Social Networks 52 (2018) 48–55

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Assessing respondent-driven sampling: A simulation study across different networks

Sandro Sperandei[a,b,*], Leonardo Soares Bastos[b], Marcelo Ribeiro-Alves[c], Francisco Inácio Bastos[a]

[a] Institute of Scientific and Technological Communication & Information in Health, Oswaldo Cruz Foundation, Brazil
[b] Scientific Computational Program, Oswaldo Cruz Foundation (FIOCRUZ), Brazil
[c] National Institute of Infectious Diseases Evandro Chagas, Oswaldo Cruz Foundation (FIOCRUZ), Brazil

ARTICLE INFO

Article history:
Available online 4 June 2017

Keywords:
Respondent-driven sampling
Hidden population
Hard-to-reach population
Simulation method
Random graph
Epidemiology

ABSTRACT

The purpose was to assess RDS estimators in populations simulated with diverse connectivity characteristics, incorporating the putative influence of misreported degrees and transmission processes. Four populations were simulated using different random graph models. Each population was "infected" using four different transmission processes. From each combination of population x transmission, one thousand samples were obtained using a RDS-like sampling strategy. Three estimators were used to predict the population-level prevalence of the "infection". Several types of misreported degrees were simulated. Also, samples were generated using the standard random sampling method and the respective prevalence estimates, using the classical frequentist estimator. Estimation biases in relation to population parameters were assessed, as well as the variance. Variability was associated with the connectivity characteristics of each simulated population. Clustered populations yield greater variability and no RDS-based strategy could address the estimation biases. Misreporting degrees had modest effects, especially when RDS estimators were used. The best results for RDS-based samples were observed when the "infection" was randomly attributed, without any relation with the underlying network structure.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Most hard-to-reach populations are marginalized, stigmatized and—depending on mores and laws—may be criminalized. Men who have sex with men (MSM), drug users, migrants belonging to ethnic/linguistic/religious minorities, people living with HIV/AIDS are some examples of these populations. Even when their members are relatively numerous in a given setting (for instance, neighborhoods where migrants from a given ethnicity cluster), it is difficult or rather impossible to use traditional sampling methods to assess them (Johnston et al., 2016; Montealegre et al., 2012a).

Such populations/groups are not easily identifiable, tend to conceal their status to protect them from actual or perceived prejudice and to avoid interactions with institutions and/or people who may be viewed as sources of additional difficulties and stigma (but see Montealegre et al., 2012b; respecting successfully HIV testing strategies for undocumented immigrant in Houston, Texas, USA, despite relevant differential rates according to education, country of origin, etc.).

Low frequencies of a given characteristic behavior and/or geographic dispersal worsens the problem because even if individuals may be candid and prone to reveal their status and habits, a large sample size and complex, costly logistics would be required to find a reasonable number of individuals (Heckathorn, 1997; Salganik and Heckathorn, 2004). Examples of such difficulties (having as a key consequence the violation of basic assumptions of random selection, an essential feature of any unbiased sampling strategy) have been documented by studies targeting rural populations, even in high-income countries (e.g., USA) where good transportation and sound infrastructure partially alleviate such hurdles and caveats (Young et al., 2014).

Currently, one of the most popular sampling technique used to assess hard-to-reach populations is respondent-driven sampling (RDS) (Heckathorn, 1997). Since the late 1990′s, its application have mushroomed and it has already proven to be efficient in finding members of several hard-to-reach populations. The recommendation and adoption of RDS by major agencies such as the Centers for the Disease Control and Prevention (CDC) (Lansky et al., 2007) and the World Health Organization (WHO) (Johnston et al., 2013)

* Corresponding author at: Rua Ferreira de Andrade, 583-202, Rio de Janeiro, RJ 20780-200, Brazil.
E-mail address: ssperandei@gmail.com (S. Sperandei).

http://dx.doi.org/10.1016/j.socnet.2017.05.004
0378-8733/© 2017 Elsevier B.V. All rights reserved.

have fostered its acceptance and widespread use (Salganik and Heckathorn, 2004).

However, although RDS is able to recruit members from a hard-to-reach population, estimates based on RDS studies remain a matter of concern and debate. Clearly, RDS is a chain-referral, non-probabilistic, sampling method, similar to snowballing (Goodman, 1961; Heckathorn, 2011), and prevalence estimates based on RDS data may be biased (Goel and Salganik, 2010). As a chain-referral method, sampling results are intrinsically dependent on the underlying network structure of the population under analysis, as well as on several other factors, such as the differential recruitment of specific subgroups, geographic heterogeneities, structural bottlenecks secondary to violence or lack of transportation, less-than-optimal bridging between different segments etc. (see, for instance, Burt and Thiede, 2014; Rudolph et al., 2015; Toledo et al., 2011).

The assessment of the accuracy and validity of RDS estimates remains a challenge, since it is very difficult (or rather impossible) to know the actual contact network of each individual. Usually, the reported number of contacts is used to weight the individual information when calculating prevalence of a given characteristic or medical condition (Gile et al., 2015; Goel and Salganik, 2010).

