METRON
https://doi.org/10.1007/s40300-023-00259-y

Deriving a zero-truncated modelling methodology to analyse capture–recapture data from self-reported social networks

Mark E. Piatek[1] · **Dankmar Böhning**[1]

Received: 13 October 2022 / Accepted: 31 October 2023

Abstract
Capture–recapture (CRC) is widely used to estimate the size (N) of hidden human populations (e.g., the homeless) from the overlap of sample units between two or more repeated samples or lists (a.k.a., capture occasions). There is growing interest in deriving CRC data from social-network data. The current paper hence explored if self-reported social networks (lists of social ties) submitted by participants from the target population could function as distinct capture occasions. We particularly considered the application of zero-truncated count distribution modelling to this type of data. A case study and follow-up simulation study focused on two methodological issues: (1) that a participant cannot be named in their own self-reported social network and hence cannot be named as many times as non-participants; and (2) positive dependence between being a participant and being named by (a social tie of) other participants. Regarding the latter, a further motivation of the simulation study was to consider the impact of using respondent-driven sampling to select participants, because all non-seed RDS participants are recruited as a social tie of another participant. Exponential random graph modelling was used to generate the simulation study's target populations. Early comparison was also made to estimates of N from Successive Sampling.

Keywords Capture–recapture · Hidden populations · Population size estimation · Zero-truncated modelling · Social networks · Exponential random graph modelling · Respondent-driven sampling

1 Introduction

Capture–recapture (CRC) analysis is widely used to estimate the size of hidden human populations, who may stay under the radar because of societal stigma or lack of exhaustive recording. Estimating their size can help governments make informed funding/policy decisions and be useful for sociological or medical research. Examples include estimating the number of people who inject drugs (PWID) (e.g., [16, 23, 47]), female sex-workers (FSW)

✉ Mark E. Piatek
mp1n17@soton.ac.uk

[1] Southampton Statistical Science Research Institute, University of Southampton, Southampton, UK

(e.g., [23, 42, 46]), men who have sex with men (MSM) (e.g., [23, 46]), the homeless (e.g., [19, 20]) and cases during the recent Covid-19 pandemic (e.g., [13, 50]).

CRC estimates population size from the overlap of sample units between two or more capture occasions (repeated samples of the target population). The presence of a sample unit at a capture occasion is called a 'capture', so if present at three capture occasions we would say it is 'captured' three times. The often-stigmatised nature of hidden human populations makes it hard to obtain capture occasions via direct sampling, so various indirect approaches have been devised. For example, a register from an external organisation (such as a subscriber list) is often used as a capture occasion (e.g., [19, 47]).

Exploration continues into ways to source CRC data of human populations. The current paper considers a novel pairing of CRC data derived from self-reported social networks of participants and a CRC analytical technique called 'zero-truncated count distribution modelling'. In the current paper's approach, a sample group of participants from the target population each submit a list of individuals they know from that population. Each of these self-reported social networks is treated as a distinct capture occasion. Because participants cannot be named in their own list, we also consider if being a participant can be treated as an extra capture occasion.

Treating each participant's self-reported social network as a distinct capture occasion means many capture occasions can be obtained from a single sample of participants. Pairing this with the zero-truncated modelling approach makes use of this strength, as it is particularly suited for analysing overlap between many capture occasions.

1.1 Advantages of using three or more capture occasions

An advantage of using three or more capture occasions is that otherwise CRC is heavily reliant on the assumed independence between capture occasions. Positive dependence between capture occasions is known to lead to underestimates of population size and vice versa, which can be severe when only using two capture occasions because estimation is limited to the Lincoln–Petersen estimator or equivalent [41]. Sensitivity to the independence assumption increases sensitivity to other CRC assumptions that play into it. For example, false matches of individuals between capture occasions can cause negative dependence whereas unequal catchability of individuals can cause positive dependence because some individuals are more likely to be captured multiple times.

Obtaining three or more capture occasions enables more nuanced methods of estimating population size with less sensitivity to the independence assumption. Examples include: log linear modelling, which can consider dependencies between three or more capture occasions (e.g., [3, 19, 20, 47, 53]); Bayesian latent class modelling (e.g., [23, 42, 44]); the 'ratio plot' diagnostic tool in zero-truncated modelling [5, 9]; and continuous time CRC modelling that, when using many capture occasions over time, can allow for a delayed onset of behaviours relating to being repeatedly captured [25].

There is hence great interest in exploring ways to source three or more capture occasions. Examples include: using a single external register of repeated entries, such as a hospital admissions register (e.g., [8, 35, 54]); selecting a sample of participants from the target population as a third capture occasion in combination with two external registers (e.g., [47, 53]); and a 3-capture-occasion adaptation of the Unique-object Multiplier method (e.g., [23, 42, 44]).

1.2 Utilising social networks in capture–recapture

Several of the above approaches involve selecting a (theoretically) representative sample from the target population as a capture occasion. In the absence of a sample frame, these approaches often utilise the social ties between members of the target population to perform respondent-driven sampling (RDS), which is thought to approach a somewhat representative sample [34].

While it is hence not uncommon for participants' social networks to be used in CRC to expand the reach of sampling, their use as capture occasions is less explored. However, interest in this area is growing. One approach by Dombrowski et al. [22] derives two capture occasions, where the first capture is being a participant and the second is being named in at least one other participant's self-reported social network. In a more recent example in Buchanan et al. [16], the first capture was being a participant and the second was being named in at least one other participant's self-reported social network who did not appear in one's own. Recent advances in population size estimation using Privatised Network Sampling have extended the Dombrowski et al. [22] approach with more nuanced population-size estimators [26, 37]. Two of these from Fellows [26], the 'cross-alter' and 'cross-network' estimators, considered the overlap of non-participants and participants between participants' self-reported social networks, which was found to produce better estimates due to increasing the number of individuals in the data [26]. The latter was a promising sign that participants' self-reported social networks could function as a viable source of captures of participants and non-participants. Besides CRC, other methods that use social-network data from a target population to estimate its size include the Snowball method [20, 27], the Network Scale-up method (e.g., [24]), Successive Sampling [32, 33] and CRC Successive Sampling [38].

While methods like Successive Sampling ask participants how many individuals they know from the target population, the current paper's approach instead asks them to list individuals. This increases data sensitivity, particularly as CRC is often used on stigmatised populations. However, a number of studies have demonstrated ways for participants to submit anonymised or pseudo-anonymised lists of social ties, showing it can be practicable [16, 22, 26, 37].

We initially present a case study, using real-world data, that attempts to estimate a known target population size. Methodology is outlined in Sect. 2, initial inspection of data is in Sect. 3.1, model fitting is in Sect. 3.2 and estimation of population size is in Sect. 3.3. A follow-up simulation study is described in Sect. 4 and further discussion is in Sect. 5.

2 Methods

2.1 Zero-truncated count distribution modelling

'Zero-truncated count distribution modelling' has long been used to estimate population size from CRC data. See Böhning et al. [11] for an introduction on its application to human populations. The method uses aggregate-level data of how many sample units are captured exactly 1 time (f_1), exactly 2 times (f_2), exactly 3 times (f_3), etc. across several capture occasions. This is summarised into a frequency distribution of discrete counts (Table 1). The number of sample units captured at least once is referred to as the observed population (n). The number of uncaptured sample units is f_0. Hence, the total size of the target population is $N = f_0 + n$.

Table 1 Underlying structure of CRC data when summarised into a frequency distribution

Number of times captured (x)	0	1	2	3	4	5	...	m	n
Frequency (f_x)	? (f_0)	f_1	f_2	f_3	f_4	f_5	...	f_m	$\sum_{x=1}^{m} f_x$

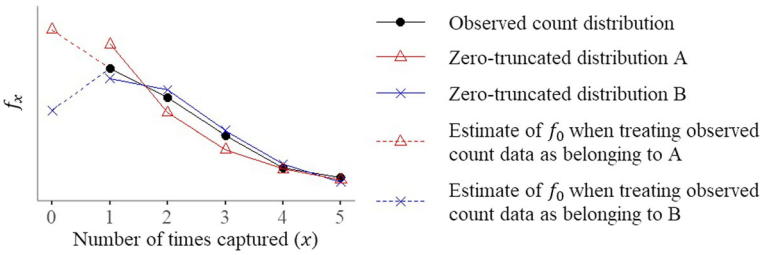

Fig. 1 Graphical display demonstrating zero-truncated modelling approach

To estimate the size (N) of the target population, the method assumes that the observed data (f_1 to f_m) follow the same shape/profile as f_0 to f_m. Hence, it assumes that the shape of f_1 to f_m can be projected further to the left to estimate f_0. To approximate the shape of f_1 to f_m, a zero-truncated (ZT) probability distribution is used (e.g., ZT Poisson). Based on estimated parameters from that ZT distribution, f_0 (and subsequently N) is estimated as though f_0 to f_m follow the untruncated version of that ZT distribution.

Several ZT distributions can be considered to see which is most consistent with the data. For example, in Fig. 1, the data is closer to ZT distribution B than ZT distribution A, suggesting estimates of N should be based on B rather than A. While a variety of distributions can theoretically be considered, the current literature tends to focus on a small number of distributions stemming from the ZT Poisson and ZT geometric, described further in Sect. 2.2.

The approach tends to involve choosing one or two ZT distributions as starting points and then using model-fitting diagnostic tests, like the chi-square goodness of fit (GoF) test, to check for inconsistency with the observed data. A conceptual drawback is that, by testing for inconsistency, ZT distributions can be deemed consistent with the data via acceptance of the null hypothesis. For this reason, the use of the GoF test here only tends to provide partial evidence of consistency between a ZT distribution and the observed data rather than a more conclusive finding. However, the 'ratio plot' diagnostic tool (described in Sect. 2.3) can help inform the results from these tests by providing a more granular inspection of how consistent the ZT distributions are to the data.

