具体地，在已生成的 BA 网络上，随机选择两条边 (v_i, v_j) 和 (v_k, v_l)，其中 v_i、v_k 的属性为 A，v_j、v_l 的属性为 B（非 A 属性）；而后将 (v_i, v_j) 和 (v_k, v_l) 断开，分别将 v_i 和 v_k、v_j 和 v_l 相连，得到边 (v_i, v_k) 和 (v_j, v_l)；重复上述过程，直至到达设定的同质性为止.

4）活跃系数（activity ratio）

活跃系数[55]用于量化不同属性节点网络平均度的系统性差异，定义为不同属性节点平均度的比值

$$AR = \frac{\sum_{i \in A} k_i/N_A}{\sum_{j \in B} k_j/N_B} \quad (17)$$

其中 N_A 和 N_B 分别表示属性为 A 和属性为 B 的节点数量. 如果 $AR = 1$，则表明节点属性与节点中心网络的规模无关.

为了评估活跃系数对 ECM 效果的影响，根据已生成的 BA 网络（$AR \approx 1.0$），采取属性交换策略，得到 $0.5 \leqslant AR \leqslant 0.9$ 和 $1.1 \leqslant AR \leqslant 1.5$ 的 10 个 BA 网络. 具体地，在已生成的 BA 网络上，若需要生成网络的活跃系数小于已生成 BA 网络，则随机选择属性为 A 的节点 v_i 和属性为 B 的节点 v_j，如果 v_i 的度 k_i 大于 v_j 的度 k_j，则交换 v_i 和 v_j 的属性，即将 v_i 的属性置为 B，将 v_j 的属性置为 A，重复上述过程，直至到达设定的活跃系数为止；若需要生成网络的活跃系数大于已生成 BA 网络，则随机选择属性为 A 的节点 v_i 和属性为 B 的节点 v_j，如果 v_i 的度 k_i 小于 v_j 的度 k_j，则交换 v_i 和 v_j 的属性，即将 v_i 的属性置为 B，将 v_j 的属性置为 A，重复上述过程，直至到达设定的活跃系数为止.

2.3　试验结果

针对 BA 模型，分别随机抽取 1 到 1 000 个中心网络（即样本），并在 $P(A)$ 分别取 0.1、0.2、0.3 和 0.4 的情况下进行仿真实验（重复 500 次），得到结果如图 2 所示. 可见，在不同的 $P(A)$ 下，ECM 方法得到的总体变量均值估计总是接近总体真实值 $P(A)$，并且，随着样本规模的增大，ECM 估计值的波动越来越小，并很快稳定在真实值附近.

进一步在不同 $P(A)$ 配置的 ER、BA 和 KOSKK 这三种不同类型的模型网络上对 ECM 性能进行仿真实验，结果如表 1 所示. 可以看出，中心网络推断方法得出的估计值基本都分布在真实值附近，即 Bias 都位于 0 附近（图 3）. 当 ECM 采集所有样本的邻居信息时，$P(A)$ 在所有场景下的误差均小于 0.001，且在多次仿真实验中，每次对 $P(A)$ 的估计值波动都不大，总体变量均值估计的最大误差都小于 0.015. 当 ECM 仅采集样本节点的 3 个或 5 个邻居节点信息时，使用式（8）得到的总体估计值 $\hat{P}_{n3}(A)$ 和 $\hat{P}_{n5}(A)$ 产生的最大误差也不超过 0.004，展示了该方法极强的鲁棒性和操作过程的灵活性.

