OPEN ACCESS Freely available online

PLOS ONE

Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam

Linus Bengtsson[1]*, Xin Lu[1,2], Quoc Cuong Nguyen[3], Martin Camitz[1], Nguyen Le Hoang[4], Tuan Anh Nguyen[5], Fredrik Liljeros[2], Anna Thorson[1]

1 Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden, **2** Department of Sociology, Stockholm University, Stockholm, Sweden, **3** Family Heath International, Hanoi, Vietnam, **4** iSEE, Hanoi, Vietnam, **5** Hanoi Medical University, Hanoi, Vietnam

Abstract

Objective: Lack of representative data about hidden groups, like men who have sex with men (MSM), hinders an evidence-based response to the HIV epidemics. Respondent-driven sampling (RDS) was developed to overcome sampling challenges in studies of populations like MSM for which sampling frames are absent. Internet-based RDS (webRDS) can potentially circumvent limitations of the original RDS method. We aimed to implement and evaluate webRDS among a hidden population.

Methods and Design: This cross-sectional study took place 18 February to 12 April, 2011 among MSM in Vietnam. Inclusion criteria were men, aged 18 and above, who had ever had sex with another man and were living in Vietnam. Participants were invited by an MSM friend, logged in, and answered a survey. Participants could recruit up to four MSM friends. We evaluated the system by its success in generating sustained recruitment and the degree to which the sample compositions stabilized with increasing sample size.

Results: Twenty starting participants generated 676 participants over 24 recruitment waves. Analyses did not show evidence of bias due to ineligible participation. Estimated mean age was 22 years and 82% came from the two large metropolitan areas. 32 out of 63 provinces were represented. The median number of sexual partners during the last six months was two. The sample composition stabilized well for 16 out of 17 variables.

Conclusion: Results indicate that webRDS could be implemented at a low cost among Internet-using MSM in Vietnam. WebRDS may be a promising method for sampling of Internet-using MSM and other hidden groups.

Citation: Bengtsson L, Lu X, Nguyen QC, Camitz M, Hoang NL, et al. (2012) Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam. PLoS ONE 7(11): e49417. doi:10.1371/journal.pone.0049417

Editor: Patricia Kissinger, Tulane University, United States of America

Received July 2, 2012; **Accepted** October 10, 2012; **Published** November 12, 2012

Funding: Funding was provided by the Swedish International Development Agency (Sida). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bengtssonlinus@gmail.com

Introduction

Men who have sex with men (MSM) has emerged as a key population in the global HIV epidemic [1,2,3]. Modeling work on the Asian epidemic points to a scenario in which 42 percent of all new HIV infections in Asia will occur among MSM by 2020 [4]. While population-based surveys in countries with generalized epidemics have generated vast amounts of data on sexual behavior [5], studies on MSM and other hidden populations struggle to generate representative samples and adequate sample sizes [1]. The lack of representative data of MSM risk-behavior severely hinders an understanding of the underlying dynamics of the MSM epidemics and prevents an evidence-based response. New methods for representative sampling of MSM and other hidden groups are thus needed.

Respondent-driven sampling (RDS) was developed to overcome sampling challenges in studies of populations for which a sampling frame is difficult or impossible to define, such as MSM, injecting drug users (IDU), and sex workers (SW) [6,7,8,9,10,11]. An RDS study starts by purposively selecting a handful of participants who are known members of the study population. These ''seeds'' are given invitation coupons (usually three) to distribute to other members of the population. These members are in turn given three new coupons to distribute. Monetary incentives are usually given both for participation and recruitment.