Since the actual contact network of each individual is unknown, simulating connected populations seems to be a valid strategy to evaluate assumptions which are key to the method, as well as their putative violations when estimators are based on studies carried out in real-life situations. Some studies have assessed the accuracy and validity of standard estimators using simulated data, profiting from actual information on degree distributions (Goel and Salganik, 2010; McCreesh et al., 2012; Mills et al., 2014; Wejnert, 2009). However, one must be keep in mind that the number of contacts in common between any two individuals is hard to assess or is unknown, and there is little, if any, information about it. Even assuming that information from two individuals about their total number of contacts are precise, the extent such contacts may overlap is usually hard or impossible to estimate in real-life conditions.

Another possible relevant source of estimation error from RDS sampling is due to the dependency between the putative transmission of a given pathogen (or any other transmissible element) and the underlying network structure of the population. For instance, the transmission of some pathogens depend on close and prolonged contact between infected and at-risk individuals (e.g., HIV/AIDS and other sexually transmitted infections/diseases), whereas other conditions are less dependent on the network structure and can be transmitted even if individuals' interaction is incidental, such as in the spread of influenza virus via the shared use of public transportation.

To the best of our knowledge, a single study has addressed the impact of information error about the number of contacts on RDS estimators. Mills et al. (2014) have shown information error may determine relevant estimation biases on RDS studies.

In the present paper, we assessed RDS estimators' performance under varying conditions of network structure, misreporting degrees, and transmission dependency.

2. Material and methods

2.1. Simulated populations

Four different populations (N = 10,000) were simulated, each using a different approach based on different families of random graph models: Erdös-Renyi (ER – Erdös and Rényi, 1959), Watts-Strogatz (WS – Watts and Strogatz, 1998), Barabasi-Albert (BA – Barabasi and Albert, 1999) and Interconnected Islands (II). For the sake of the present study, only static network have been considered. Some information about the connectivity characteristics of each model used is provided as follows:

- Erdös-Renyi (ER): the connection between two individuals is established in a completely random fashion and any two individuals will be connected with a fixed probability. The only parameter needed is the probability (P) of a link between two individuals, set at 0.001;
- Watts-Strogatz (WS): starting from a regular ring lattice, an individual will be linked to a fixed number of neighbors at each side. Here, we set this parameter to five neighbors to each side. Then, each link has a certain probability to be broken and reattached to any other individual in the population, creating "shortcuts" between groups of individuals, which was set at 0.1 in our model. This model is usually known as the *small-word* model;
- Barabasi-Albert (BA): known as the *preferential attachment* model, this model starts with one individual and adds other individuals, one by one. Each entering individual will be preferentially attached to individuals with a higher number of contacts (usually mentioned as a "rich get richer" attachment strategy). The parameter to this model is the number of connections each new member of the population will add when created and was set at five in the simulation;
- Interconected-Islands (II): the original population is initially split into a number of subpopulations (five, in our simulations). Within each subpopulation, the connectivity is determined as in the ER model and a random set of individuals in each subpopulation is chosen connecting individuals from other subpopulations. In our simulation, we set five connecting individuals, which represents a highly clustered population. The third parameter needed is the probability of a random connection between individuals, as in the ER model before and was set at 0.005.

All model parameters were set to obtain a mean degree of 10 connections, irrespectively of the model used.

2.2. Disease transmission process

Each population was challenged by four transmission processes, all of them dependent on the underlying network connections. Different numbers of infection seedings (10, 100, 500, and 1500 seeds) were randomly selected and launched to transmit the condition (to "infect") to their contacts. Following a Susceptible-Infected (SI) model (which does not consider recovery as a plausible outcome), and taking HIV/AIDS as our key example, infection was spread in the population step by step. In each step, an individual connected to an infected contact had a probability of 0.05 to become infected. The infection process follows until a theoretical prevalence of ~15% "infected" individuals, which is defined here as a theoretical ¨ ceiling value". Clearly, the greater the number of infection seedings, the lesser the dependency between infection dynamic and network connections (i.e., infections spread by a huge number of infection seedings could not be distinguished from a simple "mass effect" dissemination process, where the underlying network structure is not taken into consideration). In our simulation, the ceiling value was a 1,500 seeds infectious process, where no relationship between the condition and the underlying network of contacts was made evident, and the infection can be approximately described as "randomly assigned" (data not shown). Information about individual degree was purposely "biased" in several ways, to simulate different types of misreported degree. Besides "no bias", i.e., a hypothetically perfectly accurate degree information, which corresponds to actual population data, we defined the alternatives as follows: "random misreporting", where the degree information was extracted from normally distributed data, with coefficients of variation either equal to 0.2 or 0.6; and, "systematic misreporting",

where simulated degrees were rounded up to the "next five" number (e.g., degrees 41 or 44 would be rounded to 45; here represented as "5+"), rounded up to the "next 10" (e.g., degree 41 or 44 would be rounded to 50; here represented as "10+"), or rounded to the nearest 10 or 100 (to those degree above 100; here represented as "100"). For all misreported degree under analysis, all other simulation parameters were kept fixed and the degree information has been deliberately altered.

2.3. Sampling from each population

For each combination of "network population model" + "transmission process", one thousand simulations (each one with n = 500 individuals) were obtained using a "RDS (chain-referral recruiting) process". The RDS process was always launched with three random seeds. Then, each seed randomly recruited one to three of its contacts. The probability to recruit one, two or three contacts were set as 0.18, 0.18 and 0.64, respectively, based on empirical data from drug users in the city of Recife, BR (unpublished data). The process was continued until the total sample size were obtained. Although the literature has clearly shown that non-random recruitment may introduce biases in RDS-based studies (Gile et al., 2015), no homophily-related bias was explicitly incorporated into the present simulations.