As more generally with CRC analysis, this rests on some key assumptions: firstly, that capture occasions are independent; secondly, that all members of the target population (sample units) are equally catchable (i.e., have a homogeneous probability of being captured) at any capture occasion; thirdly, that individuals are accurately matched between capture occasions; and fourthly, that the target population is closed/static across all capture occasions.

2.2 Zero-truncated distributions considered

As f_1 to f_m are non-negative integers, the ZT Poisson tends to be a starting point for modelling them. However, with ZT Poisson the mean and variance are assumed equal, which is often not true because heterogeneity (unequal capture probabilities among sample units) can cause

over-or-underdispersion of variance. Hence, other commonly used distributions include the ZT geometric and various mixing distributions of the ZT Poisson, such as the Poisson-gamma and the Conway–Maxwell Poisson (CMP). In the current paper, the ZT Poisson and ZT geometric were used as starting points, as their popularity has led to the development of population-size estimators with robustness to heterogeneity. Also considered were the zero-truncated one-inflated (ZTOI) Poisson and ZTOI geometric, as they can factor in an overly large f_1.

The untruncated Poisson has a probability mass function (pmf) of: $p_x = \exp(-\theta)\theta^x/x!$ for $x = 0, 1, \ldots, m$ where $\hat{\theta}_{MLE} = \bar{x}$. The ZT Poisson has pmf: $p_x = \theta^x/((\exp(\theta) - 1)x!)$ for $x = 1, 2, \ldots, m$. Although $\hat{\theta}_{MLE}$ cannot be given in closed form for the ZT Poisson, it can be derived from the untruncated Poisson using the E-M algorithm [11, 21]. This oscillates between two steps, using an arbitrary initial value (e.g., 0.5) for $\hat{\theta}$. In Step 1, $\hat{f}_0 = n \times \exp(-\hat{\theta})/(1 - \exp(-\hat{\theta}))$ and, in Step 2, $\hat{\theta} = S/(n + \hat{f}_0)$ where $S = \sum_{x=1}^{m} x f_x$ and $n = \sum_{x=1}^{m} f_x$. The ZTOI Poisson has pmf:

$$p_x^{1+} = \left\{ \begin{array}{l} w + (1-w)\frac{\lambda}{\exp(\lambda)-1} \ldots \text{ if } x = 1 \\ (1-w)\frac{\lambda^x}{(\exp(\lambda)-1)x!} \ldots \text{ if } x > 1 \end{array} \right\}$$

for $x = 1, 2, \ldots, m$ where w is a weight parameter. For the ZTOI Poisson, $\hat{\lambda}_{MLE}$ and $\hat{w}_{MLE}$ cannot be given in closed form and were hence iteratively calculated using an E-M algorithm approach from Godwin and Böhning [28]. This cycled through the following steps. In step 1, we assigned arbitrary initial values (e.g., 0.5) for $\hat{N}$ and $\hat{\delta_1}$. (The $\hat{\delta_1}$ is the number of unobservable inflated 1s.) In step 2, we estimated $\hat{w}$ via $\hat{w} = \hat{\delta_1}/n$. In step 3, we estimated $\hat{\lambda}$ via $\hat{\lambda} = (\sum_{i=1}^{\hat{N}} x_i - \hat{\delta_1})/(\hat{N}(1-\hat{w}))$ where $\sum_{i=1}^{\hat{N}} x_i \equiv \sum_{x=1}^{m} x f_x$. In step 4, we estimated $\hat{\delta_1}$ via $\hat{\delta_1} = f_1 \times \hat{w}(1-\exp(-\hat{\lambda}))/\big(\hat{w}(1-\exp(-\hat{\lambda})) + (1-\hat{w})\hat{\lambda}\exp(-\hat{\lambda})\big)$. In step 5, we repeated steps 2-4 until $\hat{\delta_1}$ converged. Then, in step 6, we re-estimated $\hat{N}$ via $\hat{N} = n/(1 - \exp(-\hat{\lambda}))$. We repeated steps 2-6 until $\hat{N}$ converged, at which point $\hat{\lambda}_{MLE}$ and $\hat{w}_{MLE}$ would also have converged.

The untruncated geometric has pmf: $p_x = (1-\theta)^x\theta$ for $x = 0, 1, \ldots, m$ where $\hat{\theta}_{MLE} = 1/(\bar{x} + 1)$. The ZT geometric has pmf: $p_x = (1-\theta)^{(x-1)}\theta$ for $x = 1, 2, \ldots, m$ where $\hat{\theta}_{MLE} = 1/\bar{x} = 1/(S/n)$. The ZTOI geometric has pmf:

$$p_x^{1+} = \left\{ \begin{array}{c} w(1-\theta)^x\theta/(1-w\theta) \ldots \text{ if } x > 1 \\ \big((1-w) + w(1-\theta)^x\theta\big)/(1-w\theta) \ldots \text{ if } x = 1 \end{array} \right\}$$

for $x = 1, 2, \ldots, m$ where w is a weight parameter; $0 \leq w \leq 1$. For the ZTOI geometric, $\hat{\theta}_{MLE}$ and $\hat{w}_{MLE}$ cannot be given in closed form and were hence calculated via the nested E-M algorithm approach from Kaskasamkul and Böhning [36]. This oscillates between the following two steps, using arbitrary initial values (e.g., 0.5) for $\hat{\theta}$ and $\hat{w}$. In Step 1, $\hat{f}_0 = n \times \hat{w} \times \hat{\theta}/(1 - \hat{w} \times \hat{\theta})$ and $\hat{N} = n + \hat{f}_0$. In Step 2, $\hat{w} = 1 - (f_1/\hat{N})(1-\hat{w})/\big((1-\hat{w}) + \hat{w}(1-\hat{\theta})\hat{\theta}\big)$ and

$$\hat{\theta} = \frac{\hat{N} - f_1(1-\hat{w})/((1-\hat{w}) + \hat{w}(1-\hat{\theta})\hat{\theta})}{\hat{N} + \sum_{i=1}^{\hat{N}} x_i - 2f_1(1-\hat{w})/\big((1-\hat{w}) + \hat{w}(1-\hat{\theta})\hat{\theta}\big)}$$

where $\sum_{i=1}^{\hat{N}} x_i \equiv \sum_{x=1}^{m} x f_x$.

2.3 The ratio plot

The 'ratio plot' diagnostic tool [9] was used to help inform model-fitting diagnostics. This is a graph in which horizontality of data-points indicates the level of consistency between a given ZT distribution and observed data.

The tool is based on the power series density $p_x(\theta) = a_x\theta^x / \sum_{x=0}^{\infty}\{a_x\theta^x\}$ where $\sum_{x=0}^{\infty}\{a_x\theta^x\}$ is a normalising constant that converts $p_x(\theta)$ into proportions that sum to 1. If a_x is set to $a_x = 1/x!$, the power series density becomes the Poisson density: $p_x(\theta) = (\theta^x/x!)/\sum_{x=0}^{\infty}\{\theta^x/x!\}$. If a_x is set to $a_x = 1$, the power series density becomes the geometric density: $p_x(\theta^\star) = \theta^{\star x}/\sum_{x=0}^{\infty}\{\theta^{\star x}\}$ where $\theta^\star = 1-\theta$. Although outside the scope of the current paper, it can also model a binomial density by setting $a_x = \binom{T}{x}$ for $x = 0, \ldots T$ where T are positive integers and where $a_x = 0$ when $x > T$ [5].

A property of the power series density is that the ratio of $p_{x+1}(\theta)a_x$ to $p_x(\theta)a_{x+1}$ is always equal to θ, which is a constant. That is, $r_x = p_{x+1}(\theta)a_x/(p_x(\theta)a_{x+1}) = \theta$. This means r_x should also be a constant. Hence, to check for consistency between a set of observed data (f_1 to f_m) and a specific power series density (e.g., Poisson), the observed data can be combined with a_x to produce r_x and the level of constancy in r_x can be inspected.

Although $p_x(\theta)$ is unknown, f_x/N can be used instead as a non-parametric estimate; the quantity of N is unknown but cancels out in the ratio. Hence, r_x can be estimated as $\hat{r}_x = (a_x/a_{x+1}) \times (f_{x+1}/f_x)$ where f_{x+1}/f_x is the ratio of each adjacent pair of counts in the observed data and a_x/a_{x+1} is the inverse of their respective coefficients from the power series distribution. In the Poisson case, $a_x/a_{x+1} = x+1$ and hence $\hat{r}_x = (x+1)f_{x+1}/f_x$. In the geometric case, $a_x/a_{x+1} = 1$ and hence $\hat{r}_x = f_{x+1}/f_x$.

If $\hat{r}_x$ is a constant then a horizontal series of data-points should occur when plotting $\log(\hat{r}_x)$ against x. From this, the consistency between a given ZT distribution and the observed data can be visually appraised by inspecting the gradient of a linear regression line plotted between x and $\log(\hat{r}_x)$. A more horizontal line provides some evidence that the ZT distribution under consideration is consistent with the observed data [9]. Care needs to be taken, however, as heterogeneity (unequal catchability among sample units) can also contribute to causing a slope in the linear regression line.

