Journal of Complex Networks (2021) **6**, Advance Access Publication on 22 October 2021
doi: 10.1093/comnet/cnab034

Indirect inference of sensitive variables with peer network survey

SARAN CHEN
School of Mathematics and Big Data, Foshan University, 528000 Foshan, China
XIN LU†
College of Systems Engineering, National University of Defense Technology, 410073 Changsha, China
Department of Global Public Health, Karolinska Institutet, 17177 Stockholm, Sweden
†Corresponding author. Email: xin_lyu@sina.com
FREDRIK LILJEROS
Department of Sociology, Stockholm University, 10691 Stockholm, Sweden
Department of Global Public Health, Karolinska Institutet, 17177 Stockholm, Sweden
ZHONGWEI JIA
School of Public Health, Peking University, 100191 Beijing, China
Center for Intelligent Public Health, Institute for Artificial Intelligence, Peking University, 100191 Beijing, China
Center for Drug Abuse Control and Prevention, National Institute of Health Data Science, Peking University, 100191 Beijing, China
AND
LUIS EC ROCHA
Department of Economics & Department of Physics and Astronomy, Ghent University, B-9000 Ghent, Belgium

Edited by: Dr. Xiang Li

[Received on 28 December 2020; editorial decision on 13 September 2021; accepted 14 September 2021]

Misreporting is a common source of bias in population surveys involving sensitive topics such as sexual behaviours, abortion or criminal activity. To protect their privacy due to stigmatized or illegal behaviour, respondents tend to avoid fully disclosure of personal information deemed sensitive. This attitude however may compromise the results of survey studies. To circumvent this limitation, this article proposes a novel ego-centric sampling method (ECM) based on the respondent's peer networks to make indirect inferences on sensitive traits anonymously. Other than asking the respondents to report directly on their own behaviour, ECM takes into account the knowledge the respondents have about their social contacts in the target population. By using various scenarios and sensitive analysis on model and real populations, we show the high performance, that is low biases, that can be achieved using our method and the novel estimator. The method is also applied on a real-world survey to study traits of college students. This real-world exercise illustrates that the method is easy-to-implement, requiring few amendments to standard sampling protocols, and provides a high level of confidence on privacy among respondents. The exercise revealed that students tend to under-report their own sensitive and stigmatized traits, such as their sexual orientation. Little or no difference was observed in reporting non-sensitive traits. Altogether, our results indicate that ECM is a promising method able to encourage survey participation and reduce bias due to misreporting of sensitive traits through indirect and anonymous data collection.

Keywords: indirect inference; sensitive variables; peer network survey

1. Introduction

Surveys are widely used in public health, psychology, business and sociology to collect data about populations and to develop quantitative and qualitative understanding of the population characteristics [1]. The population proportions of traits or characteristics can be obtained by direct statistical inference techniques in which respondents self-report information and are sampled via standard sampling methods, as for example, simple random sampling, stratified sampling, clustering sampling [2]. When conducting a survey on sensitive topics such as sexual behaviour, abortion, drug use or illicit activities, a key concern is whether the respondents report their own characteristics accurately. Misreporting may occur in all types of surveys and is considered a major source of bias [3]. Due to the stigma associated to certain behaviours and fear of legal consequences in case of illicit activity, survey respondents tend to under-report behaviours associated to undesirable or illicit activities, and over-report desirable and socially, or legally, acceptable characteristics [3, 4]. For example, populations that are at higher risk of sexually transmitted infections, including sex workers, injecting drug users or men-who-have-sex-with-men tend to not respond questions about their stigmatized or illegal behaviours and consequently hide their health conditions [5–7]. Similarly, clients of sex workers or drug dealers may be very challenging to identify even using advanced sampling methods for hard-to-reach populations. In other contexts, such as voting, researchers also found that more than 20% of non-voters lied that they had participate in voting by comparing self-reported data with voting records [8]. Survey participants may also lie purposefully if they do not trust the study or if the survey is judged excessively complicated [9].

To reduce misreporting and increase the accuracy of estimates on sensitive topics, several methods have been developed and tested in different contexts. One set of methods relies on encouraging respondents to provide reliable responses either by increasing privacy or by using psychological games. In the self-administration approach, for example, respondents are instructed to interact directly with the questionnaire without mediation by an interviewer [10, 11]. The same data collection may be also conducted in private or in anonymous settings to further increase privacy [12, 13]. The bogus pipeline technique, on the other hand, encourages accurate reporting by telling respondents that they are being monitored by machines (e.g. lie detectors) or biological assays (e.g. saliva samples), and thus false reports can be automatically detected; in reality however there is not automatic detection of false reports [14, 15]. The indirect questioning methods, also called specialized question methods, specifically developed for surveys on sensitive topics, involve more complex procedures to prevent the interviewers from learning the answers of respondents on sensitive topics. Popular indirect methods include the randomized response technique (RRT) [16], item count technique [17], the three-card method [18] and their variants. For example, RRT designs a pair of questions for each respondent in which one question is sensitive and the other is unrelated and innocuous; a randomizing machine is used to decide whether a respondent will answer the sensitive question but the actual question is only known by the respondent. Although such methods are more effective on increasing the respondents willingness to report truthfully [19, 20], there are still limitations in the implementation. First, the complex design (e.g. the use of randomizing machines or multiple cards) may confuse the respondents and increase the difficulty of the implementation in real settings. Some respondents may be unable to understand how the survey works and hence may not feel comfortable enough to trust the study. Second, some methods generate high variance, for example, the variance of RRT is at least four times higher than the variance of conventional estimates in many cases [21].