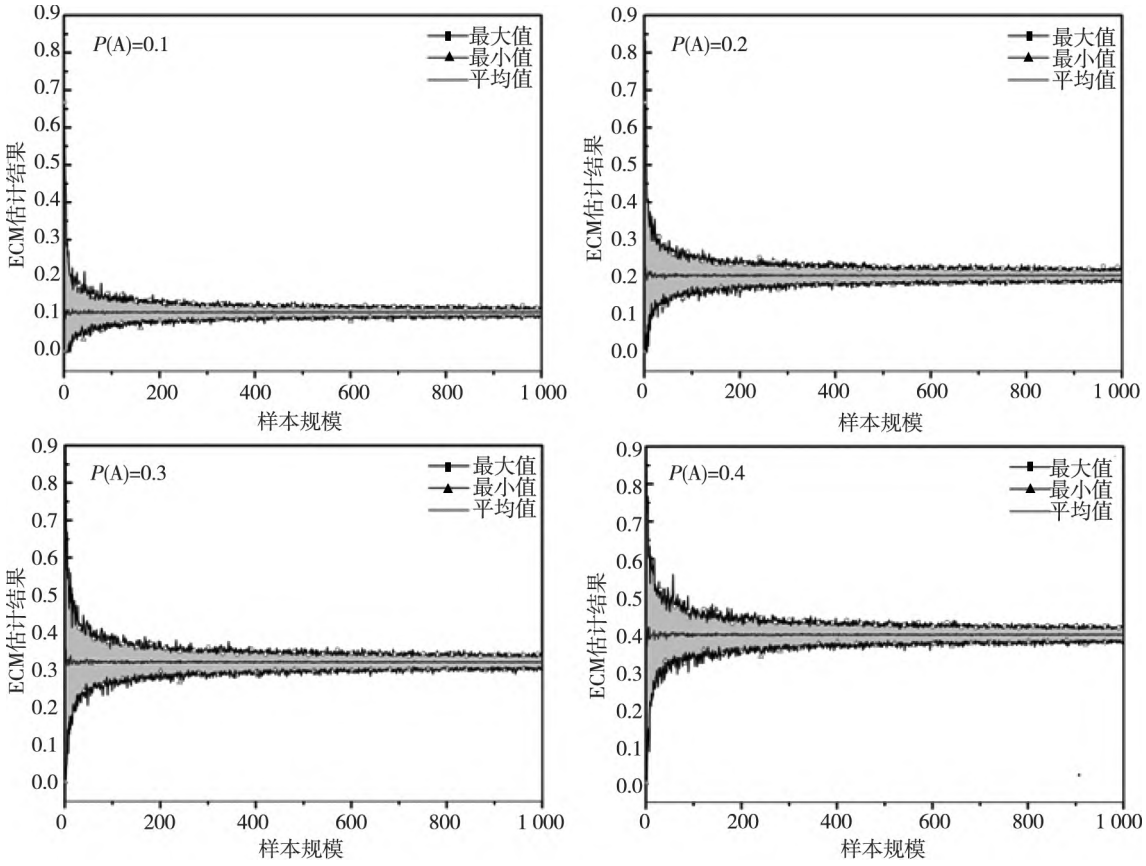

图 2　ECM 样本规模与估计结果

Fig. 2 Sample size and estimated results of ECM

表 1　ECM 在仿真网络上的实验结果

Table 1 Experimental results of ECM on simulated networks

网络模型	P(A)/%	$\hat{P}$(A)/%	$\hat{P}_{n3}$(A)/%	$\hat{P}_{n5}$(A)/%
ER	10.0	9.9	9.9	9.8
	20.0	20.1	20.1	20.1
	30.0	29.9	30.0	29.9
	40.0	39.9	39.9	39.8
BA	10.0	10.3	10.4	10.4
	20.0	20.0	19.9	20.0
	30.0	29.9	29.7	29.8
	40.0	39.9	40.3	39.8
KOSKK	10.0	10.0	9.9	10.1
	20.0	20.1	20.3	20.1
	30.0	30.0	29.9	29.9
	40.0	40.1	40.0	40.1

注：$\hat{P}_{n3}$(A) 和 $\hat{P}_{n5}$(A) 为只随机选取中心节点 3 个或 5 个邻居时的 ECM 估计值.

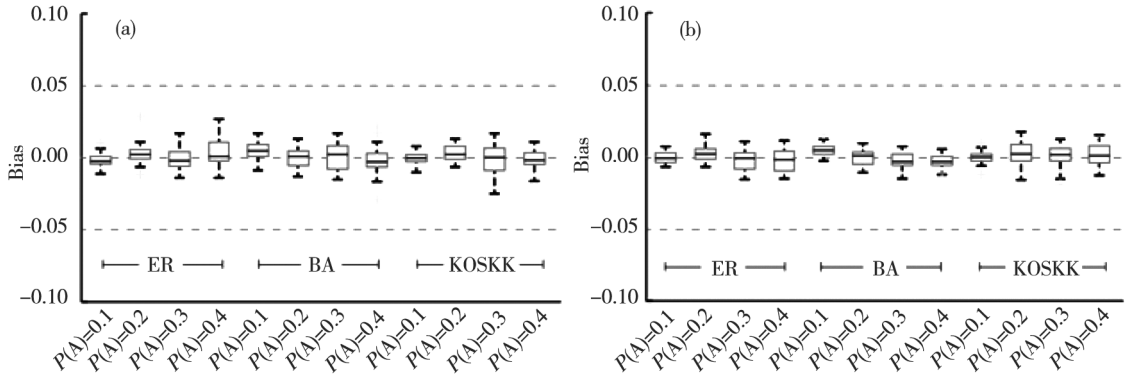

图 3　ECM 在仿真网络上的估计偏差. (a) $\hat{P}_{n3}(A)$；(b) $\hat{P}_{n5}(A)$

Fig. 3 Estimation bias of ECM on simulated network. (a) $\hat{P}_{n3}(A)$; (b) $\hat{P}_{n5}(A)$

表 2　基于 bootstrap 的 95% 置信区间估计覆盖度 Ω

Table 2 95% confidence interval estimation coverage (Ω) based on bootstrap

$P(A)$	ER 网络/%	BA 网络/%	KOSKK 网络/%
0.1	93.0	88.5	95.5
0.2	93.0	97.0	94.5
0.3	94.5	96.0	95.5
0.4	97.0	93.0	97.5

表 2 展示了在这三类模型网络中对本文提出的基于 bootstrap 的置信区间估计方法验证的结果. 虽然在 $P(A) = 0.1$ 的 BA 网络上估计得到的 95%CI 的覆盖度最小,但也接近 90%(88.5%). 而在其它不同总体变量配置的网络中,本文方法得到的置信区间的覆盖度均超过 90%,且其中 6 个的覆盖度超过 95%. 值得注意的是,在大量其他网络抽样技术中使用 bootstrap 方法进行置信区间估计的覆盖度仅能达到 60% ~ 80%[47],显然,ECM 方法的置信区间估计更加有效.

在真实在线社交网络的实验中,对网络中三个属性的节点组成情况估计结果如图 4 所示. 可以看出,该方法对每个属性的总体比例估计结果都与真实值高度一致,每个特征的估计偏差都很小. 其中,对“年龄”这一特征的估计偏差为 0.033,“所在国家”为 0.031,“所在洲”为 0.030. 这一结果表明,本文提出的隐私规避的中心网络抽样方法能够对各类人群比例进行可靠的估计,即使在真实社交网络上也仍然有效.