Random samples were generated for all populations and transmission processes, and used as "yardsticks" to cross-compare results generated by RDS simulations.

2.4. Estimators

Prevalences were estimated based on three estimators. For RDS simulations, the prevalence was first assessed using the classic prevalence estimator (thereafter called "naïve estimator"), which is obtained by simply dividing the number of infected individuals in a sample by the total sample size. Such naïve estimator was cross-compared with both RDS-I estimator (Heckathorn, 1997; Heckathorn, 2002; Salganik and Heckathorn, 2004) and RDS-II estimator (Volz and Heckathorn, 2008), which use the individual degree information to weight the prevalence estimates. RDS-I and RDS-II estimates were calculated using each one of the degree information described before.

The following assumptions by Goel and Salganik's previous simulations study (2010) were used for the sake of our own analysis: (i) symmetric relationships (i.e., if "A is a contact of B, B is a contact of A"); (ii) recruitment is randomly performed within the network contacts; (iii) recruited individuals always take part in the sample; and, (iv) the number of recruits do not depend on the size of the recruiter's network.

2.5. Outputs

For each combination of "population", "transmission" and "degree information" alternative, the mean and variance of the estimation bias (the difference between the estimated and actual population parameter) were observed.

All simulations and analyses were performed using software R (R Core Team, 2015) and the igraph package (Csardi and Nepusz, 2006).

3. Results

Fig. 1 presents the four simulated populations with 10,000 individuals each. It can be observed the similarity between WS and ER simulated populations, the more organized distribution in BA population and the very different pattern of II population, in which only few individuals have links across different subgroups.

Fig. 2 depicts the degree distribution of each simulated population and information bias. All four populations presented a mean degree of 10, as expected, when there were no information bias about individual degrees, however they may present higher mean degrees when systematic biases were added. The BA model yielded a population with a very different degree distribution, with most individuals presenting a very low number of contacts, whereas few individuals presented a very high connectivity (up to 1,800 contacts). The similarity between ER and II degree distributions is due to the fact II connectivity was defined based on the ER model.

Fig. 3 presents the box plots for the estimators, as applied to the different simulations. Each box plot was generated by one-thousand replications and it is related to one of the graph structures (Erdös-Renyi, Barabasi-Albert, Watts-Strogatz, and Interconnected Islands). The figure depicts the differences between each estimate and the actual prevalence from the reference population. Each row of sub-figures is related to one of the estimators under analysis, while each column is related to the number of infection seedings used in the dissemination (infection) process. Fig. 3 shows information about degree, but did not incorporate the effect of the before-mentioned biases. The complete (inevitably overburdened) figures for each graph structure can be found in the supplementary material (Web appendix). Some preliminary remarks about the simulations' findings are presented as follows.

Overall findings related to information biases, for different populations and transmission processes are illustrated in the Supplementary Figs. (1–4). In general, simple random sample is the "benchmark" (standard) model. Such models were not affected by the number of infection seedings used in the simulations of the transmission process, as expected (bottom line). When 1500 infection seedings were used in the dissemination process (right column), the results were very close to those obtained by random sampling.

It is possible to note that estimates from II-structured simulations show a greater dispersion of the estimates when 10 or 100 infection seedings were used, irrespectively of the estimators under analysis. BA-structured simulations yield a slightly lower dispersion. However, when the "naïve" estimator is chosen, the BA-structured simulations show a greater estimation bias (top line) vis-à-vis the other simulations.

Overall, the performance of the two RDS estimators are strongly associated with the nature of the graph model chosen to simulate the population and with the number of infection seedings used to disseminate the disease in the population. These variables seem to be more important than the information biases respecting the degree information, even when the mean degree of every simulated population are about the same and when the information bias equals half of the mean population degree (for instance, situations like 5+ and 10+).

4. Discussion

Simulation studies have a pivotal role in the assessment of complex statistical methods, especially when it is difficult (or rather impossible) to optimally standardize operations and procedures in real-life conditions. Such difficulties or impossibilities may be secondary to budgetary restrictions, operations issues, ethical conundrums, among other factors.

Obviously, the usefulness and accuracy of simulation studies depend on the soundness of methods, procedures and choice of parameters. Although "*in silico*" simulations cannot (at least, given the current status of computer science) fully emulate the complexity and subtleness of the real world, *in silico* models should always attempt to be as comprehensive and accurate as possible (Cioffi-Revilla, 2014).