Research shows that people are more willing to talk about their friends' sensitive characteristics (e.g. drug use [22] and abortion [23]) than their own, and that they can provide reliable responses about the specific number of these friends [21–23]. We exploit this social behaviour to develop a sampling method, hereafter called ego-centric sampling method (ECM), to generate estimates for sensitive variables using surveys about the respondents peer networks. Therefore, instead of asking respondents to report their

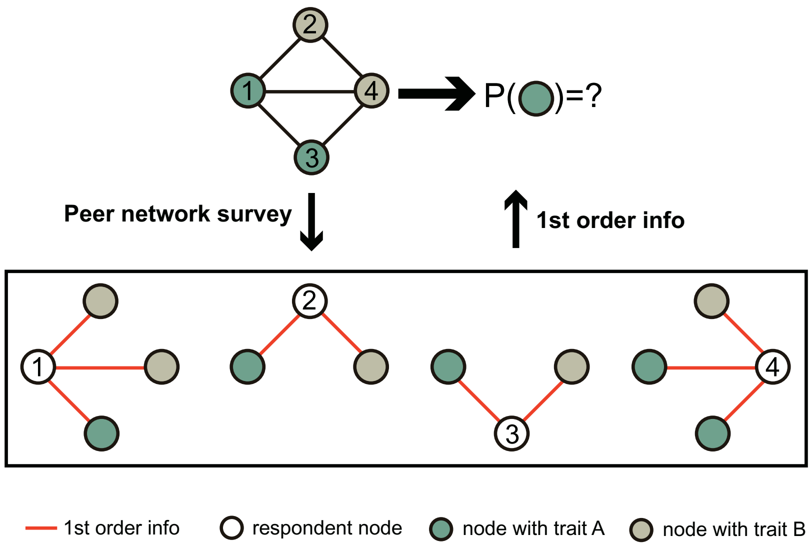

FIG. 1. Ego-centric sampling. The diagram illustrates the basic mechanisms of the ECM. ECM collects information by using the peer network survey and then makes inference taking into account the social contacts.

own traits, ECM collects ego-centric information anonymously and then makes indirect inference without the self-reported information but taking into account the social network effect for respondents and their peers (see Fig. 1). The method consists in first conducting the so-called peer network survey, which has been successfully applied in the general social survey [24] and in the study of hidden populations [25]. Different from traditional sampling, our method neither requires respondents (i.e. egos) to provide personal information on sensitive questions nor name their social contacts (i.e. friends or close acquaintances). The survey collects only the number of contacts (also called social degree) and the number of those contacts with trait or characteristic A (remaining contacts thus have the mutually exclusive trait or characteristic B). This procedure can be typically implemented with two simple questions: (1) how many friends or close acquaintances do you have in the target group of the survey? and (2) how many of your friends or close acquaintances have trait/characteristic A? Once this information is collected, ECM infers population-level statistics using the underlying social networks between people in the population.

2. Results

The performance of the ECM is tested with model populations in which the social structures and other population traits can be controlled, and in a real-world study about sexual behaviour, demographics and economics of college students in a Chinese university.

2.1 *Performance on model populations*

We first study the ECM performance considering three models of underlying social contacts between individuals: (i) the Erdős–Rényi (ER) network [26]; (ii) the Barabási–Albert (BA) network [27]; and (iii) the KOSKK social network [28, 29]. The ER network model is a reference model in which links are distributed uniformly among pairs of individuals (i.e. the nodes), and thus the number of social contacts has a characteristic value almost symmetrically distributed around the mean value. The BA network model, on the other hand, assumes a preferential attachment growing process in which individuals with more contacts tend to attract more new contacts. This gives rise to few individuals having much more contacts than the mean and many individuals with just a few contacts. Both models do not reproduce realistically social contacts but are helpful to understand the impact of the network structure on the sampling process. The KOSKK model reproduces social network structures observed in real-life and is used as a standard population model for testing processes in simulated, yet realistic, social networks (see Supplementary Information). Each model population contains $N = 10,000$ individuals (or nodes) and $E = 100,000$ social contacts (or edges). For each network, a trait A is assigned to a given fraction $P(A)$ of randomly selected individuals.

We begin the analysis by considering the census of the whole population N. This type of survey is often used for closed populations, such as schools, enterprises, compulsory rehabilitation centres, etc. Although we can conduct a census on each individual, some people may not be willing to answer sensitive questions anyways. Therefore, we apply the ECM to estimate the proportion $\hat{P}(A)$ of individuals with trait A (see Section 4) using the whole population (i.e. all egos in the network are involved). The results show that the bias BS (i.e. the average of the estimates $\hat{P}_m(A)$ minus the true proportion, $BS = \sum_{m=1}^{M}(|\hat{P}_m(A) - P(A)|)/M)$ where M is the number of simulations) is less than 0.003 in all cases (Table 1). The maximum bias of 0.003 is observed when the true population proportion $P(A) = 0.1$ in the BA network model, and the minimum bias is close to 0 for $P(A) = 0.2$ in the BA network, and $P(A) =$ 0.1 and 0.3 in the KOSKK network. For the remaining models, the biases of ECM are all less than 0.001. The results suggest that the high variability of individual contacts, captured by the BA network model, may have some impact on the estimates. Nevertheless, ECM has a relatively superior performance in the KOSKK population model that reproduces more realistically heterogeneous social network structures such as the degree distribution, social clustering and degree-correlations.

In this initial analysis, we assume that respondents are knowledgeable about the traits of all their contacts. In practice, it is more likely that respondents are unaware of sensitive information about all of their contacts. Respondents may also be more confident to inform about just a few closer contacts such as closer friends. We simulate this scenario of partial information by assuming that each ego randomly select either 3 or 5 of its contacts and inform only their traits during data collection. Table 1 shows that ECM tolerates variations in the network structure and produces biases smaller than 0.004 in all simulated scenarios. The results suggest that partial information, which is easier to collect, may be sufficient to generate satisfactory estimates based on peer networks.

When the census is not feasible to implement in the target population, for example, because they are hard-to-reach or too large, we can also use ECM to infer the $P(A)$ by using data collected via simple random sampling. We simulate this scenario by uniformly selecting 10% of the individuals (or nodes) and collecting their corresponding alters' information for the statistical estimation. Figure 2(A) shows the distribution of bias for the estimates of each trait on different population models. The estimates are centred around the true population values. For all the population models, the distribution of biases becomes wider for increasing values of the true proportion of the study variable ($P(A)$ from 0.1 to 0.4) in the population. This reflects the fact that more than one respondent may be giving information about the very same contact. The average biases for all the simulated model configurations are small. The average biases (over $M = 500$ random realizations of the sampling process) are all less than 0.005. In seven of the simulated configurations, i.e. when $P(A)$ is 0.1 and 0.3 in the ER network, 0.2 and 0.3 in the BA network and 0.1, 0.3 and 0.4 in the KOSKK network, the average biases are bellow 0.001. The variation of different simulations for each trait is also small. The maximum bias of all these model configurations is less than 0.015. In particular, when $P(A)$ is 0.1 and 0.2 in the ER network, 0.1 in the BA network, and 0.1 in the KOSKK network, the maximum bias is always below 0.01.