Rocchetti et al. [49] advise using weighted linear regression to reduce the impact of heteroskedasticity in the ratio plot, using weights (W_i) derived from diagonal components of $(cov(Y))^{-1}$:

$$W = \begin{bmatrix} \frac{1}{f_1}+\frac{1}{f_2} & \cdots & \cdots \\ \cdots & \frac{1}{f_i}+\frac{1}{f_{i+1}} & \cdots \\ \cdots & \cdots & \frac{1}{f_{m-1}}+\frac{1}{f_m} \end{bmatrix}^{-1} = \begin{bmatrix} f_1+f_2 & \cdots & \cdots \\ \cdots & f_i+f_{i+1} & \cdots \\ \cdots & \cdots & f_{m-1}+f_m \end{bmatrix}$$

These weights ($W_i = f_i + f_{i+1}$) are the same for both the ZT Poisson and ZT geometric. For example, to calculate weighted regression in the Poisson case, we take $\hat{\beta} = (X^T W X)^{-1} X^T W Y$ where

$$Y = \begin{pmatrix} \log(2f_2/f_1) \\ \log(3f_3/f_2) \\ \vdots \\ \log(mf_m/f_{m-1}) \end{pmatrix}, X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & m-1 \end{pmatrix}, W = \begin{pmatrix} f_1+f_2 \\ f_2+f_3 \\ \vdots \\ f_{m-1}+f_m \end{pmatrix}$$

A useful property of the power series density (and hence the ratio plot) is that it is the same when untruncated or zero-truncated. This brings an added utility to ratio plots, as the

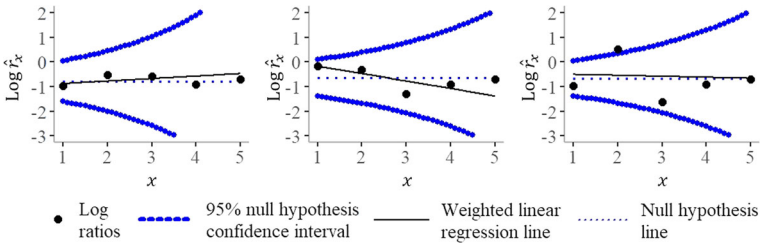

Fig. 2 Graphs demonstrating concept of ratio plot under the null. *Left panel:* demonstration of high consistency. *Central panel:* demonstration of worse consistency. *Right panel:* demonstration of overall consistency apart from one ratio

intercept of the weighted linear regression line can extrapolate the ratio plot to where $x = 0$. From this can be derived the Weighted Least Squares (WLS) estimator of population size [5].

2.4 The ratio plot under the null

Because the ratio plot ($\hat{r}_x$) is the same for both the untruncated and zero-truncated versions of a given distribution, a 95% confidence interval can be displayed pertaining to the null hypothesis that f_1 to f_m are consistent with the untruncated version of the ZT distribution being tested. This 'ratio plot under the null' is demonstrated in Fig. 2 using dummy data. See the accompanying supplementary material (Online Resources 3–4) for examples in R code.

While the null hypothesis holds in both the left and central panels of Fig. 2, evidence of consistency is stronger in the left panel because the weighted regression line is nearly horizontal and all ratios are well within the null hypothesis region. We might hence choose the ZT distribution that was being tested in the left panel over that of the centre panel. In the right panel, the weighted regression line is nearly horizontal but the ratio of f_2 to f_3 is outside the null hypothesis region. The latter gives some evidence of overall consistency apart from f_2 and/or f_3, prompting consideration of other ZT distributions.

Null hypothesis regions for the ratio plots were calculated using methodology from Böhning and Punyapornwithaya [7]. In both the Poisson and geometric cases, these were calculated as $\log \hat{r}_x \pm 1.96 \times \sqrt{var(\log \hat{r}_x)}$.

The null hypothesis in the geometric case is that f_1 to f_m are consistent with the untruncated geometric distribution (i.e., consistent with the power series density when $a_x = 1$). Setting $a_x = 1$ makes the power series density become $p_x(\theta^\star)$ from the geometric density, where $\theta^\star = 1 - \theta$. Because we are using the log scale in our ratio plot, the null hypothesis region is calculated around $\log(1 - \hat{\theta})$. Using an approximated variance term for $var(\log \hat{r}_x)$, the null hypothesis region in the geometric case is hence

$$\log(1 - \hat{\theta}) \pm 1.96 \times \sqrt{\frac{1}{n(1 - \hat{\theta})^{x+1}\hat{\theta}} + \frac{1}{n(1 - \hat{\theta})^{x}\hat{\theta}}}$$

where $\hat{\theta}$ is $\hat{\theta}_{MLE}$ from the ZT geometric. This is given in closed form as: $\hat{\theta}_{MLE} = 1/\bar{x} = 1/(S/n)$ where $S = \sum_{x=1}^{m} x f_x$ and $n = \sum_{x=1}^{m} f_x$.

The null hypothesis in the Poisson case is that f_1 to f_m are consistent with the untruncated Poisson (i.e., consistent with the power series density when $a_x = 1/x!$). Because we are using the log scale in our ratio plot, the null hypothesis region is calculated around $\log \hat{\theta}$. Using an approximated variance term for $var(\log \hat{r}_x)$, the null hypothesis region in the Poisson case

is hence

$$\log\hat{\theta} \pm 1.96 \times \sqrt{\frac{1}{n \times \exp(-\hat{\theta})\hat{\theta}^{x+1}/(x+1)!} + \frac{1}{n \times \exp(-\hat{\theta})\hat{\theta}^{x}/x!}}$$

where $\hat{\theta}$ is $\hat{\theta}_{MLE}$ from the ZT Poisson. However, as $\hat{\theta}_{MLE}$ cannot be given in closed form from the ZT Poisson, it is iteratively calculated from the untruncated Poisson using the E-M algorithm (see Sect. 2.2).

2.5 Estimators of population size

Several estimators of population size (N) are available, specific to each ZT distribution (see e.g., [11]). Some give equal weight to the whole of f_1 to f_m, such as the Turing [29] and MLE estimators. Other estimators give more importance to f_1 and f_2, as they are typically larger than f_3 to f_m and hence less impacted by fluctuation caused by heterogeneity (unequal catchability of sample units). This gives the latter some robustness to heterogeneity. A popular example is the Chao estimator [18]. A bias-corrected Chao (BC Chao) is also available for small populations [10], and both the Chao and BC Chao have been adapted for one-inflated data [12]. The WLS estimator (mentioned at the end of Sect. 2.3) is similarly robust, as it gives more weight to f_1 and f_2 [5]. The current paper hence prioritised using BC Chao and WLS.

The BC Chao estimator is: $\hat{N} = n + f_1(f_1 - 1)/(2f_2 + 2)$ for ZT Poisson, $\hat{N} = n + f_1(f_1-1)/(f_2+1)$ for ZT geometric, $\hat{N} = n+2\times(f_2^3-3f_2^2+2f_2)/(9\times(f_3+1)\times(f_3+2))$ for ZTOI Poisson and $\hat{N} = n+(f_2^3-3f_2^2+2f_2)/((f_3+1)\times(f_3+2))$ for ZTOI geometric. The WLS estimator is $\hat{N} = n + f_1 \times \exp(-\hat{\beta}_0)$ where $\hat{\beta}_0$ is the intercept of the weighted linear regression line of the ratio plot. This meant the WLS estimator was only available for the ZT Poisson and ZT geometric in the current paper. See Online Resource 1 for worked examples.

2.6 Confidence intervals for population-size estimators

In the case study, 95% confidence intervals for estimators of N were calculated via the imputed bootstrap approach. Methodology was sourced from Anan et al. [1] which had earlier roots in Buckland and Garthwaite [17], Norris and Pollock [43] and Zwane and Van der Heijden [56]. In this approach, the point-estimate of $\hat{N}$ and $\hat{f}_0$ are used to estimate probabilities of f_0 to f_m as $\hat{p}_i = \left\{\frac{\hat{f}_0}{\hat{N}}, \quad \frac{f_1}{\hat{N}}, \quad \frac{f_2}{\hat{N}}, \quad \frac{f_3}{\hat{N}}, \quad \ldots, \quad \frac{f_m}{\hat{N}}\right\}$. These are entered as the 'prob' parameter of the 'rmultinom' function in R statistical software [48], with 'size' set to $\hat{N}$ and the 'n' parameter set to the desired number of bootstrap samples. Hence, each bootstrap sample is a multinomially distributed random vector in which $\hat{N}$ is split across f_0 to f_m. The same point-estimate of $\hat{N}$ is then calculated from each bootstrap sample, producing a bell-curve around the original point-estimate. The 97.5th and 2.5th percentiles become the 95% confidence interval.

Anan et al. [1] advise a bootstrap approach as it avoids needing to assume estimators of N are normally distributed. There would be more reliance on this assumption if instead basing confidence intervals on variance estimators, as these are known to be impacted by skewness of N in CRC [18]. Nevertheless, a slight bias can enter the estimated probabilities ($\hat{p}_i$) of x_i when drawing bootstrap samples if the point-estimate of $\hat{N}$ is an over-or under-estimate.

Table 2 General structure of how captures were recorded, using dummy data

Person	Participant*	Nominated by...*					Number of times captured...	
		A	B	C	D	E	Nom. and Par.[a]	Nom.[b]
Person A	1	0	1	0	0	0	**2**	**1**
Person B	1	1	0	0	0	0	**2**	**1**
Person C	0	1	1	0	1	0	**3**	**3**
Person D	1	0	0	0	0	0	**1**	**0**
Person E	0	0	1	0	1	0	**2**	**2**

*1 = yes, 0 = no
[a] at nomination and participation capture occasions
[b] at nomination capture occasions

While Anan et al. [1] suggest only needing $\hat{N}$ number of bootstrap samples, the use of R software meant thousands could be drawn.

In the follow-up simulation study, confidence intervals were instead the 2.5th and 97.5th percentiles of $\hat{N}$ across many simulated target populations.

2.7 Case study target population and sampling approach

The target population in the case study was a cohort of 182 students on a university course. As this was an experimental setting, a participation invite was sent across the target population and a sample group of participants was formed by voluntary participation. Invites were sent at the same time, helping ensure the population was closed. Participants were asked to submit their self-reported social network (list of social ties) independently. The non-sensitive nature of the population, and its small size, meant full names were able to function as unique identifiers.

2.8 Case study derivation of capture–recapture data

In the case study, participants' self-reported social networks were treated as distinct capture occasions (referred to as 'nomination capture occasions'). For each unique individual (participant or non-participant) named in at least one self-reported social network, we counted how many they were named in (i.e., how many nomination capture occasions they were captured at). Out of K self-reported social networks, non-participants could be named up to K times. However, participants could only be named up to $K - 1$ times because they could not be named in their own social network. As this violated the assumed equal catchability of sample units at all capture occasions, we considered if being a participant could be an extra capture to make the maximum number of captures equal (referred to as the 'participation capture occasion').