RDS resembles ''snowball sampling'' [12] but differs from it in several important respects. The RDS incentive system and the limited number of invitation coupons per participant allow the creation of long recruitment chains. If the sampling conforms to methodological assumptions, the proportion of the sample with a certain characteristic stabilizes at a level determined by the characteristics of the population, independently of the characteristics of the seeds [7]. Furthermore, snowball sampling systematically oversamples individuals with many contacts. All individual

properties correlated with the number of contacts within the group under study will hence be over or under sampled in a snowball sample. In contrast, during an RDS study researchers record an estimate of each person's social network size and adjust for this bias. Participants are also not required to name or identify their contacts, as is often the case in snowball sampling. Instead participants can pass invitation coupons to any of their contacts at their own discretion. They receive a reward when their contacts participate in the survey, serving to increase participation rates and decrease selection bias. The RDS method has been shown to be analytically unbiased under a limited number of assumptions. Extensive methodological research is ongoing to investigate how well these assumptions are met during real-life implementation, how to best estimate variance and what determines the suitability of the method in the local context [9,13,14,15,16,17,18,19,20].

Although RDS in certain contexts has clear advantages over other sampling methods [21], the standard implementation of the method has several limitations, including: 1) individuals with a behavior that is stigmatized, illegal, or associated with high privacy concerns may be unwilling to access survey offices physically and may thus be underrepresented in the sample; 2) persons from middle- and upper income levels may not be sufficiently incentivized by the study rewards, given the time and effort required to participate; 3) the geographic area of study needs to be small enough to allow participants to travel to the study sites; and 4) RDS studies are, like other field survey methods, relatively expensive since they require the presence of trained staff for extended periods of time and need to be repeated at several sites to generate national or regional data,

Sampling participants though the Internet can mitigate some of these disadvantages by allowing people to participate anonymously and with little effort. Online sampling also allows for vast geographic coverage and may be carried out at markedly lower cost than standard field surveys. However, current methods of Internet-based sampling of hidden groups enroll participants through self-selection, which may cause important bias [22]. Usually, a so-called banner add is put on a web page, e.g. a site for gay men. Persons accessing this site then click the banner to volunteer for the study. These surveys can have participation rates as low as a few in a thousand to a few in a hundred out of registered users [23,24].

Web-based RDS (henceforth webRDS) can potentially circumvent both the disadvantages of standard RDS as well as disadvantages inherent in current Internet-based sampling methods of hidden groups. There are three published webRDS surveys, two involving students at Cornell University [16,25] and one among users of Facebook in the U.S [26]. The results of these studies showed that RDS estimates agreed relatively well with the true characteristics study population with the exception, in the Facebook survey, of undersampling of participants who self-identified as Hispanic/Latino, African American/Black American and were of lower education levels. These surveys did not, however, target a hidden population. We aimed to implement and evaluate webRDS for sampling and surveying of a hidden and stigmatized population, Internet-using MSM in Vietnam.

Materials and Methods

General Study Design

The survey was cross-sectional, performed online and carried out between February 18 and April 12, 2011, applying web-based respondent driven sampling (webRDS).

Inclusion Criteria and Population Delineation

Eligible participants were adult men (18 years and above) who had ever had any type of sex (including oral sex and mutual masturbation) with another man, had not previously participated in the survey, and were living in Vietnam at the time of the study. The Internet-using part of this group formed the population to which the sample aimed to generalize.

MSM and Internet Use in Vietnam

Internet access in Vietnam costs approximately 0.15 USD per hour at Internet cafés. MSM in Vietnam are stigmatized [27], and HIV prevalence in the group has been estimated at 14–20% and 14–16% in Hanoi and Ho Chi Minh City, respectively (2009) [28]. Internet use as a proportion of the population in Vietnam was 27% in 2010 (24 million persons) and 60% and 50%, respectively, in the large urban areas of Hanoi and Ho Chi Minh City [29]. Internet use among MSM in general may be considerably higher than in the general population [30]. Ninety-four percent of MSM in an offline RDS in Hanoi stated that they used the Internet [31]. The Internet in Vietnam provides an important environment in which MSM communicate with each other and meet partners.