TABLE 1 *The true and the sample estimate proportions of people with trait or property A. Simulations are repeated $M = 500$ times for each model configuration. The results are averaged over 500 times.*

Network	$P(A)$ (%)	$\hat{P}(A)$ (%)	$\hat{P}_{n3}(A)$ (%)	$\hat{P}_{n5}(A)$ (%)
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

$P(A)$ denotes the true population proportion.
$\hat{P}(A)$ denotes the estimates from the census of each respondent.
$\hat{P}_{n3}(A)$ and $\hat{P}_{n5}(A)$ denote the estimates from reports on 3 and 5 randomly selected contacts of each respondent.

2.2 *The activity ratio*

In real-life populations, individuals with different traits may dynamically affect the formation of their personal social networks [30, 31], and there might be systematic differences in the average degree of individuals with different traits. Such characteristic of social networks can be quantified by the activity ratio (AR) which is the ratio of the mean degree for all individuals (N_A) with trait A to those individuals (N_B) with trait B in the network, $AR = \langle k\rangle^A/\langle k\rangle^B$, where $\langle k\rangle^A = \sum_{i\in A} k_i/N_A$ and $\langle k\rangle^B = \sum_{i\in B} k_i/N_B$. If $AR = 1$, the traits of the respondent is independent to the size of the respondent's personal network.

In the previous section, the traits were randomly assigned to individuals and thus $AR = 1$. In this section, however, we study the effect of AR by varying it from 0.5 to 1.5 in the KOSKK network, by fixing the true population proportion at $P(A) = 0.3$ (see Section 4). In line to previous results, the estimates are distributed around the true value and the average has negligible bias if AR $= 1$ (Fig. 2(B)). On the other hand, if AR is larger or smaller than 1, the estimates deviate from the true value. If AR < 1, the estimates tend to be smaller than the true value whereas if AR > 1, the estimates tend to be larger than the true value. Such biases increase with the increasing difference between the value of AR and 1. The biases of the estimates are small in a wide range of AR values: if $0.6 <$ AR < 1.5, the biases in all model configurations are less than 0.05. When AR = 0.6 or AR = 1.5, the average bias of the estimates is 0.03 and 0.028, respectively. The largest bias is observed for AR = 0.5 and the average bias for all configurations is 0.039. The results indicate that ECM is affected by the activity ratio. However, the biases introduced by varying activity ratios are relatively small and remain within acceptable intervals. For example, the average of the relative bias when AR varies from 0.7 to 1.5 is only 10%.

2.3 *Performance on a real-world social network*

We implement the ECM on an anonymized online men who have sex with men (MSM) social network called Blued, which is the world's largest MSM dating app [32]. We collect data only on users and

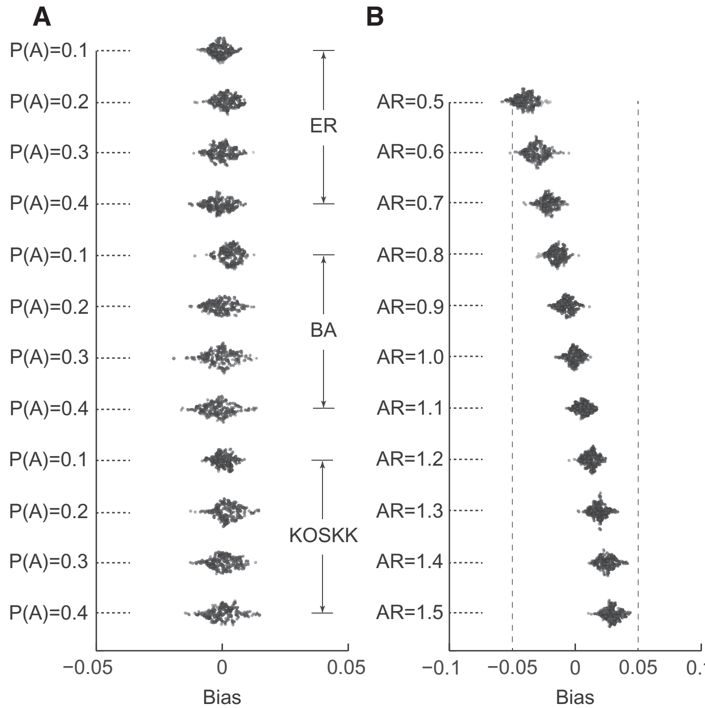

FIG. 2. Distribution of biases for each model population. The distribution of biases for the estimates of the proportion of people $\hat{P}(A)$ with property A on (A) model populations (ER, BA and KOSKK social networks) with varying proportions $P(A)$, and (B) model population (KOSKK social network) with varying activity ratios (AR). The sampling size of the simple random sampling corresponds to 10% of the full population. Simulations are repeated 500 times for each model configuration.

friendship relations (mutual following) within the giant connected component (GCC), yielding a network of $N = 556{,}627$ nodes and $E = 16{,}963{,}498$ undirected links. The degree distribution is skewed with 80% of users connecting to less than 28 other users (see Fig. 3(A)). Four traits of the users have been used for the estimation exercise: (i) *age* (less than 30 years old), (ii) *sex roles* (“Tops”), (iii) *country* (China) and (iv) *continent* (Asia).

The estimated proportions of users for each of the four traits shows the high agreement of ECM estimation and the true values, with small biases for each trait, that is, 0.033 for *age*, 0.022 for *sex roles*, 0.031 for *country* and 0.030 for *continent* (Table 2). The activity ratios of the traits are larger than 1, which may contribute to these small biases observed in the estimates according to the analysis in Section 2.2. The results suggest that ECM can provide accurate population estimates on population proportions with small biases on real social networks.