The structure of how captures were recorded is shown in Table 2. In this demonstration table, persons A, B and D are participants whereas persons C and E are non-participants. Hence, the columns 'Nominated by C' and 'Nominated by E' contain only 0s. The two right-hand columns show the total number of times each person was captured when either including or excluding the participation capture occasion. Data from either of the right-hand columns could then be summarised into a frequency distribution (f_1 to f_m) of how many

Table 3 Number of individuals captured x number of times at nomination and participation capture occasions, by whether participant

Number of times captured (x)	0	1	2	3	4	5	6	7	8	Observed population (n)
Frequency (f_x)	?(121)	29	9	11	8	0	1	2	1	61
of which...										
Participants	0	1	3	5	3	0	1	2	1	16
Non-participants	?(121)	28	6	6	5	0	0	0	0	45

Table 4 Number of individuals captured x number of times at nomination capture occasions, by whether participant

Number of times captured (x)	0	1	2	3	4	5	6	7	8	Observed population (n)
Frequency (f_x)	?(122)	31	11	9	5	1	2	1	0	60
of which...										
Participants	1	3	5	3	0	1	2	1	0	15
Non-participants	?(121)	28	6	6	5	0	0	0	0	45

Table 5 Nomination and participation captures collapsed into two capture occasions

		Nominated by one or more participants		
		Yes	No	
Participant	Yes	15	1	16
	No	45	121	166
		60	122	182

individuals were captured exactly 1 time, exactly 2 times, exactly 3 times, etc. when either including or excluding the participation capture occasion.

3 Case study results

3.1 Case study results from data-collection

16 participants took part in the study, each submitting a self-reported social network. As described above, captures were summarised into a frequency distribution (f_1 to f_m) of the number of individuals captured x number of times. Two versions of the dataset were derived that either included or excluded the participation capture occasion, shown in Tables 3 and 4 respectively.

In Table 3, 61 individuals were captured at least once, of which 29 were captured exactly 1 time, 9 were captured exactly 2 times, 11 exactly 3 times, etc. The number of uncaptured individuals (f_0) would ordinarily be unknown but in this case was known to be 121 in Table 3 and 122 in Table 4. The inclusion of the participation capture occasion increased the number of times each participant was captured by 1, which meant the 'Participants' row was shifted right in Table 3 compared to in Table 4.

Collapsing the data into two binary capture occasions (Table 5) showed there was strong positive dependence between participation and nomination. 15 out of 16 participants were

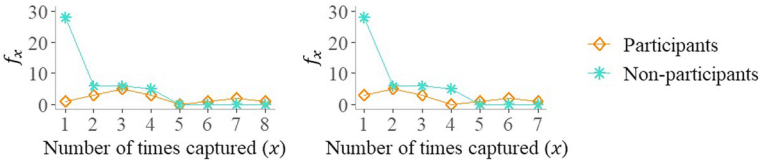

Fig. 3 Number of individuals captured x number of times. *Left panel*: from nomination and participation capture occasions. *Right panel*: from nomination capture occasions

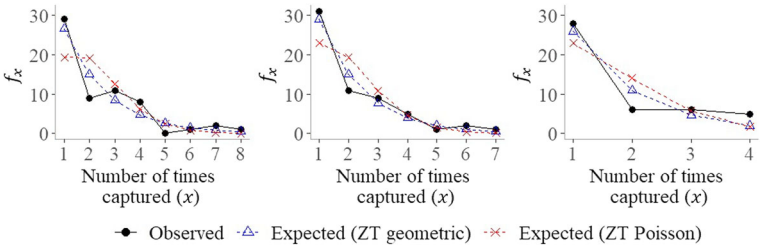

Fig. 4 Observed data (f_1 to f_m) vs expected values from ZT Poisson and ZT geometric. *Left panel:* from nomination and participation capture occasions. *Centre panel:* from nomination capture occasions. *Right panel:* from captures of just non-participants

nominated at least once. Positive dependence between capture occasions is known to produce underestimates of population size [14]. Hence, while an argument for including the participation capture occasion was that it would make the maximum number of captures equal between participants and non-participants, a counter-argument was that excluding it could be a way to effectively deflate captures of participants by 1 to help counter-balance the positive dependence that had been found.

As sample units are assumed to have equal probabilities of capture, f_1 to f_m should theoretically have a similar shape/profile across all sub-sections of the observed population. However, the shape/profile was found to substantially differ for participants and non-participants, even when only considering nomination capture occasions (Fig. 3). This suggested that estimates of population size (N) could potentially be more accurate if f_1 to f_m only included captures of non-participants, as this would have a more homogeneous shape/profile and be easier to fit with a ZT distribution. The number of participants would need to be added on to any such estimates of N afterwards because the exclusion of participants from the CRC data would mean treating participants as outside the target population during the calculation of estimators.

To summarise, initial inspection of the case study data suggested three possible ways of deriving CRC data (f_1 to f_m) with which to estimate N. These were: (a) as captures from nomination and participation capture occasions (total row from Table 3); (b) as captures from nomination capture occasions (total row from Table 4); or (c) as captures of just non-participants (from either Table 3 or 4). Model-fitting and population-size estimation was performed on all three versions to see how they compared.

3.2 Case study results from model-fitting

Visual appraisal of observed vs expected values (Fig. 4) suggested that, for all three versions of the observed data (f_1 to f_m) under consideration, the data was more consistent with the

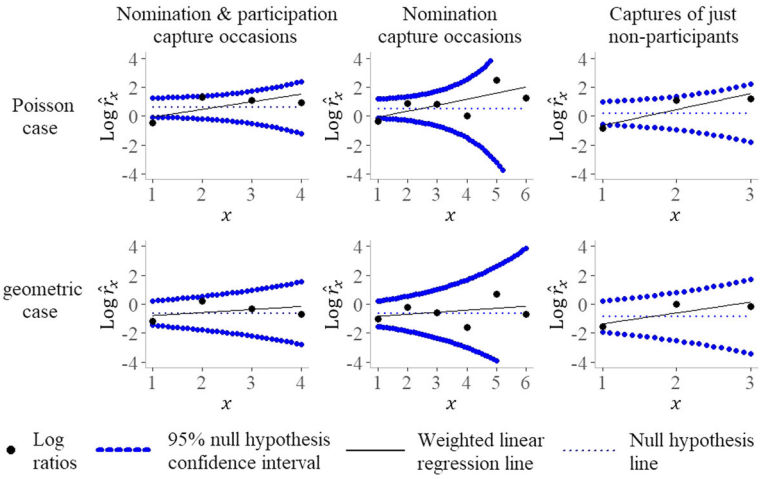

Fig. 5 Ratio plots under the null in Poisson and geometric cases. Left panels contain only 4 ratios because $f_5 = 0$, which meant f_5 to f_8 were collapsed into f_5 so there was no gap in the ratio plots. As the x-axis starts at 1, the intercept of the weighted regression line is not displayed. Intercepts were: -0.586 (top left panel); -0.5102 (top centre); -1.7566 (top right); -0.9957 (bottom left); -0.9747 (bottom centre); and -2.1106 (bottom right)

ZT geometric than the ZT Poisson. MLE parameter values ($\hat{\theta}_{MLE}$) were used for expected values as per Sect. 2.2.

Ratio plots under the null are shown in Fig. 5. For all three versions of f_1 to f_m, the weighted regression line was more horizontal in the geometric case. This again gave some evidence that the ZT geometric was more consistent with f_1 to f_m. In the Poisson case, the left-most ratio (that of f_1 to f_2) fell outside the null hypothesis region, indicating that f_1 and/or f_2 were inconsistent with the untruncated Poisson. There was thus a particular risk of bias entering population-size estimators from the ZT Poisson that prioritise f_1 and/or f_2.

To test if the observed data were inconsistent with the ZT Poisson or ZT geometric, the chi-square goodness-of-fit statistic, $\chi^2 = \sum_{x=1}^{m-1}\{(\log \hat{r_x} - \log \bar{r_x})^2/\widehat{var}(\log \hat{r_x})\}$, was used on the ratio-plot data where $\widehat{var}(\log \hat{r_x}) = 1/f_{x+1} + 1/f_x$ [5, 9, 49]. The mean ratio ($\bar{r}_x$) was used as the expected $\hat{r}_x$ value which, as described in Sect. 2.3, is expected to be a constant. The mean ratio is $\bar{r}_x = \sum_{x=1}^{m-1}\{(x+1)f_{x+1}\}/\sum_{x=1}^{m-1} f_x$ in the Poisson case and $\bar{r}_x = \sum_{x=1}^{m-1} f_{x+1}/\sum_{x=1}^{m-1} f_x$ in the geometric case. A significant p-value (< 0.05) meant rejecting the null hypothesis that f_1 to f_m were consistent with the ZT distribution being tested. Results are shown in Table 6. There was not enough evidence to reject the null hypothesis in the geometric case, which was partial evidence of consistency with ZT geometric. While a non-significant p-value occurred in the Poisson case when only using nomination capture occasions ($p = 0.13$), this was outweighed by the earlier finding in the Poisson ratio-plots that the ratio of f_1 to f_2 was outside the null hypothesis region because of the particular importance of fitting f_1 and f_2.