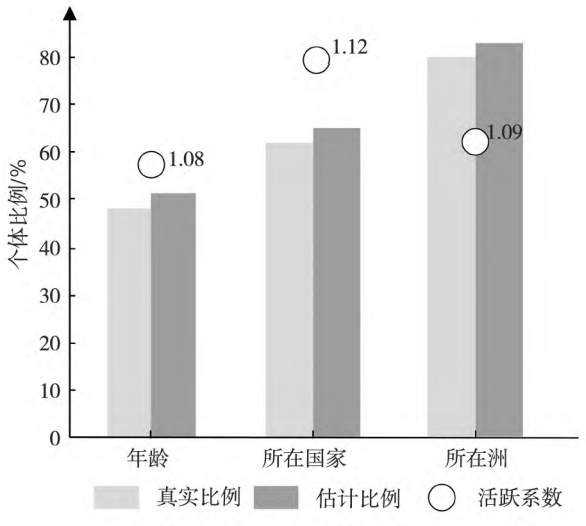

图 4　ECM 在真实在线社交网络上的实验结果

Fig. 4 Experimental results of ECM on a real-world online social network

上述不论是仿真网络实验还是真实网络实验,都很好地证明了基于隐私规避的中心网络抽样方法的性能,即使在总体比例不同、网络结构不同的模型网络中该方法仍然十分可靠,在复杂的真实社交网络结构上也同样能够取得很好的结果.

进一步,为验证不同网络结构对 ECM 抽样和估计方法的影响,本文测试了在不同网络密度、平均聚集系数、节点同质性以及活跃系数条件下 ECM 估计误差的变化情况,如图 5、附图 1 ~ 附图 4 所示. 显然,在不同 $P(\mathrm{A})$ 设置下,ECM 估计对网络密度、平均聚集系数的变化非常鲁棒,当网络密度从 0.002 增大 6 倍至 0.012、平均聚集系数从 0.007 增大至 0.017 时,ECM 估计值并未呈现显著变化,均能得到总体变量的较好估计结果. 由于 $h_{\mathrm{A}} > 0$ 表明网络中 A 类节点更倾向于与同类节点产生连接,即其邻居网络中将有更多同属性节点,根据式(7),ECM 估计将偏大。但总的来说,这种偏差并不大,在 h_{A} 从 0.00 增大到 0.20 的所有试验中,$\hat{P}(\mathrm{A})$ 的误差均小于 4%,说明 ECM 依然可以得到较好的总体变量估计结果.

由于活跃系数直接度量了不同属性节点网络规模的系统性差异,当活跃系数(AR)等于 1 时,ECM 的估计结果与总体变量的真实值差距很小. 当 AR 小于 1 或者大于 1 时,ECM 的估计结果会逐渐偏离总体变量真实值. 当 AR 大于 1 时,ECM 得到的估计结果会偏大,而当 AR 小于 1 时,其估计的结果会偏小. 但是,在 AR 并不极端的范围内,ECM 依然能取得较好的估计结果:在 $P(\mathrm{A}) = 0.1$、$P(\mathrm{A}) = 0.2$ 和 $P(\mathrm{A}) = 0.3$ 的设置下,当 $0.9 \leqslant AR \leqslant 1.1$ 时,ECM 估计的相对误差在 10%以内;在 $P(\mathrm{A}) = 0.4$ 的设置下,当 $0.9 \leqslant AR \leqslant 1.2$ 时,ECM 估计的相对误差仍在 10%以内.

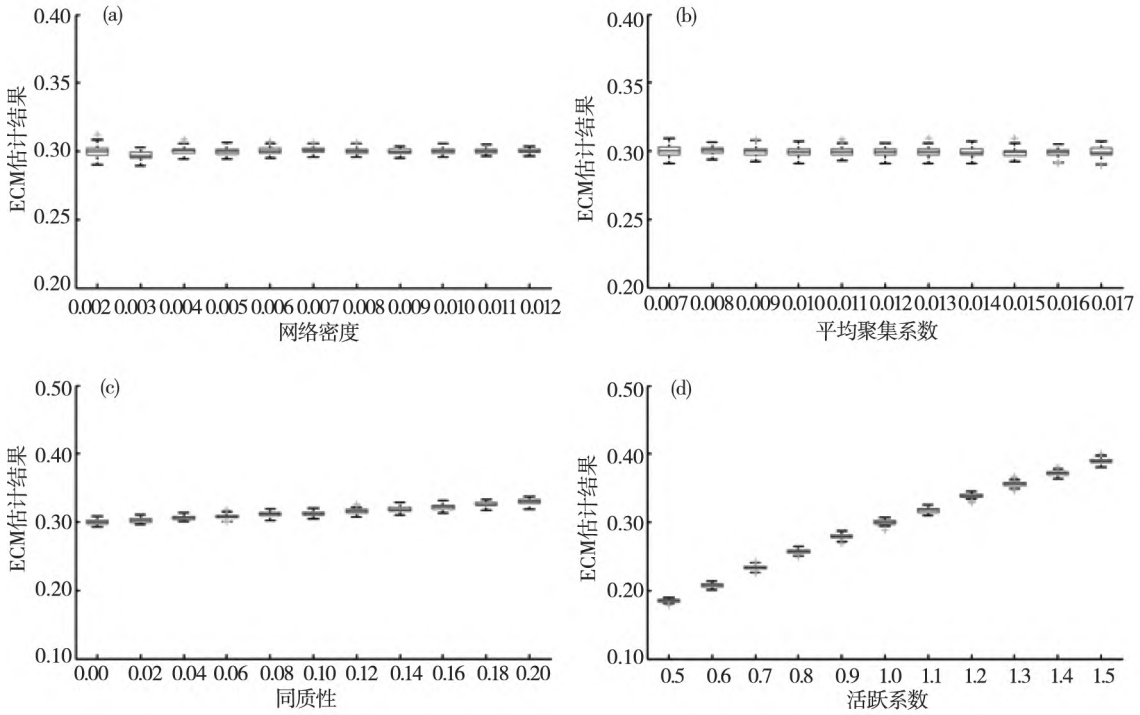

图 5 网络结构特征对 ECM 估计结果的影响,$P(\mathrm{A}) = 0.3$

Fig. 5 Impact of network structure on ECM estimation results, $P(\mathrm{A}) = 0.3$