2.4 *Reporting error in the real-world network*

In the above analysis, we assume that all respondents (user nodes) can report their social contacts accurately. In the real life, the respondents may mis-estimate their personal network size. To evaluate this potential bias, we simulate ECM on the MSM social network with degree reporting error $p_A^{\text{miss}} \in [0, 0.2]$ and $p_B^{\text{miss}} \in [0, 0.2]$, that is, a maximum of 20% social contacts with trait A or B (mutual exclusive trait

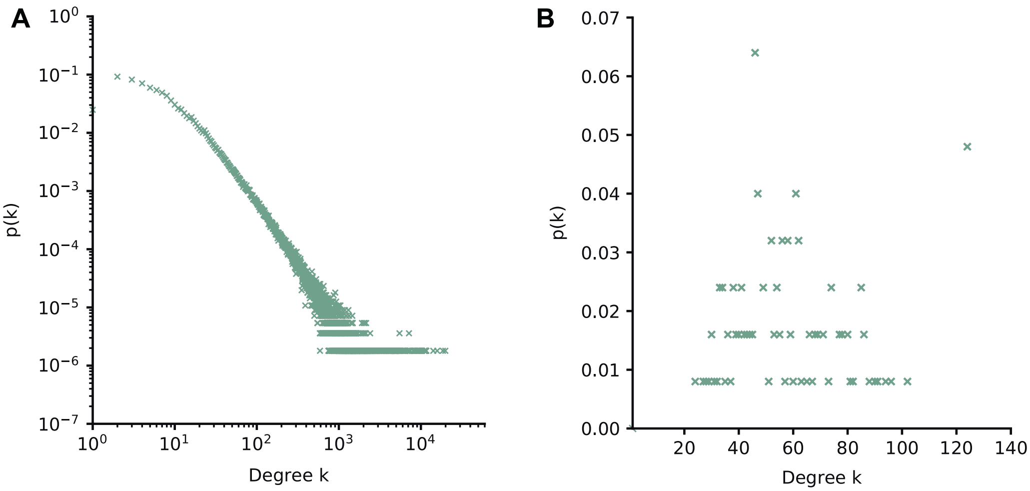

FIG. 3. Degree distributions of (A) the MSM social network and (B) respondents in real-world survey. The average degrees are 30.5 for MSM social network and 59.2 for respondents in real-world survey.

TABLE 2 *The population value and ECM estimates of four traits on the MSM social network*

(%)	*Age*	*Sex roles*	*Country*	*Continent*
Tue value	47.7	19.5	61.4	79.3
Estimate	51.0	21.7	64.5	82.3
Activity ratio	1.08	1.03	1.12	1.09

of A) may be unidentified in the respondents personal networks. Figure 4 shows that the biases of ECM range within [0.00, 0.075] even with miscounting 20% of social contacts. The highest biases are 0.072 for *age*, 0.051 for *sex roles*, 0.064 for *country* and 0.046 for *continent* with miscounting 20% of social contacts with trait A.

Another reporting error which may occur in the implementation of ECM, is that the respondents inaccurately report the traits of their social contacts. We evaluate the performance of ECM by varying trait reporting error p_A^{error} and p_B^{error} from 0 (i.e. the number of social contacts with trait A or B is reported correctly) to 0.2 (i.e. 20% of these are misclassified). Figure 5 shows that ECM is relatively more sensitive to trait reporting error in comparison to its robustness to degree reporting error (Fig. 4), though the biases become large only when there is a systematic difference in the reporting error regarding trait, for example, the bias is large only when one type of the trait is significantly misclassified. When both traits are misclassified, their effect on the bias will compensate each other and the final bias become small to moderate. In the worst case scenarios, the biases of ECM on four traits are 0.11 for *age*, 0.15 for *sex roles*, 0.08 for *country* and 0.12 for *continent*. These results suggest that ECM provides the estimates of population proportions with low and acceptable biases even in case of the degree and trait reporting errors.

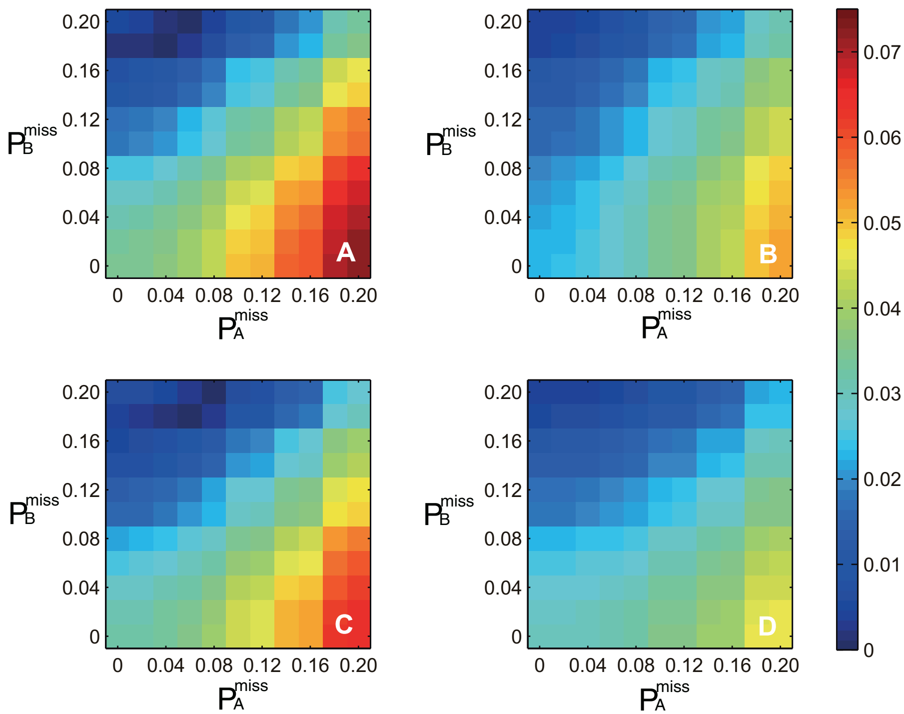

FIG. 4. Bias of ECM on four traits of the MSM social network with degree reporting error. (A) *age*, (B) *sex roles*, (C) *country* and (D) *continent*. p_A^{miss} and p_B^{miss} give the fraction of miscounted contacts, respectively, for traits A and B.