For all three versions of f_1 to f_m, a sizeable jump between f_2 and f_1 suggested possible consistency with a one-inflated distribution. We therefore considered if the ZTOI geometric or ZTOI Poisson offered a closer fit than their ZT counterparts. In each case, the likelihood ratio test (LRT) was used with $\alpha = 0.05$. This test is calculated as $LRT = -2 \times (l_0(0, \tilde{\theta}) - l_A(\hat{w}, \hat{\theta}))$ where l_0 and l_A are the log likelihoods under the null and alternative hypotheses respectively. In the null hypothesis case, $\tilde{\theta}$ was the $\hat{\theta}_{MLE}$ from the ZT distribution whereas, in

Table 6 Chi-square goodness of fit test results

CRC data from...	Distribution	χ^2 value	*df*	*p* value
Nomination and participation capture occasions[a]	ZT Poisson	11.653	4	0.02*
	ZT geometric	5.755	4	0.2
Nomination capture occasions	ZT Poisson	9.864	6	0.13
	ZT geometric	4.183	6	0.65
Captures of just non-participants	ZT Poisson	10.585	3	0.01**
	ZT geometric	5.75	3	0.11

*Significant at 0.05 level; ** significant at 0.01 level
[a]When f_5 to f_8 are collapsed into f_5

Table 7 LRT of ZTOI vs ZT geometric and of ZTOI vs ZT Poisson

Capture occasions	ZTOI vs ZT geometric			ZTOI vs ZT Poisson		
	$l_A(\hat{w}, \hat{\theta})$	$l_0(0, \tilde{\theta})$	LRT	$l_A(\hat{w}, \hat{\lambda})$	$l_0(0, \tilde{\lambda})$	LRT
Nom. and Par.[a]	− 95.534	− 95.880	0.692	− 94.702	− 101.689	13.975
Nom.[b]	− 85.635	− 85.886	0.501	− 85.219	− 90.206	9.974
Just non-participants	− 52.692	− 53.139	0.893	− 51.182	− 54.444	6.525

[a]CRC data derived from nomination and participation capture occasions
[b]CRC data derived from nomination capture occasions

the alternative hypothesis case, $\hat{\theta}$ and $\hat{w}$ were the $\hat{\theta}_{MLE}$ and $\hat{w}_{MLE}$ from the ZTOI distribution. All MLEs were calculated as per Sect. 2.2.

In the geometric case, log likelihoods were calculated as per methodology in Kaskasamkul and Böhning [36] with $l_A(\hat{w}, \hat{\theta}) = f_1 \log\{((1 - \hat{w}) + \hat{w}(1 - \hat{\theta})\hat{\theta})/(1 - \hat{w}\hat{\theta})\} + \sum_{x=2}^{m} f_x \log\{\hat{w}(1 - \hat{\theta})^x\hat{\theta}/(1 - \hat{w}\hat{\theta})\}$. In the null hypothesis case, $w = 1$ so hence $l_0(0, \tilde{\theta}) = \sum_{x=1}^{m} f_x \log\{(1 - \tilde{\theta})^x\tilde{\theta}/(1 - \tilde{\theta})\}$.

In the Poisson case, log likelihoods were calculated as per methodology in Godwin and Böhning [28], with $l_A(\hat{w}, \hat{\lambda}) = f_1 \log\{\hat{w} + (1 - \hat{w})\hat{\lambda}/(\exp(\hat{\lambda}) - 1)\} + \sum_{x=2}^{m} f_x \log\{(1 - \hat{w})\hat{\lambda}^x/((\exp(\hat{\lambda}) - 1)x!)\}$. In the null hypothesis case, $w = 0$ so hence $l_0(0, \tilde{\lambda}) = \sum_{x=1}^{m} f_x \log\{\tilde{\lambda}^x/((\exp(\tilde{\lambda}) - 1)x!)\}$.

See also Böhning and van der Heijden [6] for a simplified LRT approach that uses the 'zero-one truncated' likelihood instead of the ZTOI likelihood, as this can act as an equivalent and is more straightforward to calculate.

As the LRT approximates a two-tailed χ^2 test with 1 degree of freedom, a critical value of 2.706 was used ($\alpha = 0.05$). Results are shown in Table 7. For all three versions of f_1 to f_m, the LRT in the geometric case was less than 2.706, indicating there was not enough evidence to say the ZTOI geometric was a closer fit than ZT geometric. However, in the Poisson case the LRT was larger than 2.706, indicating the ZTOI Poisson was a closer fit than ZT Poisson.

In summary, for all three versions of f_1 to f_m, model-fitting diagnostics showed partial evidence of consistency with the ZT geometric as well as the ZTOI Poisson. Hence, estimates of N in Sect. 3.3 were based on both.

3.3 Case study results from population size estimation

Estimates of population size (N) are shown in Table 8. When the CRC data only contained

Table 8 Estimated size of the case study population ($\hat{N}$ with 95% confidence intervals)

	Nom. and Par.[a]	Nom.[b]	Just non-participants
N	182	182	182
Observed population (n)	61 (34% of N)	60 (33% of N)	45 (27% of ($N - 16$))
ZT geometric BC Chao	142 (88, 281)	138 (87, 261)	169 (95, 401)[c]
ZT geometric WLS	139 (83, 286)[d]	142 (90, 290)	292 (112, 1,221)[c]
ZTOI Poisson BC Chao	62 (59, 68)	62 (58, 77)	61 (61, 69)[c]

[a]CRC data derived from nomination and participation capture occasions
[b]CRC data derived from nomination capture occasions
[c]When based on non-participant captures, $\hat{N}$ was increased by 16 to factor in participants
[d]Here f_5 to f_8 were collapsed into f_5

captures of non-participants (right column), the number of participants (16) was added on to estimators of N because excluding participants from the data meant they were treated as outside the target population during the calculation of estimators.

N was consistently underestimated when participants were included in the CRC data (left and centre columns of Table 8). This was as expected due to the positive dependence found between participation and nomination. When instead excluding participants (right column of Table 8), underestimation was less severe in the ZT geometric's BC Chao estimator and there was in fact overestimation in the ZT geometric's WLS estimator. However, excluding the participants reduced the size of the observed population (n) (i.e., the number of individuals captured at least once in the data) from 61 to 45, leading to wider confidence-intervals for both the ZT geometric's estimators. A guide to the minimum required size of n relative to N (a.k.a., the minimum capture proportion) is given in Xi et al. [55] which advises that n be at least 58% of N when $N = 200$. This was used as a rough guide of the required n when $N = 182$. This was not satisfied by any of the three versions of the CRC data, as the capture proportion was only 34% (61/182) when the data included nomination and participation capture occasions, 33% (60/182) when just including nomination capture occasions and 27% (45/(182 − 16)) when just including captures of non-participants. In the latter case, the capture proportion was calculated as $n/(N-$number of participants) because excluding participants from the CRC data meant treating participants as outside the target population when calculating estimators of N. The ZTOI Poisson's BC Chao estimator produced a particularly low underestimate of N with all three versions of the CRC data, suggesting it had heavy reliance on a sufficient capture proportion.

4 Follow-up simulation study

To help inform the case study, the approach was next performed on thousands of simulated target populations where N was 182 or 500. In each case, we simulated a target population as a complete social network, drew a sample of participants from it and derived a CRC dataset (f_1 to f_m) from the overlap of participants' alters (social ties). Each participant's network of alters (which could include other participants and/or non-participants) was treated as equivalent to a nomination capture occasion from the case study. For example, being an alter of three participants meant being captured three times via nomination. Like in the case study, being a participant was also considered as a possible extra capture occasion to make the maximum captures equal.

A further motivation was to consider the impact of using respondent-driven sampling (RDS) to select participants. In real-world settings, the often-elusive nature of target populations may often necessitate using a multi-wave sampling method like RDS to select participants. However, RDS poses a difficulty for the current paper's approach because all non-seed participants are recruited via being an alter of another participant (described in Sect. 4.2), hence creating positive dependence between participation and nomination. We therefore derived CRC datasets from participants selected via either RDS or simple random sampling (SRS) to see how estimates of N would compare.

4.1 Simulation study target populations

Each simulated target population was a complete undirected social network, of either 182 or 500 actors, simulated by an exponential random graph model (ERGM). In this method, a network of actors (182 or 500 in this case) is simulated via a probability function that cycles through random pairs of actors (with replacement), each time generating a binary outcome of whether they are tied (1 = tie; 0 = no tie) (see e.g., [40, 52]). The probability function can include several parameterised terms that each make 1 or 0 more likely. While similar to logistic regression, a key difference is that the probability function often uses terms that cause partial dependence between binary outcomes.

The current study's ERGM used four terms: 'edges', 'k-star', 'k-triangle' and 'k-2path'. This has been advised as an effective baseline for modelling social networks [39, 52], particularly for modelling heterogeneity of degree size and transitivity (a clustering effect wherein ties are more likely between actors who share mutual ties). Further description is given in Online Resource 1.

Parameter values for edges, k-star, k-triangle and k-2path were set to -4, 0.2, 1 and -0.2 respectively, each with effect size (λ) = 2. These values were sourced from Pattison et al. [45], who found they were in line with findings from empirical datasets. These were thus treated as somewhat representative of typical social networks. Simulation was via R statistical software [48], using the 'ergm' package [31] from the 'statnet' suite. In the ergm package, the k-triangle and k-2path terms are substituted with the 'geometrically weighted edgewise shared partners' (GWESP) and 'geometrically weighted dyadwise shared partners' (GWDSP) terms respectively, which are equivalent terms but use $\lambda = \log(2) = 0.693$ [52]. This combination of terms produced an average degree size of 3.13 in networks of 182 actors and 4.88 in networks of 500 actors.

As the ERGM's probability function proceeds, it forms a Markov Chain Monte Carlo (MCMC) process wherein, after a suitable burn-in period, social network characteristics should reach convergence and settle into a regular pattern. Snapshots of complete social networks can then be taken, which is referred to as 'sampling the graph' [40]. We used sampled graphs as simulated target populations of either 182 or 500 actors. A wide gap between sampled graphs protected against autocorrelation. Like Pattison et al. [45], we sampled every 100,000th graph with a burn-in of 1,000,000. For example, to generate 1000 simulated target populations of 182 actors, the ERGM was specified as:

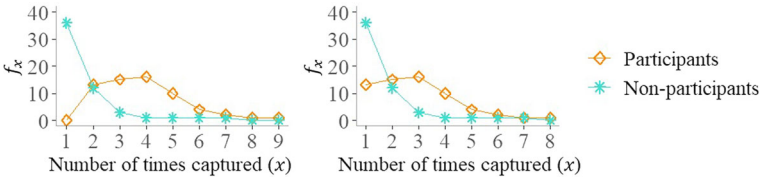

Fig. 6 Average number of individuals captured x times from CRC datasets where participant selection was via RDS, the number of participants $= 60$ and $N = 182$. *Left panel*: nomination and participation capture occasions. *Right panel*: nomination capture occasions

```
simulate(~edges + altkstar(lambda=2,T) + gwesp(decay=0.693,
T) + gwdsp(decay=0.693,T), coef = c(-4, 0.2, 1, -0.2),
nsim = 1000, basis = network(x = 182, directed = FALSE),
control = control.simulate.formula.ergm(MCMC.burnin =
1000000, MCMC.interval = 100000))
```

4.2 Simulation study sampling approach

When selecting participants via RDS, this began by randomly selecting five seed participants (without replacement) from the target population. From each participant's alters, two further participants were randomly selected where possible from among those who were not already sampled. From this new wave of participants was selected a further wave of participants in the same way, and so on across several waves until the desired number of total participants was reached. This almost always resulted in four waves and a partial fifth wave.