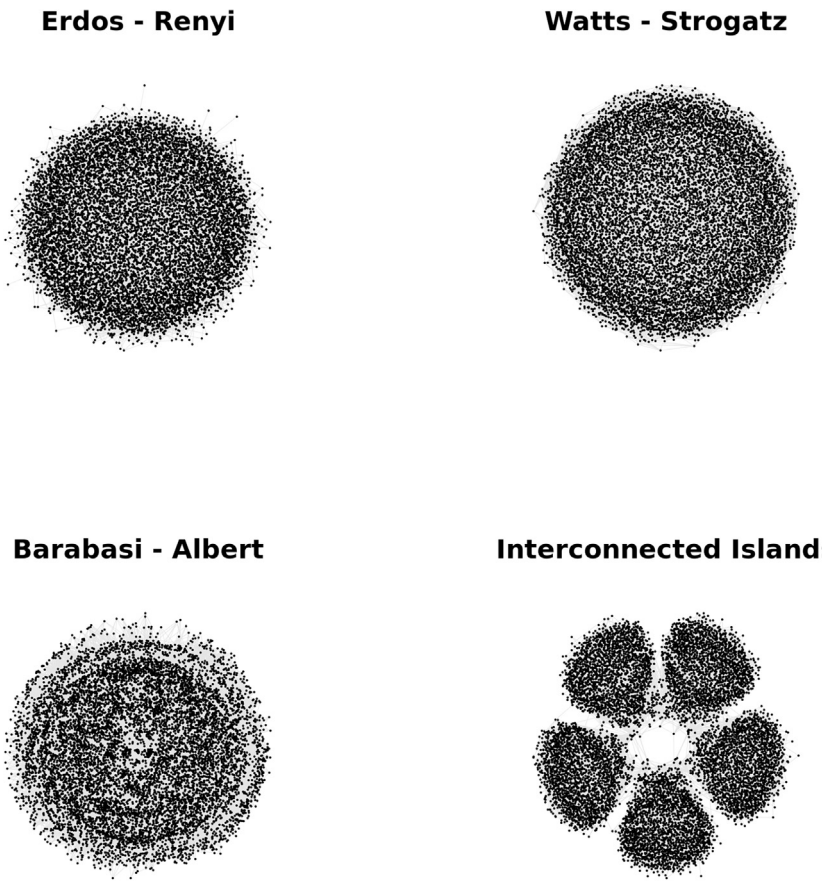

Fig. 1. Graphical representation of the simulated populations (N = 10,000).

Despite the forceful exclusion of some elements, our simulations successfully emulated the expected prevalence of a hypothetical infection in the contexts of different structures of connectivity between individuals and different transmission dynamics. The key characteristic of such models is flexibility since the detailed structure of connectivity between individuals in the context of large populations remains an elusive goal, especially considering its very dynamic nature.

Our simulation platform may help to better understand the operational characteristics and the performance of RDS under different conditions and can be used to assess the accuracy of its standard estimators, as well as to cross-compare the accuracy of RDS vis-à-vis other methodological strategies that have been used for hard-to-reach populations, such as Time Location Sampling. Recent papers have compared the two methods under different conditions and targeting several populations (Paz-Bailey et al., 2013; Tran et al., 2015; Zhao et al., 2015), and such cross-comparative studies (both empirical and *in silico*) seem to be a promising field of study, given the relevance of several hard-to-reach populations for policy making and the absence of a gold standard method to assess them.

The performance of the estimators under random sampling simulations speaks in favor of the good quality of the simulation platform, with a good match between expected and observed values for the parameters, without any discernible estimation bias and low dispersion.

Several authors have addressed the relevance of misreporting degrees to the performance of RDS estimators (Gile et al., 2015; Goel and Salganik, 2009; Wejnert, 2009), but to the best of our knowledge, only Mills et al. (2014) and Lu et al. (2012) cross-compared the influence of different degree distributions on the accuracy of estimates. However, in Mills et al.'s study connectivity was always modeled after a random distribution, that may or may not correspond to specific populations and contexts. Notwithstanding, Mills et al. (2014) have consistently found that information bias have a more relevant effect when distributions had heavier tails. Lu et al. (2012), on the other hand, reported effects on bias almost exclusively bellow 0.02 or 2% in prevalence.

In our study, four different connectivity structures were assessed, varying from random links (Erdös-Renyi model) to a heavy tail distribution (Barabasi-Albert model). The relevance of such different connectivity structures can be visualized in Fig. 2.

Other studies that assessed the performance of RDS estimators (Goel and Salganik, 2010; Wejnert, 2009) profit from simulations based on a hypothetically known network. In this sense, the capacity of a randomly-generated network, based on the distribution of degrees to reproduce real world contact networks is unknown.

A key aim of our study was to comprehensively assess the hypothetical influence of misreporting degrees on estimates derived from RDS samples. In this sense, our findings were auspicious since the impact of misreporting degrees was found to be relatively modest, compared, for instance, with the more pronounced influence of some underlying network structure. The addition of a random information error was found to be associated with modest increments of dispersion. Even such modest increments have emerged only as a consequence of errors of a great magnitude ($\geq$60% of the reference [proper] information). Information bias was particularly more pronounced among BA-based simulations, as previously demonstrated by Mills et al. (2014). Moreover, misreporting degrees showed effect only with Barabasi-Albert and Erdös-Renyi networks. In the first case, it can be related to the