2.5 *Real-world survey*

To corroborate the applicability of the proposed sampling method (ECM) in real settings, we conduct a real-world survey about demographics and sexual behaviour of college students (see Section 4). To compare the performance of ECM with standard sampling, self-reported traits of respondents and traits of their contacts are collected. An anonymous paper questionnaire is sent to each student in the whole grade. Out of 155 invitations, only 125 students returned valid responses. Figure 6 shows the ego-networks of each of the 125 students. Some students have a relatively larger number of contacts than the majority, nevertheless, the degree distribution of the sample (see Fig. 3(B)) reveals that the distribution of degrees is not heterogeneous or skewed, most likely because this is a closed population with high sociality.

Table 3 shows the self-reported and estimated proportions of students for each of the seven sensitive and non-sensitive traits (see Section 4). For the non-sensitive traits, the true proportion is calculated using authorized and anonymous data obtained from the central university registers. The results indicate relatively small biases when estimating non-sensitive traits such as *age*, *gender* and *province* of origin. In the case of the self-reported data, the biases are 0.003 (*age*), 0.012 (*gender*) and 0.03 (*province*). These results support our hypothesis that people more openly tell the truth when the questions involve non-sensitive topics. Although the accuracy of ECM estimates for non-sensitive traits are not as good as for the self-reported case, the estimates are sufficiently close to the true proportions with relatively small biases, that is, 0.009 (*age*), 0.017 (*gender*) and 0.035 (*province*). The activity ratios of each trait vary between 0.95 and 1.35 (see Supplementary Information). These variations may contribute to generate the

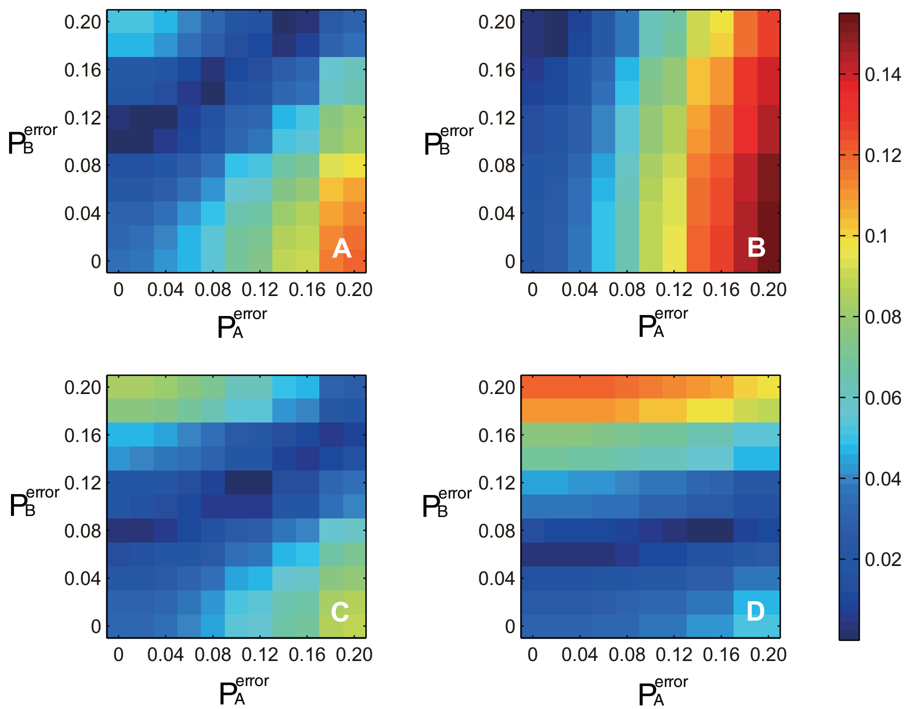

FIG. 5. Bias of ECM on four traits of the MSM social network with trait reporting error. (A) *age*, (B) *sex roles*, (C) *country* and (D) *continent*. p_A^{error} and p_B^{error} give the fraction of misclassified contacts, respectively, for traits A and B.

small bias observed in the estimators. Overall, the results indicate that in our real-world study, the ECM provides reliable population estimates without the need of using the respondents' own information.

Since we do not have the true population values for the four sensitive traits, we are limited to compare ECM with self-reported estimates. In contrast to non-sensitive traits, in which ECM estimates are all smaller than self-reported estimates, the estimation of sensitive traits are larger in ECM than in the self-reported case. In fact, the only exception is the estimation of *living expenses*, in which self-reports result on larger estimated proportions. We have observed small differences when separating the sample in man and woman (see Supplementary Information). These results suggest that the traits that may be socially embarrassing or stigmatized, such as *virginity*, *homosexuality* (sexual orientation) or *having sex with strangers*, may be under-reported, even if privacy and confidentiality of respondents are in place, as is the case for our method. Since we do not have the true value for the sensitive information, it may well be the case that in reality, ego's under-estimate living expenses while over-estimating traits related to sexual behaviour. A definitive answer is outside of the scope of our sampling method which aims to provide accurate estimations and not necessarily identify the source of potential biases.

To increase the confidence in our estimates, we also estimate the 95% confidence intervals (95% CI) for ECM using a bootstrap method [33, 34] (see details in Supplementary Information). The confidence intervals are relatively narrow for all variables. For the non-sensitive traits, due to the bias estimates of the corresponding variables, the true values are not contained in the 95% CI, but the difference between the CI and the true value is small, 0.003 (*age*), 0.008 (*gender*) and 0.025 (*province*). For the sensitive traits, the most critical for our purposes, the point estimates are all within the estimated 95% CIs.