When instead selecting participants via SRS, a random number generator was used to select the desired number of participants (without replacement) from the target population in just one sampling wave.

4.3 Simulation study derivation of capture–recapture data

As mentioned near the beginning of Sect. 4, a particular challenge when using RDS was that all non-seed participants were recruited via being an alter of another participant and were hence captured at least once via nomination. Moreover, the average shape/profile of f_1 to f_m showed that participants had a higher probability of being captured multiple times than non-participants (Fig. 6). Therefore, like in the case study, three ways of deriving f_1 to f_m were considered: (a) as captures from nomination and participation capture occasions; (b) as captures from nomination capture occasions; or (c) as captures of just non-participants. Option 'b' would include participants in the CRC data but effectively deflate their captures by not counting participation as an extra capture, partially offsetting positive dependence. Option 'c' would remove more positive dependence by removing all captures of participants, though at the cost of reducing the size of n. Meanwhile, when using SRS to select participants, only option 'a' was used because participation was independent of being nominated by (an alter of) others.

Table 9 Estimated size of simulated populations (mean $\hat{N}$) where participant selection was via SRS and CRC datasets were derived from nomination and participation capture occasions in all columns

Total CRC datasets	10,000	10,000	10,000
N	182	182	500
SRS participants	16	40	50
Observed population ($\bar{n}$)	56 (31% of N)	112 (62% of N)	227 (45% of N)
Datasets where only f_1 and f_2 were larger than 0			
Number of datasets	5045	0	3
ZT Poisson BC Chao	229 (106, 552)	n/a	501
ZT geometric BC Chao	402 (163, 1042)	n/a	779
Datasets where f_1 to f_m were consistent with ZT Poisson			
Number of datasets	1971	6998	7149
ZT Poisson BC Chao	143 (89, 208)	183 (142, 235)	514 (404, 657)
ZT Poisson WLS	135 (65, 256)	178 (130, 243)	483 (311, 698)
Datasets where f_1 to f_m were consistent with ZT geometric			
Number of datasets	2864	2776	2692
ZT geometric BC Chao	369 (171, 747)	309 (220, 439)	933 (691, 1237)
ZT geometric WLS	1555 (191, 5717)	329 (214, 571)	989 (659, 1610)
Datasets where f_1 to f_m were consistent with ZTOI Poisson			
Number of datasets	228	253	265
ZTOI Poisson BC Chao	57 (48, 70)	130 (109, 157)	292
Datasets where f_1 to f_m were consistent with ZTOI geometric			
Number of datasets	60	5	0
ZTOI geometric BC Chao	61	179	n/a
Datasets where f_1 to f_m were inconsistent with any of the above			
Number of datasets	0	207	155

Confidence intervals are shown when number of datasets $> N$

4.4 Simulation study model-fitting diagnostics

As with the case study, model-fitting diagnostics were used to check how consistent each CRC dataset was with the ZT/ZTOI Poisson/geometric although only if $\sum_{x=3}^{m} f_x > 0$. Any gaps partway through f_1 to f_m were removed by collapsing data from further along the tail. For a dataset to be deemed consistent with the ZT Poisson, it needed a non-significant χ^2 GoF p-value for ZT Poisson, a more horizontal ratio plot in the Poisson case than the geometric, and a non-significant LRT for ZTOI Poisson. Consistency with ZT geometric was equivalently assessed. For a dataset to be consistent with ZTOI Poisson or ZTOI geometric, the LRT needed to be > 2.706. Like the case study, some datasets were consistent with the ZT version of one distribution and the ZTOI version of the other (e.g., consistent with ZT geometric and ZTOI Poisson).

4.5 Simulation study results

Table 9 shows mean estimates of N from CRC datasets where participant selection was via SRS and N was 182. The 2.5th and 97.5th percentiles of $\hat{N}$ were used as 95% confidence intervals. As participant selection via SRS was independent of being nominated by (an alter

of) others, f_1 to f_m were derived from nomination and participation capture occasions in all three columns.

The left column of Table 9 was somewhat comparable to the case study, as each dataset was derived using 16 participants and the number of individuals (participants and non-participants) captured at least once (n) was, on average, 56. This was similar to the case study wherein n had been 61. Accuracy of estimators was low in this column, indicating that, even if there had been independence between nomination and participation, the case study would have needed more than 16 participants to produce reliable estimates of N.

Using 40 participants (centre column of Table 9) increased $\bar{n}$ to 112 (62% of N). This was now roughly in line with the advised minimum capture proportion in Xi et al. [55], which was that n be at least 58% of N when $N = 200$. Estimates of N in this column were somewhat more accurate overall, but especially when CRC datasets were consistent with ZT Poisson. This also occurred in the right column of Table 9 wherein N was 500 and each dataset was derived using 50 participants, producing an $\bar{n}$ of 227 (45% of 500) that was in line with the advice in Xi et al. [55] that n be at least 44% of N when $N = 500$.

Tables 10 and 11 show estimates of N when instead using RDS to select participants, with N being 182 and 500 respectively. When CRC datasets only contained captures of non-participants (right column of both tables) the number of participants was added on to estimators of N because excluding participants from the data meant they were treated as outside the target population during the calculation of estimators.

In all columns of Tables 10 and 11 apart from the right column of Table 10, the number of participants was such that $\bar{n}$ met the advised minimum capture proportion. Exclusion of participants from CRC datasets in the right column of both tables meant the capture proportion was $\bar{n}/(N-$number of participants) and, because n only included non-participants, it was more difficult to meet the advised minimum capture proportion. When N was 500 (right column of Table 11), 140 participants were used because this produced, on average, 219 non-participants and hence a capture proportion of $219/(500 - 140) = 61\%$. This satisfied the advised minimum proportion (58%) pertaining to $N = 200$, which was used as the threshold because $500-140$ was between 200 and 500. However, when N was 182, the average number of non-participants did not reach the advised minimum capture proportion no matter how many participants were used. Instead, for demonstration purposes, the right column of Table 10 used 60 participants, producing a capture proportion of only $51/(182 - 60) = 42\%$. This was lower than the advised minimum (69%) pertaining to $N = 100$, which was used as the threshold because $182 - 60$ was between 100 and 200.

Returning to Table 9, there were indications in its centre and right columns that, when nomination and participation were independent and n was sufficiently large, the current paper's approach would more often produce a Poisson shape across f_0 to f_m. Approximately 70% of CRC datasets in the centre and right columns were consistent with ZT Poisson and produced accurate average estimates of N from ZT Poisson estimators. There was also greater accuracy from the ZT Poisson's BC Chao estimator than that of the ZT geometric when only f_1 and f_2 were larger than 0. Meanwhile, approximately 28% of datasets in these columns were consistent with ZT geometric and overestimated N using ZT geometric estimators. Further work is needed to explore if this would occur outside the current simulation setting and, if so, how to obtain better estimates from datasets consistent with ZT geometric. For example, the ZT Conway–Maxwell Poisson (CMP) distribution can function as a mixture of ZT Poisson and ZT geometric and has its own WLS estimator of N [2, 51].

This suggested a similar trend towards ZT Poisson might occur when using RDS to select participants and using captures of just non-participants. This was indeed found in the right column of Table 11, wherein 7479 out of 10,000 datasets were consistent with ZT Poisson and

Table 10 Estimated size of simulated populations (mean $\hat{N}$) where participant selection was via RDS and $N = 182$

	Nomination and participation capture occasions	Nomination capture occasions	Captures of just non-participants
Total CRC datasets	10,000	10,000	10,000
N	182	182	182
RDS participants	60	60	60
Observed population ($\bar{n}$)	111 (61% of N)	111 (61% of N)	51 (42%)[a]
Datasets where f_1 to f_m were consistent with ZT Poisson			
Number of datasets	1125	2416	4173
ZT Poisson BC Chao	126 (107, 145)	140 (117, 163)	153 (123, 196)[b]
ZT Poisson WLS	126 (108, 145)	143 (120, 166)	148 (112, 206)[b]
Datasets where f_1 to f_m were consistent with ZT geometric			
Number of datasets	8152	7107	5388
ZT geometric BC Chao	168 (135, 216)	201 (153, 262)	239 (167, 366)[b]
ZT geometric WLS	161 (130, 200)	196 (151, 259)	346 (176, 895)[b]
Datasets where f_1 to f_m were consistent with ZTOI Poisson			
Number of datasets	8287	4502	579
ZTOI Poisson BC Chao	119 (103, 143)	122 (104, 149)	116 (100, 141)[b]
Datasets where f_1 to f_m were consistent with ZTOI geometric			
Number of datasets	0	9	48
ZTOI geometric BC Chao	n/a	139	110[b]
Datasets where f_1 to f_m were inconsistent with any of the above			
Number of datasets	63	180	15

Confidence intervals are shown when number of datasets $> N$

[a]When datasets only included non-participant captures, $\bar{n}$ was 42% of $(N - 60)$

[b]When based on non-participant captures, $\hat{N}$ was increased by 60 to factor in participants

Table 11 Estimated size of simulated populations (mean $\hat{N}$) where participant selection was via RDS and $N = 500$