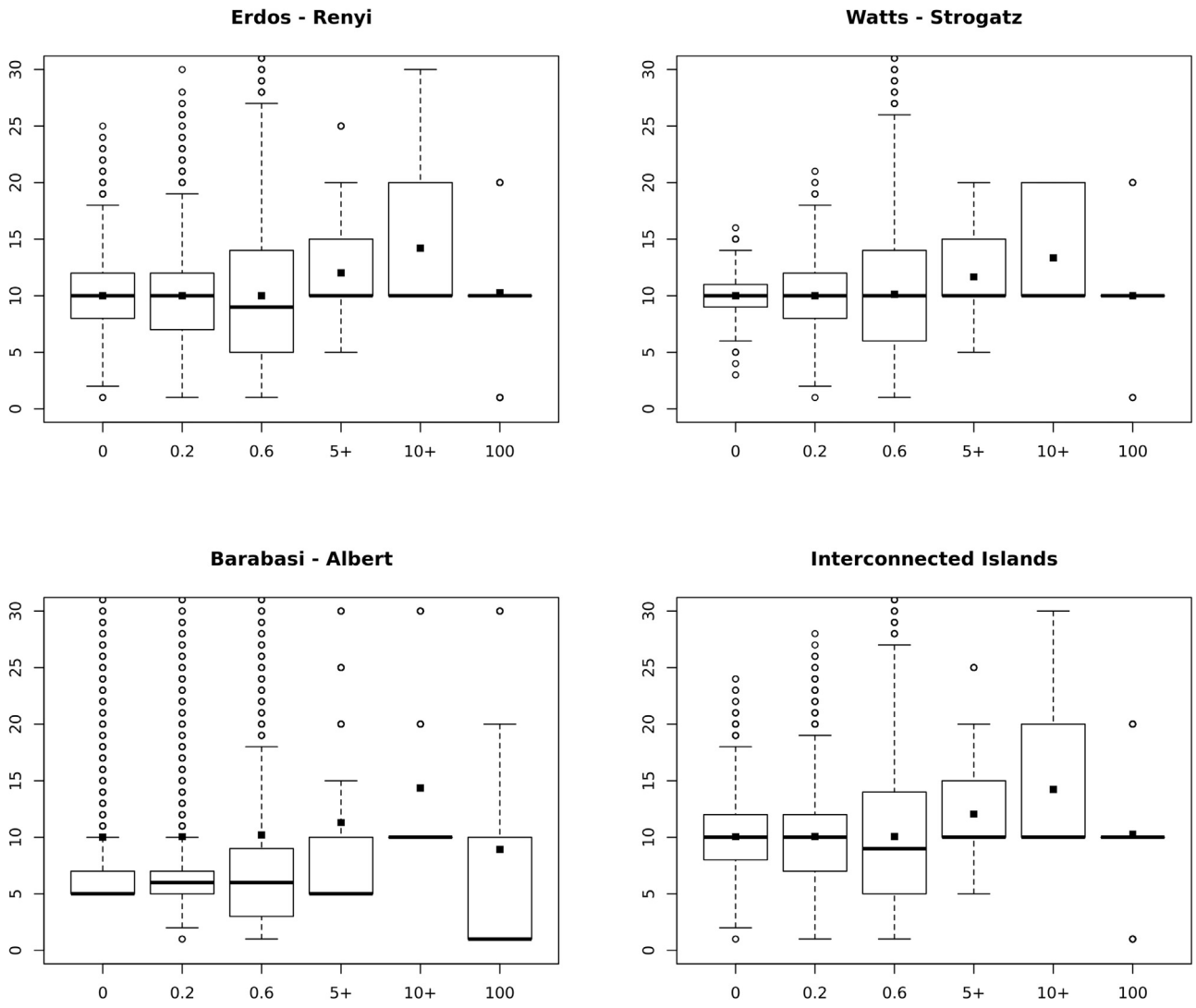

Fig. 2. Degree distribution of the simulated populations according to the information biases. 0: no bias; 0.2: random bias proportional to 20% of the real degree; 0.6: random bias proportional to 20% of the real degree; 5+: degree rounded to the "next five"; 10+: degree rounded to the "next ten"; 100: degrees until 100 are rounded to the nearest 10 and degrees above 100 are rounded to the nearest 100. Squares indicate mean degree.

greater dispersion of degrees. In the second, it happened only at the "100" simulated case.

Systematic errors have also a modest influence on estimates: even an addition of 50% of the mean degree of a given population (+5 with a mean degree of 10) has not been associated with a relevant change of the estimates. Just substantial rounding affecting degrees greater than 100 were found to be associated with relevant changes in most populations under analysis, especially among BA-generated simulations, i.e. those with a heavier tail. Our findings differ here from Mills et al.' findings (2014), whom have observed a much larger effect of the same information biases under analysis in our paper. We believe that the structure of the contacts could be responsible for this difference, since Mills et al.'s simulations were generated randomly from degree distributions, with no control about the structure of the contacts.

Regarding transmission process, a surprising result was that, in the case of completely random infection, no bias was registered and was the best performance for all estimators. As the RDS method depends on the underlying network structure, the condition of interest among individuals are interdependent, and this interdependence is reflected in the poor prevalence estimates. This result is important for studies with hard-to-reach populations where the condition of interest is not so strongly associated with the contact network, since the RDS method is simpler, easier and cheaper than classic sampling strategies (actually, as mentioned before, classic probability sampling methods can be simply impossible to apply to some populations). Similar findings were reported by Mills et al. (2014). Interconnected Islands model represented a particular case, where the standard deviation of the estimates showed a marked decrease as the number of infection seeds increase. Given the highly clustered pattern of the population, an infection process less dependent of contact network will decrease the chance of the disease being trapped in one subpopulation and the sample trapped in another.