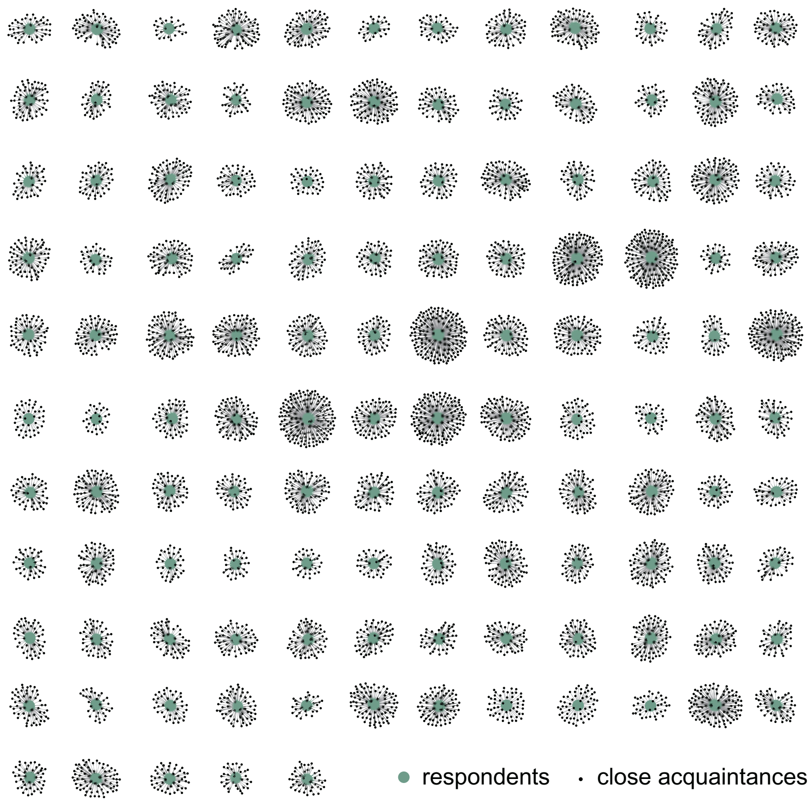

FIG. 6. Respondent's ego-networks. Each sociogram shows the respondent (the green central discs) connected (black lines) to their reported contacts (the black outer discs). The number of sampled students (i.e. egos) is 125 and the total number of contacts (i.e. alters) is 7462.

TABLE 3 *The true values, self-reported statistics (simple mean) and ECM estimates (Eq. (4.4)) using 125 respondents*

(%)	*Age*	*Gender /male*	*Province*	*Living expense*	*Virgin*	*Homosexual*	*Had sex with strangers*
True value	61.9	14.8	65.8	—	—	—	—
Self-reported	61.6	13.6	68.8	14.4	61.6	5.6	7.2
Estimate	61.0	13.1	62.3	13.2	64.0	7.4	8.2
95% CI	[59.6 ,61.6]	[12.2 ,14.0]	[61.5 ,63.2]	[12.7 ,13.7]	[62.7Â·,65.4]	[6.5 ,7.4]	[7.8 ,8.6]

3. Discussion and conclusion

Surveying populations about sensitive topics is challenging [35]. In various contexts researchers need to collect information that respondents are not willing to share due to the stigma associated or because of fear of potential legal consequences. For example, although efficient sampling methods such as venue-based or respondent-driven sampling have been successfully used to survey the behaviour of sex workers, their clients are many times unwilling to volunteer in survey studies for fear of being identified and stigmatized. Similarly, political tensions and religious influence in some locations may discourage self-reporting of political and sexual orientation. Generally speaking, sensitive topics are ubiquitous and thus one needs easy-to-implement effective methods that guarantee individual privacy, confidentiality and statistical reliability.

In this article, we propose a novel method (named ECM) which exploits social ties to anonymously obtain sensitive information from studied populations. We also propose an estimator for the proportion or prevalence of traits in such populations together with a bootstrap method to calculate the estimator's confidence interval.

The method consists in three steps: (i) census or population sample is used to identify potential respondents; (ii) two questions are asked to respondents, one about the number of social contacts in the target population and the second about the number of social contacts in the target population with a given trait; (iii) inference of population proportions is made using our novel estimator. We test our method under various simulated and real-world scenarios to study potential limitations and the applicability of the method. The results suggest that the method is generally robust in all studied scenarios and feasible for real-life applications.

A careful sensitive analysis indicated that the method is affected by the activity ratio, that is, the ratio of the average degrees (number of contacts) of the two populations with traits A and B (not A). If the average degree is relatively higher for one group, this group may be over-estimated and the second group may be under-estimated. Nevertheless, the bias is relatively small and within an acceptable interval when the activity ratio is not too far from 1 (e.g. $\pm30\%$), which is typically the case in real-world scenarios. The degree distribution, typically of concern, does not seem to affect the estimations significantly (biases less than 0.003). Our analysis also indicates that information about a few contacts is sufficient to obtain reliable estimations. This approach is important because respondents may not be aware of traits of all their social contacts and in practice asking about all contacts may induce errors. At the same time, asking about more than a single contact guarantees a level of confidentiality that does not compromise the social relation between respondents and their contacts, raising ethical concerns [36]. Furthermore, the method can provide estimates with small biases even facing the degree reporting error and trait reporting error.

The real-world exercise indicated that the method is relatively easy-to-implement and involves few amendments to standard sampling surveys. In particular, it introduces two simple and direct questions with little space to misinterpretations. We also observed that providing a list of potential contacts disconnected to the questionnaire, though not essential, may encourage and aid respondents to identify all their contacts and associate them to specific traits yet keeping privacy. The transparency and simplicity of the method (without for example randomizing devices or complex instructions), and its indirect and averaging nature make respondents less confused and more engaged. Before the real-world survey, we conduct a pilot on the respondents' willingness to participate in the survey. About 83% of the respondents (128 students) considered that the novel approach, that is, the peer network survey, guarantees their privacy and encourages them to complete the questionnaire until the end. In the real-world survey, 125 completed questionnaires were returned, corresponding to a response rate of 81%. The results also provided evidence that individuals tend to under-report traits that are stigmatized and many times not publicly known. Since the true values are missing, these results may indicate miss-perception of respondents towards contacts. A definitive answer is outside of the scope of our sampling method. It is however important to note that the proposed method is limited for situations when sensitive information is not shared between all peers but to closer friends. One approach to reduce these potential biases is to ask respondents to report on a few social contacts. A compromise however is necessary to achieve both statistically significant averages and reliability in the flow of information.