3 实证研究

3.1 调查目的与问卷设计

在企业运营过程中,薪酬待遇、职场关系、员工职业规划、企业文化认同等成为不容回避和忽视的人力资源管理问题[6, 56]. 为了在真实人群的抽样调查中验证 ECM 方法的可行性和可靠性,选取广西某家具有四十余所连锁店面的企业,以该企业某区域店面的所有基层员工及管理层为调查对象,对职员的婚姻状况、学历、工龄、职位类型、税前年薪、薪酬满意度、职业成长、职场关系、离职

倾向等一般和隐私性问题进行了问卷调查,并通过基于中心网络信息的间接估计方法实现了对该群体的总体估计. 通过设计自报告和他报告问卷,对受访者发放包含了“敏感”问题及“非敏感”问题的匿名纸质问卷,并通过本文提出的间接估计方法实现了对该群体的总体估计,具体过程如图 6 所示.

问卷一共包含两个部分,第一部分主要通过“自我报告”的方式对个体自身信息进行采集,即直接询问受访者的自身信息,包括(i)婚姻状况(是否为已婚),(ii)学历(是否为高中以下),(iii)工龄(是否少于 5 年),(iv)职位类型(是否为店员),(v)税前年薪(是否在 4 万以内),(vi)薪酬满意度(对薪水是否满意),(vii)职业成长(上司是否栽培),(viii)职场关系(同事关系是否融洽),(ix)离职倾向(是否有离职意向). 第二部分则是通过“他报告”的方式对个体中心网络的信息进行采集,即询问受访者熟人或者朋友的总数量(对应为社交网络中的“度”),以及具有上述特征的朋友数量. 本文在问卷上对“朋友”进行了详细的定义,以避免因个体对“朋友”的理解误差对实验结果造成干扰(见附件 1). 另外,从该单位人事部获取受访者“非敏感”属性,即变量 i ~ iv 的真实值,与用户自报告信息和 ECM 方法的推断结果进行比对,检验用户自报告数据的准确性以及基于中心网络的推断方法的性能.

为了鼓励受访者参与、提高响应率和数据的可靠性,在调查实施时采取了一定的措施. 首先,受访者会被告知,数据分析人员不参与数据收集. 其次,问卷是匿名填写的,确保受访者的隐私. 第三,在问卷实施前对调查方法进行科普,进一步明确将只会收集受访者朋友的数量信息,并不会涉及到任何身份信息. 第四,受访者会被提供一张调查总体的名册,确保在问卷填写过程中计算出相对准确的社交网络规模.

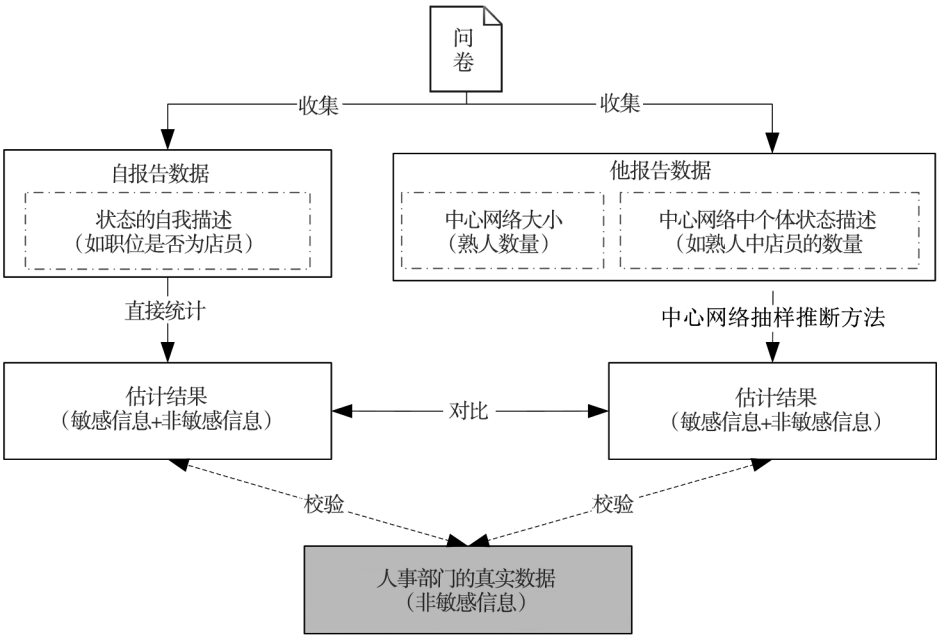

图 6　实证调查的基本流程

Fig. 6 The basic process of empirical questionnaire surveys

3.2　样本数据分析

数据采集之后,对存在明显错误的问卷进行筛选,然后利用收集到的自我报告信息和中心网络信息可以分别计算出含有各特征的总体比例,将“自我报告”直接统计方法和中心网络抽样推断方法对总体比例的估计结果进行对比. 一共发放了 52 份问卷,其中有效返回且无数据填写错误的共 42 份,回收率为 80.8%. 采集到的 42 个受访者的中心网络如图 7 所示. 计算发现,受访者中心网络的平均度(即网络平均大小)为 10.8,并且

其中一些受访者的“朋友”网络明显大于大多数受访者的“朋友”网络,符合网络科学的“朋友悖论”[57,58],即朋友的好友平均数总是大于自身的好友数量. 但是,受访者的度分布并没有呈现出异质或胖尾的特征,这与一般的实证社交网络的特性有所不同. 造成这样结果的主要原因很有可能是这个封闭群体的社会性很高,个体之间经常会产生联系.