In our simulation, the effect of network connectivity can be noted when comparing the four supplemental figures. In this sense, it is important to stress II networks and its effect on the estimates. This network model creates loosely connected to each other subpopulations. Such effect can be observed when working with populations living in territories dominated by rival factions such as discussed by Toledo et al. (2011) or biases by differential locations vis-a-vis recruitment centres (McCreesh et al., 2011). The extent this translate an homophily-related bias or other constraints, such as clustering, accessibility, etc remains open, as discussed by Rocha

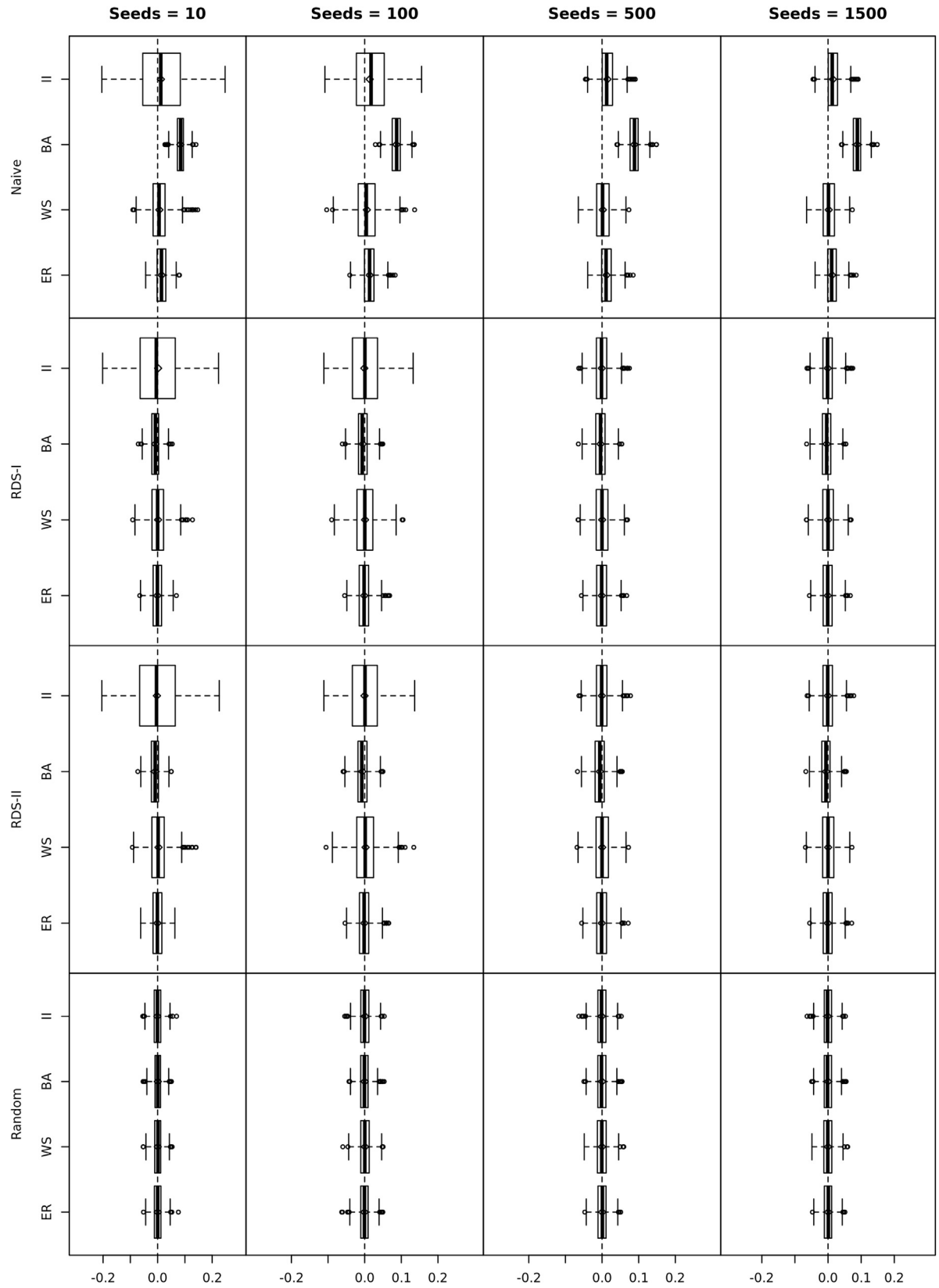

Fig. 3. Distribution of the biases in the estimation according to population, infection seedings and estimator used. The figure depicts the differences between each estimate and the actual prevalence from the reference population. Only the result to the "no information bias" condition is shown.

et al. (2017). In II networks, the low connectivity among subpopulations caused that not only the simulated infection often stays restricted to a single subpopulation, but also a selected sample belongs entirely only to one subpopulation. In a scenario where the sample was restricted to a subpopulation and infection restricted to a different sub-population, the prevalence is underestimated.

Analogously, when sample and infection were confined within the same subpopulation, the estimate prevalence of infection is much higher than its true value in the population. This simulation can be useful in the development and improvement of estimators able to handle this type of problem.