Sampling

The study was performed in collaboration with a local research organization in Vietnam working to promote LGBT and ethnic minority rights (iSEE). iSEE has an extensive knowledge and contact network among MSM community groups and a close collaboration with web administrators of Vietnamese LGBT web sites. Fifteen seeds, who were recruited through these networks, initiated the survey and a further five seeds were added two weeks later to increase the speed of recruitment. Six seeds came from Ho Chi Minh City, ten from Hanoi and four from Hoa Binh. Nineteen out of the 20 seeds had attended some kind of education after high school (vocational training, college or university). Participants received, from their recruiter, an invitation message with a login code and a web address. They logged in, accessed detailed information about the study, approved participation and eligibility and answered a written questionnaire. Participants could then compare their own answers to aggregated answers of earlier participants, displayed in informative bar charts. On the last page the participants were encouraged to recruit MSM friends by providing an e-mail or Yahoo! Messenger address (popular for communications in Vietnam), and being automatically sent four invitation messages, which could be forwarded to MSM friends. The messages were also displayed on the screen and could be copied for sending by other preferred means. Text both on the web site and in the email/Yahoo! chat messages emphasized that only MSM living in Vietnam and of age 18 years or above were allowed to participate. A warning was included saying that advanced checks were applied and that failure to follow the recruitment rules would mean loss of compensation. No restriction was given as to whether the recruiter knew each other in real life or only through the Internet. Reminders to recruit were sent out two and four days after completing the survey. Participants were informed that they had seven days to recruit and were given rewards for recruitments that took place during that time. Some participants took the survey at a later time point. They were retained in the sample and the persons they recruited were given standard compensation.

Web Site

The graphic design of the web site aimed at giving a professional and friendly impression without strong MSM connotations.

Incentives and Recruitment Stimuli Included the Following

1) 2.45 USD (50,000 VND) as credit on the participant's SIM card and the same amount for each successful recruitment of an MSM friend (maximum four); 2) the option of donating the monetary reward to an MSM community organization chosen by the participant; 3) a lottery with the possibility of winning an iPad; 4) text emphasizing participation in order to support MSM in Vietnam; and 5) being able to compare one's own answers to those of other participants in simple, informative and anonymous charts. Eight questions were included in the questionnaire specifically to stimulate the participants' interest in comparing themselves with other participants.

Piloting and Early Versions of the System

The web site and recruitment system was extensively pilot tested. Interviews and focus-group discussions among MSM were performed to understand social networks among MSM, online interaction and to decide on appropriate incentives. Two versions of the webRDS site were used for sampling before the study described in this paper was carried out. These webRDS systems differed in that they had a less advanced graphic design and smaller incentives. In the first survey in 2009, recruitment died out after a maximum of 5 waves (25 participants, 15 seeds). The second time, recruitment improved but stopped after 5 waves (84 participants, 15 seeds).

Data Collection

The questionnaire contained 17 questions, including number of sexual partners in the past 6 months, sexual partner preferences (prefer as sexual partners only men, men to women, women to men or only women), the duration of the respondent's longest relationship, opinion on legalizing same-sex marriage in Vietnam (for or against), frequency of Internet-use, socio-demographic characteristics, network size (see separate heading), relationship between the participant and his recruiter (stranger, acquaintance, friend, close friend, lover/ex-lover, or relative), and the social context in which the participant got to know his recruiter. Logical checks with error messages were used for interdependent questions. Only positive integers were allowed for numeric answers. All questions included a ''don't want to answer'' option and all questions needed to be answered. Participants who wanted to receive rewards filled out contact details and a personal identifier (telephone number, email or Yahoo! Messenger address, and the last three digits of their nine-digit ID number). Time points at which each participant loaded the web pages was stored to facilitate identification of ineligible submissions, including unserious attempts to answer the questionnaire or the same person trying to answer more than one questionnaire to receive additional rewards.