3.3　预测结果及比较

统计采集到的自我报告信息,利用中心网络推断方法,可以得出问卷中这九个变量的总体比例估计(表 3). 对于“非敏感”特征而言,即问题(i)婚姻状况,(ii)学历,(iii)工龄,(iv)职位类型,可以看出自我报告估计结果和中心网络推断得出的估计结果与真实比例都十分接近. 其中,自我报告对四个“非敏感”特征总体比例的估计与真实比例的差别分别为 0.013、0.011、0.010 和 0.027,偏差非常小. 这说明对于个体“非敏感”信息而言,受访者往往会愿意提供自身真实信息. 利用中心网络信息对“学历”和“工龄”的推断结果虽然准确率略逊于自我报告的估计结果,但其与真实比例的拟合度仍十分可观,估计偏差分别为 0.028 和 0.013. 而对于利用中心网络信息对“婚姻状况”和“职位类型”的估计偏差分别为 0.013 和 0.007,它们则持平于或更优于自我报告的估计结果.

对于“敏感”特征(“税前年薪”、“薪酬满意度”、“职业成长”、“职场关系”和“离职倾向”),由于这些信息的真实比例事先无法得知,因此对自我报告估计结果和中心网络信息的推断结果进行一个综合比较. 可以发现,对于“职业成长”和“离职倾向”这两个属性,中心网络估计结果与自我报告统计结果基本一致,差别仅为 0.01. 而对于“薪酬满意度”,中心网络估计结果略高于基于自我报告的统计结果,对于“税前年薪”和“职场关系”,中心网络估计结果均低于基于自我报告的统计结果. 这在一定程度上说明,即使受访者隐私得到了充分的保证,但在调查过程中,面对有关经济状况或社会认可性的问题时(如“税前年薪”、“职场关系”),受访者的自我报告往往偏向“过表达”,而问及隐私性较强或不易于公开表露的特征时(如“薪酬满意度”),受访者偏向于“欠表达”.

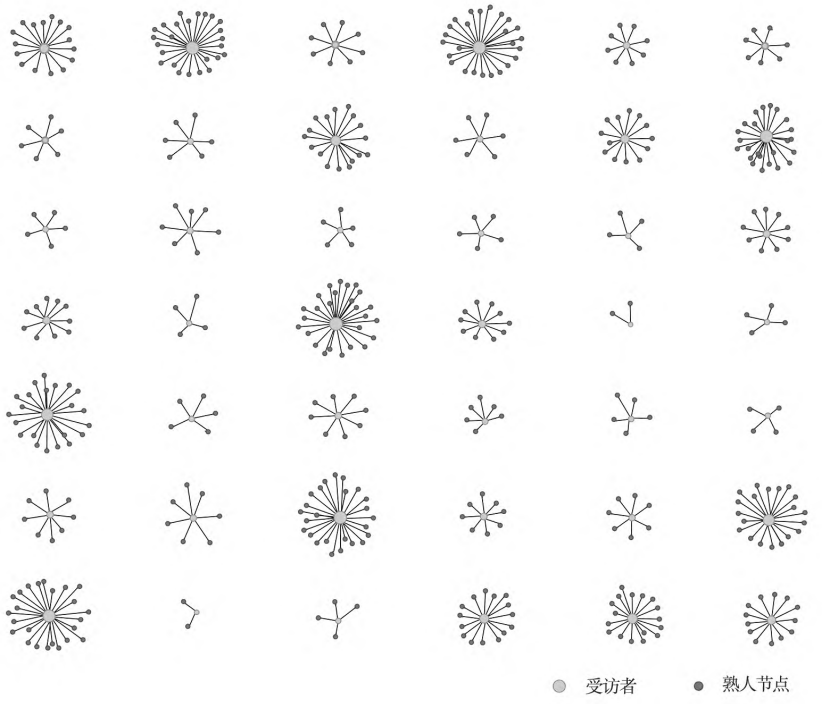

图 7　问卷收集的 42 个受访者的中心网络情况

Fig. 7 Ego-networks of 42 respondents collected by the questionnaire

表 3　实证调查总体真实值、自报告数据与 ECM 估计结果对比

Table 3 Comparison of the true value of empirical survey, self-reported data and ECM estimated results

属性	婚姻状况/%	学历/%	工龄/%	职位/%	税前年薪/%	薪酬满意度/%
真实比例	84.6	55.8	53.8	71.1	—	—
自我报告	83.3	54.7	54.8	73.8	66.7	78.6
ECM 估计	85.9	53.0	55.1	70.4	54.1	80.4
CI (95%)	[84.8,86.9]	[51.5,54.7]	[53.4,56.7]	[69.3,71.4]	[52.6,55.4]	[79.3,81.4]

属性	职业成长/%	职场关系/%	离职倾向/%
真实比例	—	—	—
自我报告	83.3	92.8	14.3
ECM 估计	83.4	86.8	14.2
CI (95%)	[82.4,84.4]	[85.9,87.8]	[13.1,15.2]

为了提高估计比例的可信度,本文还计算了中心网络方法的总体估计量的 95% 置信区间(95%CI). 总体来看,所有总体变量(各属性/特征的总体比例)的置信区间都相对较窄. 对于“非敏感”特征,虽然中心网络估计结果较自我报告结果的偏差稍大,除“工龄”外,总体真实比例并未落在 95% 置信区间内,但与真实比例的差距甚小. 其中在“婚姻状况”特征上,95% 置信区间与真实值的差距只有 0.002;在“学历”特征上,差距为 0.011. 而对于“敏感性”特征,虽然没有真实比例作为参考,但可以看出,中心网络推断方法的点估计都位于 95% 置信区间以内.

上述结果表明基于受访者中心网络信息的隐私规避调查与推断方法可以在实际抽样调查中有效应用,且能保证对总体比例进行可靠的推断和估计. 中心网络抽样过程中采集到的受访者“朋友”信息比以自我报告形式收集自身属性信息更加容易和准确. 同时,将中心网络与实际抽样调查过程相结合操作简单,只需要加入两个简单的小问题就能轻松实现. 另外,实证结果也印证了受访者对于个体正向信息会倾向于提供“过表达”的答案,而对于个体负面信息倾向于提供“欠表达”的答案.