Rocha et al. (2017) assessed this clustering effect on RDS estimates in a series of simulation experiments. The authors reported association between the performance of the estimators and the refusal rate of the recruited individuals, variable not controlled in our work. According to the authors, a good response rate is able to cancel the clustering effect. However, their experiment did not control the sample size used, which can also explain the differences in relation to current results.

The present simulation study is limited by the fact homophily is not explicitly taken in consideration despite the literature shown may be an important source of bias in RDS studies. Gile and Handcock (2010) simulated a scenario were infected persons would be 20% more likely to be recruited and, not surprisingly, it produced a bias in the estimate toward greater prevalences, averaging about 5% more. Also, Lu et al. (2012) simulated homophily, but using the activity between peers as the strength of probability of recruitment. They reported an effect they classified as “moderate”, up to 3%. Compared to the effects of the transmission process and network structure presented here, this magnitude of effect can be considered small. However, more work is needed to investigate the effect of homophily in RDS studies.

Limitations are also secondary to the fact our study is based on static networks. A former study by Boily et al. (2004) has clearly demonstrated that social networks might be dynamic. But such dynamics is under dependence of antiretroviral therapy availability and coverage, for instance. A third limitation refers to the absence of simulation of different types of bottlenecks.

5. Conclusion

The estimators RDS I and RDS II show to be robust to the degree information bias in almost all simulated scenarios. Even where the degree information bias affected the estimate, the RDS estimators performed better than the naïve estimator that ignores the sampling design. In any case the RDS estimates performed better than the naïve estimator under simple random sampling. However, the practical aspects associated with RDS sampling seem to be sufficient to justify its use in hard-to-reach populations. The presence of population clusters appears to compromise the performance of these estimators, what should be evaluated by further simulations. Finally, the RDS sampling seems to be an effective way to obtain valid estimates of outcomes that are not associated with the network contacts of individuals.

Acknowledgements

SS is the recipient of a postdoctoral scholarship from Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.socnet.2017.05.004.

References

Barabasi, A., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Boily, M.-C., Bastos, F.I., Desai, K., Mâsse, B., 2004. Changes in the transmission dynamics of the HIV epidemic after the wide-scale use of antiretroviral therapy could explain increases in sexually transmitted infections: results from mathematical models. Sex Transm. Dis. 31, 100–113, http://dx.doi.org/10.1097/01.OLQ.0000112721.21285.A2.

Burt, R.D., Thiede, H., 2014. Assessing differences in groups randomized by recruitment chain in a respondent-driven sample of Seattle-area injection drug users. Ann. Epidemiol. 24, 861–867, http://dx.doi.org/10.1016/j.annepidem.2014.09.002 (e14).

Cioffi-Revilla, C., 2014. Introduction to Computational Social Science, Texts in Computer Science. Springer London, London, http://dx.doi.org/10.1007/978-1-4471-5661-1.

Csardi, G., Nepusz, T., 2006. The Igraph Software Package for Complex Network Research. InterJournal Complex Sy, pp. 1695.

Erdös, P., Rényi, A., 1959. On random graphs, I. Publ. Math. 6, 290–297.

Gile, K.J., Handcock, M.S., 2010. Respondent-driven sampling: an assessment of current methodology. Sociol. Methodol. 40, 285–327, http://dx.doi.org/10.1111/j.1467-9531.2010.01223.x.

Gile, K.J., Johnston, L.G., Salganik, M.J., 2015. Diagnostics for respondent-driven sampling. J. R. Stat. Soc. Ser. A (Stat. Soc.) 178, 241–269, http://dx.doi.org/10.1111/rssa.12059.

Goel, S., Salganik, M.J., 2009. Respondent-driven sampling as Markov chain Monte Carlo. Stat. Med. 28, 2202–2229, http://dx.doi.org/10.1002/sim.3613.

Goel, S., Salganik, M.J., 2010. Assessing respondent-driven sampling. Proc. Natl. Acad. Sci. U. S. A. 107, 6743–6747, http://dx.doi.org/10.1073/pnas.1000261107.

Goodman, L.A., 1961. Snowball sampling. Ann. Math. Stat. 32, 148–170, http://dx.doi.org/10.1214/aoms/1177705148.

Heckathorn, D.D., 1997. Respondent-driven sampling: a new approach to the study of hidden populations. Soc. Probl. 44, 174–199, http://dx.doi.org/10.2307/3096941.

Heckathorn, D.D., 2002. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Social Probl. 49, 11–34.

Heckathorn, D.D., 2011. Snowball versus respondent-driven sampling. Sociol. Methodol. 41, 355–366, http://dx.doi.org/10.1111/j.1467-9531.2011.01244.x.

Johnston, L.G., Chen, Y.-H., Silva-Santisteban, A., Raymond, H.F., 2013. An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. AIDS Behav. 17, 2202–2210, http://dx.doi.org/10.1007/s10461-012-0394-8.

Johnston, L., Oumzil, H., El Rhilani, H., Latifi, A., Bennani, A., Alami, K., 2016. Sex differences in HIV prevalence, behavioral risks and prevention needs among anglophone and francophone sub-Saharan african migrants living in rabat, Morocco. AIDS Behav. 20, 746–753, http://dx.doi.org/10.1007/s10461-015-1115-x.