Analyses of Duplicated Submissions, Data Cleaning and Analysis

9.6% of completed surveys (65 surveys) included a stated age below 18 years, or a telephone number, e-mail or Yahoo! Chat address that had previously been registered in the system. We defined these as ''invalid''. We excluded seeds (customary in RDS analysis [8]) together with the aforementioned invalid submissions to produce a cleaned sample. From this sample we estimated, in Matlab, population proportions using the current state of the art estimator, RDSII, which requires only information on the sample compositions and the social network sizes of the participants [10]. We have not included confidence intervals in this paper since there is currently no consensus on how to best estimate RDS design effects.

We checked all surveys for other signs of duplication or invalidity by flagging surveys containing a repeated IP number, deviating answers (as described below), or short completion times. We analyzed the sensitivity of the estimates to inclusion and exclusion of these flagged submissions. Specifically we compared the RDS II estimates generated from the full sample of non-seed submissions with valid age with the RDS II estimates generated from groups with progressively stricter inclusion criteria according to the following: 1) exclusion of submissions with a repeated email, Yahoo! Chat ID or telephone number (forming the cleaned sample above); 2) additionally excluding repeated IP numbers; and 3) additionally excluding submissions with short completion times ($<$ three minutes), submissions stating no education (rare in Vietnam), or submission stating six-month partner numbers above 1,000. Differences were small between the groups. Details are included in the supplementary material. For all estimates in the supplementary material the maximum absolute differences when comparing the full sample to the groups with progressively stricter inclusion criteria were 6.6%.

Personal Network Size

We asked participants for the number of MSM they had interacted with in any way during the past seven days (including on the phone, Internet, or in person). We then asked how many of these persons they believed used the Internet. We chose the seven-day timeframe to reflect the potential high frequency of contacts online. We used the second network question to define the participants' personal network. We replaced missing personal network size data with the RDSII-estimated average network size from submissions with non-missing network data. The average network size was 5.5 persons.

Evaluation and Analyzes of Equilibrium

As there is no gold standard by which to validate the sampling, we evaluated the system in terms of its success in generating sustained recruitment, the degree to which the sample compositions stabilized with increasing sample size (independence of the sample from the seeds), and finally, in the discussion, we contrast the sample compositions with results from other surveys.

We analyzed whether equilibrium was achieved in two ways. We first used the standard criterion from the RDS study literature [32]. This criterion requires the sampling process to have reached a certain number of waves. The number of waves required is, for each variable, determined by the number of steps required by a first-order Markov process to reach a less than a two percent relative difference between its value at a given step and its value after an infinite number of steps. The transition probabilities used to calculate the values of the Markov process are the averaged transition probabilities in the study's recruitment chains. Second, we produced plots of the changes in the sample compositions as sample size increased.

Ethics

All information about the survey was available on all web pages except the log-in page and could be accessed at any time. All pages included a log-out button, which automatically removed traces of the survey from the computer and transported the user to a search engine. Browser history has to be deleted manually by the user and participants were given detailed instructions on how to do so. Telephone and chat support were available. IP addresses were converted to a unique anonymous code using a one-way encryption algorithm, and the original IP numbers were deleted.