4　结束语

在管理决策中,管理对象的真实状态往往因隐私、敏感等因素导致自我报告数据质量不高,数据采集难以大规模开展,进而难以掌握目标对象的真实情况,对管理政策的制定和实施等带来极大挑战. 与此同时,数字经济和数字治理格局的发展对数据隐私保护提出了更高要求. 在此背景下,隐私规避的统计推断方法成为保障调查数据可靠性和隐私性的重要技术途径. 针对敏感问题自我报告数据难获取、可靠性低且非概率抽样方法无法实现对总体的无偏估计等问题,本文提出了中心网络抽样统计推断(ECM)方法,实现了在不考虑中心节点(管理对象)属性信息的情况下,依靠中心节点邻居信息来间接推断总体参数,并在模型网络和真实网络上进行了可行性和有效性验证.

ECM 方法实施简单,可对调查对象进行随机采样或实施普查,除采集样本的自我报告数据外,同时采集每个样本关于其密切社交对象的报告数据,从而避免了管理对象自身因敏感原因等不愿提供数据或提供不真实数据的问题. 同时,在进行

相对苛刻的实证数据获取时(如抽样调查),该方法也优于一般的“自我报告”式统计方法,能够简单高效地适用于管理调查场景中涉及到调查成本高昂、调查问题内容因隐私或敏感性不可靠等问题的处理中. 另一方面,尽管 ECM 方法对总体的估计误差不易受网络密度、网络聚集性等结构特征的影响,在统计推断过程中应密切关注目标群体特征是否与个体网络存在交互作用,当个体的社交关系存在较强的同质性或不同属性节点的平均度存在较大差异时,即中心网络的组成和规模不独立于属性 A 时,应该谨慎地使用 ECM 的总体估计结果.

本文研究可以为有效收集敏感性问题的调查结果提供技术支撑,从而更好地服务社会、经济、公共卫生等领域的管理决策制定. 这种“间接”的依靠中心网络信息去解决问题的思路还可被应用于在线社交媒体分析、舆情商情分析、网络重构等方面. 后续工作可围绕上述方面开展进一步的实证研究,通过网络爬虫结合中心网络抽样推断方法来具体实施.

参 考 文 献:

[1]粟 芳, 邹奕格, 韩冬梅. 中国农村地区互联网金融普惠悖论的调查研究——基于上海财经大学 2017 年“千村调查”[J]. 管理科学学报, 2020, 23(9): 76 – 94.
Su Fang, Zou Yige, Han Dongmei. Investigation research for the paradox of Internet financial inclusion in Chinese rural area: Based on “A Thousand Villages Investigation” of 2017 by SUFE[J]. Journal of Management Sciences in China, 2020, 23(9): 76 – 94. (in Chinese)

[2]Shahzad U, Ahmad I, Al-Noor N H, et al. Särndal approach and separate type quantile robust regression type mean estimators for nonsensitive and sensitive variables in stratified random sampling[J]. Journal of Mathematics, 2022, 2022: 1430488.

[3]Rehm J, Kilian C, Rovira P, et al. The elusiveness of representativeness in general population surveys for alcohol[J]. Drug and Alcohol Review, 2021, 40(2): 161 – 165.

[4]Hart E, VanEpps E M, Schweitzer M E. The (better than expected) consequences of asking sensitive questions[J]. Organizational Behavior and Human Decision Processes, 2021, 162: 136 – 154.

[5]Zahl-Thanem A, Burton R J F, Vik J. Should we use email for farm surveys? A comparative study of email and postal survey response rate and non-response bias[J]. Journal of Rural Studies, 2021, 87: 352 – 360.

[6]Kwak D H A, Ma X, Kim S. When does social desirability become a problem? Detection and reduction of social desirability bias in information systems research[J]. Information & Management, 2021, 58(7): 103500.

[7]陈国青, 任 明, 卫 强, 等. 数智赋能: 信息系统研究的新跃迁[J]. 管理世界, 2022, 38(1): 180 – 195.
Chen Guoqing, Ren Ming, Wei Qiang, et al. Data-intelligence empowerment: A new leap of information systems research[J]. Management World, 2022, 38(1): 180 – 195. (in Chinese)

[8]陈国青, 张 瑾, 王 聪, 等. “大数据 – 小数据”问题: 以小见大的洞察[J]. 管理世界, 2021, 37(2): 203 – 213 + 14.
Chen Guoqing, Zhang Jin, Wang Cong, et al. The“Big Data-Small Data”problem: Insights for the big through the small[J]. Management World, 2021, 37(2): 203 – 213 + 14. (in Chinese)

[9]陈松蹊, 毛晓军, 王 聪. 大数据情境下的数据完备化: 挑战与对策[J]. 管理世界, 2022, 38(1): 196 – 206.
Chen Songxi, Mao Xiaojun, Wang Cong. Missing data completion in the big data era: Challenges and solutions[J]. Manage-

ment World, 2022, 38(1): 196 - 206. (in Chinese)

[10]Lu X. Respondent-driven Sampling: Theory, Limitations & Improvements[D]. Stockholm: Karolinska Institutet, 2013.

[11]陈　收, 蒲　石, 方　颖, 等. 数字经济的新规律[J]. 管理科学学报, 2021, 24(8): 36 - 47.
Chen Shou, Pu Shi, Fang Ying, et al. The new rules of digital economy[J]. Journal of Management Sciences in China, 2021, 24(8): 36 - 47. (in Chinese)

[12]Hedt B L, Pagano M. Health indicators: Eliminating bias from convenience sampling estimators[J]. Statistics in Medicine, 2011, 30(5): 560 - 568.