	Nomination & participation capture occasions	Nomination capture occasions	Captures of just non-participants
Total CRC datasets	10,000	10,000	10,000
N	500	500	500
RDS participants	70	70	140
Observed population ($\bar{n}$)	230 (46% of N)	231 (46% of N)	219 (61%)[a]
Datasets where f_1 to f_m were consistent with ZT Poisson			
Number of datasets	0	64	7479
ZT Poisson BC Chao	n/a	348	484 (436, 538)[b]
ZT Poisson WLS	n/a	361	488 (422, 563)[b]
Datasets where f_1 to f_m were consistent with ZT geometric			
Number of datasets	6056	7933	2320
ZT geometric BC Chao	504 (407, 603)	613 (479, 771)	686 (589, 805)[b]
ZT geometric WLS	523 (415, 640)	631 (466, 838)	668 (562, 820)[b]
Datasets where f_1 to f_m were consistent with ZTOI Poisson			
Number of datasets	8787	9591	535
ZTOI Poisson BC Chao	288 (232, 397)	265 (231, 321)	416 (364, 501)[b]
Datasets where f_1 to f_m were consistent with ZTOI geometric			
Number of datasets	1572	821	0
ZTOI geometric BC Chao	381 (265, 602)	332 (257, 492)	n/a
Datasets where f_1 to f_m were inconsistent with any of the above			
Number of datasets	259	9	183

Confidence intervals are shown when number of datasets $> N$

[a]When datasets only included non-participant captures, $\bar{n}$ was 61% of $(N - 140)$

[b]When based on non-participant captures, $\hat{N}$ was increased by 140 to factor in participants

Table 12 Successive Sampling population size estimate (SS-PSE) from 1000 simulated populations (mean $\hat{N}$ with 95% confidence intervals)

N	182	500	500
RDS participants	60	70	140
SS-PSE using low prior	127 (86, 183)	230 (149, 326)	268 (192, 391)
SS-PSE using accurate prior	162 (97, 249)	372 (231, 526)	334 (220, 522)
SS-PSE using high prior	265 (133, 407)	763 (495, 1047)	533 (324, 859)

produced only slight underestimates of N. However, this only occurred when the advised minimum capture proportion was satisfied, as the same trend was not present in the right column of Table 10 wherein the advised minimum proportion had not been met.

Further exploration of different sized simulated target populations found that, when using captures of just non-participants, n could, on average, only meet the advised minimum capture proportion when $N >= 300$ (described in Online Resource 1). Hence, for target populations where $N < 300$, it instead seemed optimal to use captures of participants and non-participants from just nomination capture occasions, which would partially offset positive dependence between participation and nomination while sustaining a large enough n to feasibly meet the minimum capture proportion. The centre column of Table 10 showed that, at least in the current simulation setting, this could lead to relatively accurate estimates of N from the approximately 70% of datasets consistent with ZT geometric in that column. This was because the more general tendency for ZT geometric estimators to overestimate N (seen earlier in the centre and right columns of Table 9) was somewhat counter-balanced by underestimation caused by positive dependence in the data.

Using RDS to select participants meant the Successive Sampling population size estimator (SS-PSE) of N could also be performed as a comparison [32, 33]. The SS-PSE is Bayesian and requires a prior estimate of N. For this, we used high, low or accurate priors based on those used in a simulation study in Handcock et al. [32]. Each prior was a beta distribution with a median of either $2 \times N$ in the high case, (N+number of participants)/2 in the low case or N in the accurate case. The median of the posterior distribution of $\hat{N}$ was taken as a point-estimate from each simulated target population. Calculation was via the 'sspse' R package [30]. See Online Resource 7 for an example. While the 'visibility' parameter of 'sspse' can improve estimates of N by imputing an adjusted degree size that factors in self-reporting bias, this was set to 'false' because participants' social ties were not self-reported in this study.

SS-PSE estimates of N are shown in Table 12. The 2.5th and 97.5th percentiles of the point-estimates from across simulated target populations were used as 95% confidence intervals. We found that, when the median of the prior was accurate (i.e., N), the median of the posterior distribution tended to underestimate N. This may have been partly due to either transitivity in the ERGM model that generated the target populations, the relatively small mean degree size (3.13 when $N = 182$ and 4.88 when $N = 500$) and/or using only five seed participants. This seeming sensitivity of the SS-PSE to one or more of these factors contrasted with the high level of accuracy we had seen from ZT Poisson estimators in the centre and right columns of Table 9 and the right column of Table 11, which appeared to be comparatively shielded by being based on only non-participant data.

5 Discussion

Results from the case study and simulation study gave a mixed picture of the current paper's approach. Selecting participants via RDS in the simulation study and via voluntary participation in the case study both brought substantial positive dependence between participation and nomination, leading to deviation from the assumed independence of capture occasions and the assumed equal catchability of participants and non-participants.

However, the accuracy of ZT Poisson estimators in the right column of Table 11 was a promising early finding that, when using RDS to select participants, many CRC datasets produced quite accurate estimates of N when only including captures of non-participants and having a sufficiently large n after excluding participants. The accuracy of ZT geometric estimators in the centre column of Table 10 was also a promising finding that, when needing to include participants in the data to keep n sufficiently large, approximately 70% of CRC datasets produced relatively accurate estimates of N by including captures of participants and non-participants from just nomination capture occasions. Nevertheless, further work is needed to obtain better estimates of N from the remaining 30% of CRC datasets in both of those columns.

More work is also needed to explore other RDS settings and methodological extensions. For example, for target populations with more clustering, a system of design weights could potentially be used to offset clustering in each participant's list of social ties. The current paper's approach also needs a fuller comparison to the SS-PSE [32, 33] and/or Privatised Network Sampling (PNS) [26, 37], as they can factor in the multi-wave structure of RDS data.

Simulating target populations via ERGMs opens the door to exploring more varied target populations that could better reflect real-world settings. For example, while the average degree size in the current simulation study (3.13 when $N = 182$ and 4.88 when $N = 500$) may have been in line with general human populations, degree size may be smaller among elusive populations and make one-inflation more prevalent. As another example, an ERGM simulating a population of drug users could incorporate covariate information, like a metric of drug use, that could be a major additional source of heterogeneity in individuals' degree size and hence their capture-rate.

An advantage of sourcing captures from participants' self-reported social networks is that many non-participants can be captured indirectly. Frank and Snijders [27] highlight a similar advantage in the Snowball Method, noting that it allows time and resources to be devoted to fewer but longer interviews that may encourage participation from stigmatised populations.

However, self-reported lists of social ties have a known susceptibility to memory-bias [4, 15]. This could be exacerbated if participants are asked to list individuals using pseudo-anonymised identifiers derived from several demographic variables, which may be necessary with stigmatised populations [16, 22, 26, 37]. For example, in Buchanan et al. [16], participants were asked to list individuals by combining abbreviations of initials, gender, age and district. Some populations might not be familiar enough to know such details about each other or remember them accurately, potentially leading to false matches. One possible solution is the Telefunken approach of Dombrowski et al. [22], in which participants are asked to include the last four digits of peoples' phone numbers (if known) when compiling unique identifiers.

As the target population in the case study were sitting the same university course, they could feasibly appear in each other's social networks. This was necessary to satisfy the equal-catchability assumption. For a larger or more fragmented target population, a stratified approach could potentially be used. For example, the ZT Poisson's BC Chao estimator could

be performed on M subsections of a target population via $\hat{N} = \sum_{i=1}^{m}\{n_i + f_{1i}(f_{1i} - 1)/(2f_{2i} + 2)\}$. Similarly, stratification could be used to factor in a variable (e.g., ethnicity) if it was suspected to affect the capture-rate of sample units.

6 Conclusion

There has been growing interest in deriving CRC data from self-reported social networks, and the current paper adds to this by considering a methodology for applying zero-truncated modelling to this type of data. This included an early exploration of how the approach could be applied when selecting participants via RDS, which was an important practical consideration for real-world settings. The approach still needs to be more fully compared to others, particularly as RDS brings a multi-wave structure to data that methods like Successive Sampling and Privatised Network Sampling can factor in. Further work is also needed to explore more varied target populations and key limitations such as task-fatigue, memory-bias, stratification of large populations and the necessary level of social ties/cohesion in the target population.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40300-023-00259-y.

Funding The authors received no financial support for the research and/or authorship of this article.

Declarations

Conflict of interest The authors have no conflicting interests to disclose.

Ethics approval Ethics approval was obtained via the University of Southampton's Ethics and Research Governance Online (ERGO) system. ERGO Ethical Approval Ref: 26150.