In summary, the proposed ECM shows practical advantages over more complex standard methods to estimate population proportions on sensitive topics in both census or sampled populations. The sensitive topics about sex orientation, virginity and living expense can be well estimated by ECM. The method requires two simple questions about the respondent's contacts to assure privacy for all individuals. The method encourages participation and reduces misreporting during data collection, providing robust results in both simulated and real-world scenarios. The method has potential to be a valuable tool for both qualitative and quantitative research in settings where privacy must be guaranteed. Further validation of the method is necessary, particularly to test its performance in settings where traditional methods have been challenged, as for example, in the study of clients of sex-workers, drug dealers via drug users, same-sex sexual contacts in countries where such activities are criminalize or in corruption.

4. Materials and methods

4.1 *Ego-centric sampling method*

The ECM consists in three steps: (i) census or population sample is used to identify potential respondents; (ii) the peer network survey: two questions are asked to each respondent i, one about the number of social contacts (i.e. the degree k_i) in the target population and the second about the number of social contacts k_i^A in the target population with a given trait A; (iii) inference: the peer network survey is used to make anonymized inferences of the proportion $P(A)$ of individuals with trait A in the target population. Steps (i) and (ii) can be applied in both qualitative and quantitative surveys. In this article, we study quantitative surveys and thus derive below a point estimator $\hat{P}(A)$ for the prevalence (step (iii)). A bootstrap procedure is suggested in the Supplementary Information, and used to analyse the real-life survey, to generate 95% confidence intervals.

We first define a social network as a set of N nodes i representing individuals connected by social ties (or links, i.e. (i,j)) in case they identify each other as friends or close acquaintances [37]. Assuming that these social ties are reciprocal (i.e. network links are undirected), the number of links from all type A nodes with degree k can be written as:

$$\sum_{i\in\{i|k_i=k\}} k_i^A = n_k p(A|k)k, \tag{4.1}$$

where $p(A|k)$ is the probability of being a type A node for nodes with degree k and n_k is the number of nodes with degree k. And the expectation of the number of nodes with trait A is:

$$\mathrm{E}(N_A) = \sum_{k=1}^{k_{\max}} p(A|k)n_k, \tag{4.2}$$

where $k_{\max}$ is the maximum degree of the network. The proportion of nodes with trait A in the network, $P(A)$, can thus be estimated by:

$$\hat{P}(A) = \frac{\sum_{k=1}^{k_{\max}} \sum_{i \in \{i | k_i = k\}} k_i^A / k}{N}. \tag{4.3}$$

We can also extend the above equation when only n random alters are used:

$$\hat{P}_R(A) = \frac{\sum_{k=1}^{k_{\max}^R} \sum_{R_i \in \{R_i | k_{R_i} = k\}} k_{R_i}^A / k}{n}, \tag{4.4}$$

where $\hat{P}_R(A)$ is the estimator for n random alters, $k_{\max}^R$ is the maximum degree of the random samples, and $k_{R_i}^A$ is the degree with trait A for a random respondent R_i.

4.2 *Activity ratio of social networks*

The activity ratio (AR) is the ratio of the mean degree for the group of nodes (individuals) that has trait A to the mean degree of the group of nodes that has trait B in the population, that is $\mathrm{AR} = \langle k \rangle^A / \langle k \rangle^B$. It quantifies how nodes with different properties affect the formation of their personal (ego) networks. Based on the KOSKK social network model (see Supplementary Information and refs. [28, 29]), we generate a set of networks with varying activity ratios by using the following algorithm: Let w be the activity ratio of the current network and w' the target value. When $w > w'$, a node i with trait A and a node j with trait B are randomly selected. Let the degree of nodes i and j are k_i and k_j respectively. If $k_i > k_j$, we swap the traits of the two nodes, that is node i gets trait B and j gets trait A. On the other hand, when $w < w'$, we swap the traits of the nodes only if $k_i < k_j$. This process is repeated until w reaches the desired w'. To generate different statistical scenarios, the network structure is kept fixed and a new trait list is randomly generated.

4.3 *Real-world survey*

To evaluate the effectiveness of ECM, we implement a empirical study at a university in Hunan Province, China, by inviting all third-year undergraduate students from the same program to participate the survey (see Supplementary Information for details). The peer network survey is conducted to collect both sensitive and non-sensitive ego-centric data. In addition, personal self-reported data are also collected for comparison. The questionnaire is divided into two parts. The first part collects self-reported data by directly asking the characteristics of the respondent. The questions include: (i) age (if more than 22 years old), (ii) gender (if male), (iii) birth province (if Hunan), (iv) living expenses (if less than 800 RMB – approx. 127 USD – per month), (v) virginity (if yes), (vi) sexual orientation (if self-identified as homosexual), and (vii) the history of having sex with strangers (if yes). The second part collects information about peers, that is, it asks the respondent to inform the number of his or her close acquaintances (social contacts) with the above traits (e.g. k_i^{age}) and the total number of his or her close acquaintances (i.e. k_i).

Some steps are taken to further encourage participation and improve the response rate and the reliability of the collected data. First, the interviewers inform the respondents that the researchers analysing the data are not involved in data collection. Second, the questionnaire is filled out anonymously to assure confidentiality and privacy. Third, besides the anonymous questionnaire, a list of names of all students is given to each respondent, to aid on counting the total number of close acquaintances and the number of close acquaintances who has a given trait. For a total of 155 students from the same program, 125 of them returned complete and consistent questionnaires, giving a response rate of 81%.

Supplementary data

Supplementary data are available at *COMNET* online.

Funding

The National Nature Science Foundation of China (71771213, 72025405 and 82041020 to X.L. and 71901067 to S.C.); the Hunan Science and Technology Plan Project (2020TP1013 and 2020JJ4673); the Shenzhen Basic Research Project for Development of Science and Technology (JCYJ20200109141218676), the National Key Research and Development Program of China (2020YFC0849200) and the Key Joint Project for Data Center of the National Natural Science Foundation of China and Guangdong Provincial Government (U1611264), in part.

Acknowledgements

The authors would like to thank Professor Yong Li and Dr. Beiling Ma for the help in implementing the empirical study.

References

1. Groves, R. M., Fowler, Floyd J., J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2011) *Survey Methodology*. New Jersey: Wiley.