[13]Magnani R, Sabin K, Saidel T, et al. Review of sampling hard-to-reach and hidden populations for HIV surveillance[J]. Aids, 2005, 19: S67 - S72.

[14]Topp L, Barker B, Degenhardt L. The external validity of results derived from ecstasy users recruited using purposive sampling strategies[J]. Drug and Alcohol Dependence, 2004, 73(1): 33 - 40.

[15]Andrade C. The inconvenient truth about convenience and purposive samples[J]. Indian Journal of Psychological Medicine, 2021, 43(1): 86 - 88.

[16]Etikan I, Musa S A, Alkassim R S. Comparison of convenience sampling and purposive sampling[J]. American Journal of Theoretical and Applied Statistics, 2016, 5(1): 1 - 4.

[17]Wullenkord M C, Tröger J, Hamann K R S, et al. Anxiety and climate change: A validation of theclimate anxiety scale in a German-speaking quota sample and an investigation of psychological correlates[J]. Climatic Change, 2021, 168(3): 1 - 23.

[18]Dosek T. Snowball sampling and facebook: How social media can help access hard-to-reach populations[J]. PS: Political Science & Politics, 2021, 54(4): 651 - 655.

[19]De Brier N, Van Schuylenbergh J, Van Remoortel H, et al. Prevalence and associated risk factors of HIV infections in a representative transgender and non-binary population in Flanders and Brussels (Belgium): Protocol for a community-based, cross-sectional study using time-location sampling[J]. Plos One, 2022, 17(4): e0266078.

[20]Karon J M, Wejnert C. Statistical methods for the analysis of time: Location sampling data[J]. Journal of Urban Health, 2012, 89(3): 565 - 586.

[21]刘　雯. 分类敏感问题分层抽样调查的统计方法及应用[D]. 苏州: 苏州大学, 2012.
Liu Wen. Statistical Methods for Qualitative Sensitive Questions Survey in Stratified Sampling and its Application[D]. Suzhou: Soochow University, 2012. (in Chinese)

[22]Narjis G, Shabbir J. An improved two-stage randomized response model for estimating the proportion of sensitive attribute[J]. Sociological Methods & Research, 2021: 00491241211009950.

[23]Simons W R, Horvitz D G, Shah B V. The unrelated question ranmomized response model[J]. Proceedings in the Social Statistics Seetion, Ameriean Statistical Association, 1967: 65 - 72.

[24]Greenberg B G, Abul-Ela A L A, Simmons W R, et al. The unrelated question randomized response model: Theoretical framework[J]. Journal of the American Statistical Association, 1969, 64(326): 520 - 539.

[25]Mangat N S. An improved randomized response strategy[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1994, 56(1): 93 - 95.

[26]Kim J M, Warde W D. A mixed randomized response model[J]. Journal of Statistical Planning and Inference, 2005, 133(1): 211 – 221.

[27]刘 寅, 田国梁. 敏感性问题的抽样调查中非随机化响应技术的新进展[J]. 应用概率统计, 2019, 35(2): 200 – 217.
Liu Yin, Tian Guoliang. Advances of the non-randomized response techniques in sample surveys with sensitive questions[J]. Chinese Journal of Applied Probability and Statistics, 2019, 35(2): 200 – 217. (in Chinese)

[28]Yu J W, Tian G L, Tang M L. Two new models for survey sampling with sensitive characteristic: Design and analysis[J]. Metrika, 2008, 67(3): 251 – 263.

[29]Groenitz H. A new privacy-protecting survey design for multichotomous sensitive variables[J]. Metrika, 2014, 77(2): 211 – 224.

[30]Tian G L. A new non-randomized response model: The parallel model[J]. Statistica Neerlandica, 2014, 68(4): 293 – 323.

[31]江 浩. 基于负调查的敏感信息收集方法及其应用研究[D]. 合肥: 中国科学技术大学, 2019.
Jiang Hao. Research on Sensitive Information Collection Methods Based on the Negative Survey and Their Applications[D]. Hefei: University of Science and Technology of China, 2019. (in Chinese)

[32]方 舒. 基于负调查的保护位置隐私的评论模型[D]. 武汉: 武汉理工大学, 2019.
Fang Shu. A Review Model for Protecting Location Privacy Based on Negative Surveys[D]. Wuhan: Wuhan University of Technology, 2019. (in Chinese)

[33]刘 洋. 关于敏感性问题调查技术的研究[D]. 北京: 首都经济贸易大学, 2014.
Liu Yang. Research on Survey Techniques for Sensitive Questions[D]. Beijing: Capital University of Economics and Business, 2014. (in Chinese)

[34]Spira C, Raveloarison R, Cournarie M, et al. Assessing the prevalence of protected species consumption by rural communities in Makira Natural Park, Madagascar, through the unmatched count technique[J]. Conservation Science and Practice, 2021, 3(7): e441.

[35]Droitcour J A, Larson E M. An innovative technique for asking sensitive questions: The three-card method[J]. Bulletin of Sociological Methodology, 2002, 75(1): 5 – 23.

[36]Yan T. Consequences of asking sensitive questions in surveys[J]. Annual Review of Statistics and Its Application, 2021, 8: 109 – 127.

[37]Nuno A, John F A V S. How to ask sensitive questions in conservation: A review of specialized questioning techniques[J]. Biological Conservation, 2015, 189: 5 – 15.

[38]Fisher J C D, Flannery T J. Designing randomized response surveys to support honest answers to stigmatizing questions[J]. Review of Economic Design, 2022: 1 – 33.

[39]Raifman S, DeVost M A, Digitale J C, et al. Respondent-driven sampling: A sampling method for hard-to-reach populations and beyond[J]. Current Epidemiology Reports, 2022: 1 – 10.