Lansky, A., Abdul-Quader, A.S., Cribbin, M., Hall, T., Finlayson, T.J., Garfein, R.S., Lin, L.S., Sullivan, P.S., 2007. Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System. Public Health Rep. 122 (Suppl), 48–55.

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. Ser. A (Stat. Soc.) 175, 191–216, http://dx.doi.org/10.1111/j.1467-985X.2011.00711.x.

McCreesh, N., Johnston, L.G., Copas, A., Sonnenberg, P., Seeley, J., Hayes, R.J., Frost, S.D.W., White, R.G., 2011. Evaluation of the role of location and distance in recruitment in respondent-driven sampling. Int. J. Health Geogr. 10, 56, http://dx.doi.org/10.1186/1476-072X-10-56.

McCreesh, N., Frost, S.D.W., Seeley, J., Katongole, J., Tarsh, M.N., Ndunguse, R., Jichi, F., Lunel, N.L., Maher, D., Johnston, L.G., Sonnenberg, P., Copas, A.J., Hayes, R.J., White, R.G., 2012. Evaluation of respondent-driven sampling. Epidemiology 23, 138–147, http://dx.doi.org/10.1097/EDE.0b013e31823ac17c.

Mills, H.L., Johnson, S., Hickman, M., Jones, N.S., Colijn, C., 2014. Errors in reported degrees and respondent driven sampling: implications for bias. Drug Alcohol Depend. 142, 120–126, http://dx.doi.org/10.1016/j.drugalcdep.2014.06.015.

Montealegre, J.R., Risser, J.M., Selwyn, B.J., McCurdy, S.A., Sabin, K., 2012a. Prevalence of HIV risk behaviors among undocumented Central American immigrant women in Houston, Texas. AIDS Behav. 16, 1641–1648, http://dx.doi.org/10.1007/s10461-011-0130-9.

Montealegre, J.R., Risser, J.M., Selwyn, B.J., Sabin, K., McCurdy, S.A., 2012b. HIV testing behaviors among undocumented Central American immigrant women in Houston, Texas. J. Immigr. Minor. Health 14, 116–123, http://dx.doi.org/10.1007/s10903-011-9534-x.

Paz-Bailey, G., Miller, W., Shiraishi, R.W., Jacobson, J.O., Abimbola, T.O., Chen, S.Y., 2013. Reaching men who have sex with men: a comparison of respondent-driven sampling and time-location sampling in Guatemala city. AIDS Behav., http://dx.doi.org/10.1007/s10461-013-0589-7.

R Core Team, 2015. A Language and Enviroment for Statistical Computing.

Rocha, L.E.C., Thorson, A.E., Lambiotte, R., Liljeros, F., 2017. Respondent-driven sampling bias induced by community structure and response rates in social networks. J. R. Stat. Soc. Ser. A (Stat. Soc.) 180 (1), 99–118, http://dx.doi.org/10.1111/rssa.12180.

Rudolph, A.E., Young, A.M., Lewis, C.F., 2015. Assessing the geographic coverage and spatial clustering of illicit drug users recruited through respondent-driven sampling in New York City. J. Urban Health 92, 352–378, http://dx.doi.org/10.1007/s11524-015-9937-4.

Salganik, M.J., Heckathorn, D.D., 2004. Sampling and estimation in hidden populations using respondent-driven sampling. Sociol. Methodol. 34, 193–240, http://dx.doi.org/10.1111/j.0081-1750.2004.00152.x.

Toledo, L., Codeço, C.T., Bertoni, N., Albuquerque, E., Malta, M., Bastos, F.I., 2011. Putting respondent-driven sampling on the map: insights from Rio de Janeiro, Brazil. J. Acquir. Immune Defic. Syndr. 57, S136–43, http://dx.doi.org/10.1097/QAI.0b013e31821e9981.

Tran, H.V., Le, L.-V.N., Johnston, L.G., Nadol, P., Van Do, A., Tran, H.T.T., Nguyen, T.A., 2015. Sampling males who inject drugs in Haiphong, Vietnam: comparison of time-location and respondent-driven sampling methods. J. Urban Health 92, 744–757, http://dx.doi.org/10.1007/s11524-015-9966-z.

Volz, E., Heckathorn, D.D., 2008. Probability based estimation theory for respondent-driven sampling. J. Off. Stat. 24, 79–97.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. Nature 393, 440–442, http://dx.doi.org/10.1038/30918.

Wejnert, C., 2009. An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. Sociol. Methodol. 39, 73–116, http://dx.doi.org/10.1111/j.1467-9531.2009.01216.x.

Young, A.M., Rudolph, A.E., Quillen, D., Havens, J.R., 2014. Spatial, temporal and relational patterns in respondent-driven sampling: evidence from a social network study of rural drug users. J. Epidemiol. Commun. Health 68, 792–798, http://dx.doi.org/10.1136/jech-2014-203935.

Zhao, J., Cai, R., Chen, L., Cai, W., Yang, Z., Richardus, J.H., de Vlas, S.J., 2015. A comparison between respondent-driven sampling and time-location sampling among men who have sex with men in Shenzhen, China. Arch. Sex. Behav. 44, 2055–2065, http://dx.doi.org/10.1007/s10508-014-0350-y.