2. Kalton, G. (2020) *Introduction to Survey Sampling*. California, US: Sage Publications.

3. Tourangeau, R. & Yan, T. (2007) Sensitive questions in surveys. *Psychol. Bull.*, **133**, 859–883.

4. Fisher, R. J. (1993) Social desirability bias and the validity of indirect questioning. *J. Consumer Res.*, **20**, 303–315.

5. Baral, S., Beyrer, C., Muessig, K., Poteat, T., Wirtz, A. L., Decker, M. R., Sherman, S. G. & Kerrigan, D. (2012) Burden of HIV among female sex workers in low-income and middle-income countries: a systematic review and meta-analysis. *Lancet Infect. Dis.*, **12**, 538–549.

6. Aceijas, C., Stimson, G. V., Hickman, M. & Rhodes, T. (2004) Global overview of injecting drug use and HIV infection among injecting drug users. *AIDS*, **18**, 2295–2303.

7. Beyrer, C., Baral, S. D., Griensven, F. V., Goodreau, S. M., Chariyalertsak, S., Wirtz, A. L. & Brookmeyer, R. (2012) Global epidemiology of HIV infection in men who have sex with men. *Lancet*, **380**, 367.

8. Belli, R. F., Traugott, M. W. & Beckmann, M. N. (2001) What Leads to Voting Overreports? Contrasts of overreporters to validated voters and admitted nonvoters in the American National Election Studies. *J. Off. Stat.*, **17**, 479–498.

9. Böckenholt, U. & Heijden, P. G. M. V. D. (2007) Item randomized-response models for measuring noncompliance: risk-return perceptions, social influences, and self-protective responses. *Psychometrika*, **72**, 245–262.

10. Aquilino, W. S. & Losciuto, L. A. (1990) Interview mode effects in drug use surveys. *Public Opin. Q.*, **54**, 362–395.

11. Corkrey, R. & Parkinson, L. (2002) A comparison of four computer-based telephone interviewing methods: getting answers to sensitive questions. *Behav. Res. Methods Instrum. Comput.*, **34**, 354.

12. Tourangeau, R., Rasinski, K., Jobe, J. B., Smith, T. W. & Pratt, W. F. (1997) Sources of error in a survey on sexual behavior. *J. Off. Stat.*, **13**, 341–365.

13. Couper, M. P., Singer, E. & Tourangeau, R. (2003) Understanding the effects of audio-CASI on self-reports of sensitive behavior. *Public Opin. Q.*, **67**, 385–395.

14. Jones, E. E. & Sigall, H. (1971) The bogus pipeline: a new paradigm for measuring affect and attitude. *Psychol. Bull.*, **76**, 349–364.

15. Tourangeau, R., Smith, T. W. & Rasinski, K. A. (2010) Motivation to report sensitive behaviors on surveys: evidence from a bogus pipeline experiment. *J. Appl. Soc. Psychol.*, **27**, 209–222.

16. Warner, S. L. (1965) Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.*, **60**, 63–69.

17. Dalton, D. R., Wimbush, J. C. & Daily, C. M. (1994) Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Pers. Psychol.*, **47**, 817–829.

18. Droitcour, J. A. & Larson, E. M. (2002) An innovative technique for asking sensitive questions: the three-card method. *BMS*, **75**, 5–23.

19. Tourangeau, R. & Smith, T. W. (1996) Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opin. Q.*, **60**, 275–304.

20. Nuno, A. & John, F. A. V. S. (2015) How to ask sensitive questions in conservation: a review of specialized questioning techniques. *Biol. Conserv.*, **189**, 5–15.

21. Miller, J. D. (1985) The nominative technique: a new method of estimating heroin prevalence. *NIDA Res. Monogr.*, **57**, 104.

22. Fishburne, P. M. (1980) Survey techniques for studying threatening topics: a case study on the use of heroin. *Doctoral Dissertation.*

23. Rossier, C. (2010) Measuring abortion with the anonymous third party reporting method, chapter 7. *Methodologies for Estimating Abortion Incidence and Abortion-related Morbidity: A Review*. (S. Singh & L. T. A. Remez, eds). New York, US: Guttmacher Institute, IUSSP, pp. 99–106.

24. Burt, R. S. (1984) Network items and the general social survey. *Soc. Netw.*, **6**, 293–339.

25. Heckathorn, D. D. (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Soc. Probl.*, **44**, 174–199.

26. Erdős, P. & Rényi, A. (1959) On random graphs. *Publ. Math.*, **6**, 290–297.

27. Albert, R. & Barabási, A. (2001) Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, xii.

28. Kumpula, J. M., Onnela, J. P., Saramäki, J., Kaski, K. & Kertész, J. (2007) Emergence of communities in weighted networks. *Phys. Rev. Lett.*, **99**, 228701.

29. Toivonen, R., Kovanen, L., Kivelä, M., Onnela, J. P., Saramäki, J. & Kaski, K. (2009) A comparative study of social network models: network evolution models and nodal attribute models. *Soc. Netw.*, **31**, 240–254.

30. Gile, K. J. & Handcock, M. S. (2010) Respondent driven sampling: an assessment of current methodology. *Sociol. Methodol.*, **40**, 285.

31. Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B. J., Thorson, A. & Liljeros, F. (2012) The sensitivity of respondent - driven sampling. *J. R. Stat. Soc. A*, **175**, 191–216.

32. Huang, G., Cai, M. & Lu, X. (2019) Inferring opinions and behavioral characteristics of gay men with large scale multilingual text from blued. *Int. J. Environ. Res. Public Health*, **16**, 3597.

33. Efron, B. & Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1**, 54–75.

34. Baraff, A. J., Mccormick, T. H. & Raftery, A. E. (2016) Estimating uncertainty in respondent-driven sampling using a tree bootstrap method.. *Proc. Natl. Acad. Sci. USA*, **113**, 14668.

35. Barnett, J. (1998) Sensitive questions and response effects: an evaluation. *J. Manag. Psychol.*, **13**, 63–76.

36. Cronin B., Perra N., Rocha L. E. C., Zhu Z., Pallotti F., Gorgoni S., Conaldi G. & De Vita R. (2021) Ethical implications of network data in business and management settings. *Soc. Netw.*, **67**, 29–40.

37. Wasserman, S. & Faust, K. (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.