[40]Lu X, Bengtsson L, Britton T, et al. The sensitivity of respondent-driven sampling[J]. Journal of the Royal Statistical Society: Series A (Statistics in Society), 2012, 175(1): 191 – 216.

[41]Salganik M J, Heckathorn D D. Sampling and estimation in hidden populations using respondent-driven sampling[J]. Socio-

logical Methodology, 2004, 34(1): 193 – 240.

[42] Fishburne P M. Survey Techniques for Studying Threatening Topics: A Case Study on the Use of Heroin[D]. New York: New York University, 1980.

[43] Rossier C. The anonymous third party reporting method[J]. Methodologies for Estimating Abortion Incidence and Abortion-related Morbidity: A Review, 2010: 99 – 106.

[44] Lu X. Linked ego networks: Improving estimate reliability and validity with respondent-driven sampling[J]. Social Networks, 2013, 35(4): 669 – 685.

[45] Krivitsky P N, Morris M. Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US[J]. The Annals of Applied Statistics, 2017, 11(1): 427.

[46] Krivitsky P N, Morris M, Bojanowski M. Impact of survey design on estimation of exponential-family random graph models from egocentrically-sampled data[J]. Social Networks, 2022, 69: 22 – 34.

[47] Baraff A J, McCormick T H, Raftery A E. Estimating uncertainty in respondent-driven sampling using a tree Bootstrap method [J]. Proceedings of the National Academy of Sciences, 2016, 113(51): 14668 – 14673.

[48] 王宗润, 汪武超, 陈晓红, 等. 基于 BS 抽样与分段定义损失强度操作风险度量[J]. 管理科学学报, 2012, 15(12): 58 – 69.

Wang Zongrun, Wang Wuchao, Chen Xiaohong, et al. Application of LDA based on Bootstrap sampling and piecewise-defined severity distribution in operational risk measurement[J]. Journal of Management Sciences in China, 2012, 15(12): 58 – 69. (in Chinese)

[49] Liu Y, Sanhedrai H, Dong G G, et al. Efficient network immunization under limited knowledge[J]. National Science Review, 2021, 8(1): nwaa229.

[50] Berx J. Hierarchical deposition and scale-free networks: A visibility algorithm approach[J]. Physical Review E, 2022, 106 (6): 064305.

[51] Kojaku S, Hébert-Dufresne L, Mones E, et al. The effectiveness of backward contact tracing in networks[J]. Nature Physics, 2021, 17(5): 652 – 658.

[52] Toivonen R, Kovanen L, Kivelä M, et al. A comparative study of social network models: Network evolution models and nodal attribute models[J]. Social Networks, 2009, 31(4): 240 – 254.

[53] Huang G, Cai M, Lu X. Inferring opinions and behavioral characteristics of gay men with large scale multilingual text from blued[J]. International Journal of Environmental Research and Public Health, 2019, 16(19): 3597.

[54] Cai M, Huang G, Kretzschmar M E, et al. Extremely low reciprocity and strong homophily in the world largest MSM social network[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(3): 2279 – 2287.

[55] Gile K J, Handcock M S. Respondent-driven sampling: An assessment of current methodology[J]. Sociological Methodology, 2010, 40(1): 285 – 327.

[56] 王　磊. 多分类敏感问题 RRT 模型下整群抽样调查的统计方法及其效度信度模拟评价[D]. 苏州: 苏州大学, 2012.

Wang Lei. The RRT Model for Inquiring Multichotomous Sensitive Questions Under Cluster Sampling and Assessment of Validity and Reliability Through Computer Simulation[D]. Suzhou: Soochow University, 2012. (in Chinese)

[57] Bong K W, Utreras-Alarcón A, Ghafari F, et al. A strong no-go theorem on the Wigner's friend paradox[J]. Nature Physics, 2020, 16(12): 1199 – 1205.

[58] Cantwell G T, Kirkley A, Newman M E J. The friendship paradox in real and model networks[J]. Journal of Complex Networks, 2021, 9(2): cnab011.

Network survey and indirect inference with privacy avoidance

LU Xin[1], *LIU Chu-chu*[1,2], *CAI Meng-si*[1], *CHEN Sa-ran*[3]

1. College of Systems Engineering, National University of Defense Technology, Changsha 410073, China;
2. School of Computing, National University of Singapore, Singapore 117417, Singapore;
3. State Key Laboratory on Blind Signal Processing, Chengdu 610041, China

Abstract: In the process of management decision-making, the real state of the management objects is often subject to low-quality self-reported data or large sampling biases due to concerns regarding privacy or sensitivity, which makes it difficult to know the real situation of the target objects. To solve this problem, yet to meet the data privacy protection demand in the era of the digital economy, this paper develops a data collection method based on social network indirect reports, and designs an ego-centric sampling method (ECM) based on indirectly reported sample data on the basis of network sampling and statistical inference theory. This method is simple to implement such that it can be deployed by either randomly sampling the survey objects or conducting a census. In addition to collecting the self-reported data of the samples, it also collects data of each sample's close social contacts, so as to avoid the problem that people are unwilling to provide some data or provide untrue data due to sensitive privacry reasons. The proposed method can achieve a high-precision estimation of the population, and it can realize the interactive verification of self-reported data and cross-reported data. The research method is fully validated on the online social network of a hard-to-reach population with up to 556 627 active users. The sampling experiment shows that the estimation bias of ECM is less than 3% for the average number of friends and overall characteristics of the whole network. Furthermore, this paper conducts an empirical study by implementing a questionnaire survey on general and sensitive variables for employees in an enterprise, and derives the overall estimation of the study objects through the indirect estimation method; the results verify the practicality and effectiveness of ECM.

Key words: privacy; data authenticity; network sampling; indirect estimation; statistical inference