References

1. Anan, O., Böhning, D., Maruotti, A.: Uncertainty estimation in heterogeneous capture–recapture count data. J. Stat. Comput. Simul. **87**(10), 2094–2114 (2017)
2. Anan, O., Böhning, D., Maruotti, A.: Population size estimation and heterogeneity in capture–recapture data: a linear regression estimator based on the Conway–Maxwell–Poisson distribution. Stat. Methods Appl. **26**(1), 49–79 (2017)
3. Bales, K., Murphy, L.T., Silverman, B.W.: How many trafficked people are there in Greater New Orleans? Lessons in measurement. J. Hum. Traffick. **6**(4), 375–387 (2020). https://doi.org/10.1080/23322705.2019.1634936
4. Bell, D.C., Belli-McQueen, B., Haider, A.: Partner naming and forgetting: recall of network members. Soc. Netw. **29**(2), 279–299 (2007)
5. Böhning, D.: Ratio plot and ratio regression with applications to social and medical sciences. Stat. Sci. **31**(2), 205–218 (2016)
6. Böhning, D., van der Heijden, P.G.: The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain. Ann. Appl. Stat. **13**(2), 1198–1211 (2019)

7. Böhning, D., Punyapornwithaya, V.: The geometric distribution, the ratio plot under the null and the burden of dengue fever in Chiang Mai province. In: Böhning, D., van der Heijden, P.G.M., Bunge, J. (eds.) Capture–recapture methods for the social and medical sciences, pp. 55–60. CRC Press, Boca Raton (2018)
8. Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W., et al.: Estimating the number of drug users in Bangkok 2001: a capture–recapture approach using repeated entries in one list. Eur. J. Epidemiol. **19**(12), 1075–1083 (2004)
9. Böhning, D., Baksh, M.F., Lerdsuwansri, R., et al.: Use of the ratio plot in capture–recapture estimation. J. Comput. Graph. Stat. **22**(1), 135–155 (2013)
10. Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., et al.: A generalization of Chao's estimator for covariate information. Biometrics **69**(4), 1033–1042 (2013)
11. Böhning, D., Bunge, J., van der Heijden, P.G.M.: Basic concepts of capture–recapture. In: Böhning, D., van der Heijden, P.G.M., Bunge, J. (eds.) Capture–recapture methods for the social and medical sciences, pp. 3–17. CRC Press, Boca Raton (2018)
12. Böhning, D., Kaskasamkul, P., van der Heijden, P.G.M.: A modification of Chao's lower bound estimator in the case of one-inflation. Metrika **82**(3), 361–384 (2019)
13. Böhning, D., Rocchetti, I., Maruotti, A., et al.: Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods. Int. J. Infect. Dis. **97**, 197–201 (2020)
14. Brenner, H.: Use and limitations of the capture–recapture method in disease monitoring with two dependent sources. Epidemiology **6**(1), 42–48 (1995)
15. Brewer, D.D.: Forgetting in the recall-based elicitation of personal and social networks. Soc. Netw. **22**(1), 29–43 (2000)
16. Buchanan, R., Meskarian, R., van der Heijden, P.G.M., et al.: Prioritising hepatitis C treatment in people with multiple injecting partners maximises prevention: a real-world network study. J. Infect. **80**(2), 225–231 (2020)
17. Buckland, S.T., Garthwaite, P.H.: Quantifying precision of mark-recapture estimates using the bootstrap and related methods. Biometrics **47**(1), 255–268 (1991)
18. Chao, A.: Estimating the population size for capture–recapture data with unequal catchability. Biometrics **43**(4), 783–791 (1987)
19. Coumans, A.M., Cruyff, M., van der Heijden, P.G.M., et al.: Estimating homelessness in the Netherlands using a capture–recapture approach. Soc. Indic. Res. **130**(1), 189–212 (2017)
20. David, B., Snijders, T.A.B.: Estimating the size of the homeless population in Budapest, Hungary. Qual. Quant. **36**(3), 291–303 (2002)
21. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. **39**(1), 1–38 (1977)
22. Dombrowski, K., Khan, B., Wendel, T., et al.: Estimating the size of the methamphetamine-using population in New York City using network sampling techniques. Adv. Appl. Sociol. **2**(4), 245–252 (2012)
23. Doshi, R.H., Apodaca, K., Ogwal, M., et al.: Estimating the size of key populations in Kampala, Uganda: 3-source capture–recapture study. JMIR Public Health Surveill. **5**(3) (2019). https://doi.org/10.2196/12118. Erratum in: JMIR Public Health Surveill. **6**(2) (2020). https://dx.doi.org/10.2196/19893
24. Ezoe, S., Morooka, T., Noda, T., et al.: Population size estimation of men who have sex with men through the network scale-up method in Japan. PLoS ONE (2012). https://doi.org/10.1371/journal.pone.0031184
25. Farcomeni, A., Scacciatelli, D.: Heterogeneity and behavioural response in continuous time capture–recapture, with application to street cannabis use in Italy. Ann. Appl. Stat. **7**(4), 2293–2314 (2013). https://doi.org/10.1214/13-AOAS672
26. Fellows, I.E.: Estimating population size from a privatized network sample. J. Surv. Stat. Methodol. **10**(5), 1346–1369 (2022). https://doi.org/10.1093/jssam/smac010
27. Frank, O., Snijders, T.A.B.: Estimating the size of hidden populations using snowball sampling. J. Off. Stat. **10**(1), 53–67 (1994)
28. Godwin, R.T., Böhning, D.: Estimation of the population size by using the one-inflated positive Poisson model. J. R. Stat. Soc. Ser. C Appl. Stat. **66**(2), 425–448 (2017)
29. Good, I.J.: The population frequencies of species and the estimation of population parameters. Biometrika **40**(3–4), 237–264 (1953)
30. Handcock, M.S., Gile, K.J., Kim, B.J., et al.: sspse: Estimating Hidden Population Size Using Respondent Driven Sampling Data. Los Angeles, CA. R package version 1.0.3 (2022). https://CRAN.R-project.org/package=sspse
31. Handcock, M.S., Hunter, D.R., Butts, C.T., et al.: ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks. The Statnet Project (https://statnet.org). R package version 4.3.2. https://CRAN.R-project.org/package=ergm (2022)

32. Handcock, M.S., Gile, K.J., Mar, C.M.: Estimating hidden population size using respondent-driven sampling data. Electron. J. Stat. **8**(1), 1491–1521 (2014)
33. Handcock, M.S., Gile, K.J., Mar, C.M.: Estimating the size of populations at high risk of HIV using respondent-driven sampling data. Biometrics **71**(1), 258–266 (2015)
34. Heckathorn, D.D.: Respondent-driven sampling: a new approach to the study of hidden populations. Soc. Probl. **44**(2), 174–199 (1997)
35. Hser, Y.-I.: Population estimation of illicit drug users in Los Angeles County. J. Drug Issues **23**(2), 323–334 (1993)
36. Kaskasamkul, P., Böhning, D.: Population size estimation for one-inflated count data based upon the geometric distribution. In: Böhning, D., van der Heijden, P.G.M., Bunge, J. (eds.) Capture–recapture methods for the social and medical sciences, pp. 191–209. CRC Press, Boca Raton (2018)
37. Khan, B., Lee, H.-W., Fellows, I., et al.: One-step estimation of networked population size: respondent-driven capture–recapture with anonymity. PLoS ONE (2018). https://doi.org/10.1371/journal.pone.0195959
38. Kim, B.J., Handcock, M.S.: Population size estimation using multiple respondent-driven sampling surveys. J. Surv. Stat. Methodol. **9**(1), 94–120 (2021)
39. Koskinen, J., Daraganova, G.: Exponential random graph model fundamentals. In: Lusher, D., Koskinen, J., Robins, G. (eds.) Exponential random graph models for social networks: theory, methods and applications, pp. 49–76. Cambridge University Press, Cambridge (2013)
40. Koskinen, J., Snijders, T.: Simulation, estimation and goodness of fit. In: Lusher, D., Koskinen, J., Robins, G. (eds.) Exponential Random Graph Models for Social Networks: Theory, Methods and Applications, pp. 141–166. Cambridge University Press, Cambridge (2013)
41. Lincoln, F.C.: Calculating Waterfowl Abundance on the Basis of Banding Returns. US Department of Agriculture (118) (1930)
42. Nguyen, L.T., Patel, S., Nguyen, N.T., et al.: Population size estimation of female sex workers in Hai Phong, Vietnam: use of three source capture–recapture method. J. Epidemiol. Glob. Health **11**(2), 194–199 (2021)
43. Norris, J.L., III., Pollock, K.H.: Including model uncertainty in estimating variances in multiple capture studies. Environ. Ecol. Stat. **3**(3), 235–244 (1996)
44. Okiria, A.G., Bolo, A., Achut, V., et al.: Novel approaches for estimating female sex worker population size in conflict-affected South Sudan. JMIR Public Health Surveill. (2019). https://doi.org/10.2196/11576
45. Pattison, E.P., Robins, G.L., Snijders, T.A.B., et al.: Conditional estimation of exponential random graph models from snowball sampling designs. J. Math. Psychol. **57**(6), 284–296 (2013)
46. Paz-Bailey, G., Jacobson, J.O., Guardado, M.E., et al.: How many men who have sex with men and female sex workers live in El Salvador? Using respondent-driven sampling and capture–recapture to estimate population sizes. Sex. Transm. Infect. **87**(4), 279–282 (2011)
47. Plettinckx, E., Crawford, F.W., Antoine, J., et al.: Estimates of people who injected drugs within the last 12 months in Belgium based on a capture–recapture and multiplier method. Drug Alcohol Depend. (2021). https://doi.org/10.1016/j.drugalcdep.2020.108436
48. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). https://www.R-project.org/
49. Rocchetti, I., Bunge, J., Böhning, D.: Population size estimation based upon ratios of recapture probabilities. Ann. Appl. Stat. **5**(2), 1512–1533 (2011)
50. Sarría-Santamera, A., Abdukadyrov, N., Glushkova, N., et al.: Towards an accurate estimation of COVID-19 cases in Kazakhstan: back-casting and capture–recapture approaches. Medicina (2022). https://doi.org/10.3390/medicina58020253
51. Shmueli, G., Minka, T.P., Kadane, J.B., et al.: A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. J. R. Stat. Soc. Ser. C Appl. Stat. **54**(1), 127–142 (2005)
52. Snijders, T.A.B., Pattison, P.E., Robins, G.L., et al.: New specifications for exponential random graph models. Sociol. Methodol. **36**(1), 99–153 (2006)
53. Sukrat, B., Okascharoen, C., Rattanasiri, S., et al.: Estimation of the adolescent pregnancy rate in Thailand 2008–2013: an application of capture–recapture method. BMC Pregnancy Childbirth **20**(1), 1 (2020). https://doi.org/10.1186/s12884-020-2808-3
54. Van der Heijden, P.G.M., Cruyff, M.J.L.F., van Houwelingen, H.C.: Estimating the size of a criminal population from police records using the truncated Poisson regression model. Stat. Neerl. **57**(3), 289–304 (2003)
55. Xi, L., Watson, R., Yip, P.S.F.: The minimum capture proportion for reliable estimation in capture–recapture models. Biometrics **64**(1), 242–249 (2008)
56. Zwane, E.N., van der Heijden, P.G.M.: Implementing the parametric bootstrap in capture–recapture models with continuous covariates. Stat. Probab. Lett. **65**(2), 121–125 (2003)