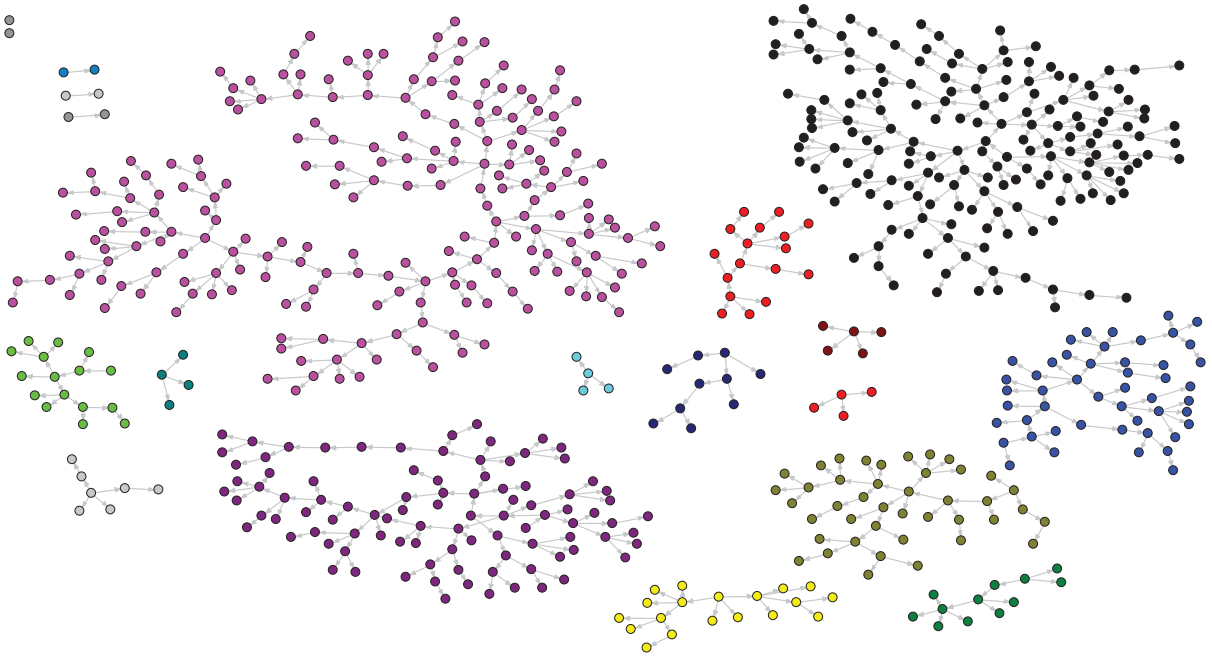


Figure 1. The recruitlment chains of submitted surveys. Each color represents a separate recruitment chain. Two seeds did not generate further participants.
doi:10.1371/journal.pone.0049417.g001

Login passwords were only valid for a single session and could not be used on two computers simultaneously. Communication between the users and the server was encrypted. Graphs with aggregated survey results were updated with data from a large number of participants at a time and displayed so that it was impossible to understand what others answered. The study was approved by the Hanoi Medical University Review Board for Bio-Medical Research.

Results

Recruitment Dynamics

676 study participants submitted a survey during the study period. The length of recruitment chains varied from 1 to 24 waves (excluding seed wave). Eight recruitment chains (out of 20) reached more than five waves (Fig. 1).

Five seeds were added 14 days after the first group (see methods). For clarity of presentation we backdated the start date of these five seeds 14 days so that all seeds could be considered to have started on the same day. Using this adjustment, the site received slightly less than 500 submissions during the first two weeks of sampling. The daily number of submissions then gradually decreased and about 100 surveys were submitted during the last 20 days, after which submissions stopped by itself (Fig. 2).

Equilibrium

Using the standard criteria in the literature [32], equilibrium was reached for all variables after a maximum of seven waves and a median of two waves. We also plotted the sample compositions with increasing sample sizes. Selected variables are shown in Fig. 3 and all variables are available in the supplementary material. Judging from these plots, the sample compositions stabilized well for all variables in the survey, with the exception of home province. The maximum absolute difference in RDSII estimated proportions comparing the full sample and the last 200 respondents among all the variables in the supplementary material, was 4.3% for estimates of proportions and 0.67 for estimated numeric values (sexual partner numbers, age and social network sizes).

Characteristics of the Sample

The majority of the sample consisted of young persons with an estimated mean and median age of 22 years. The estimated proportion with education at vocational school, college or university was 87%. An estimated 67% used the Internet every day during the past month and an estimated 82% came from the two large metropolitan areas of Ho Chi Minh City and Hanoi (81% of the sample). The recruitment chains also penetrated

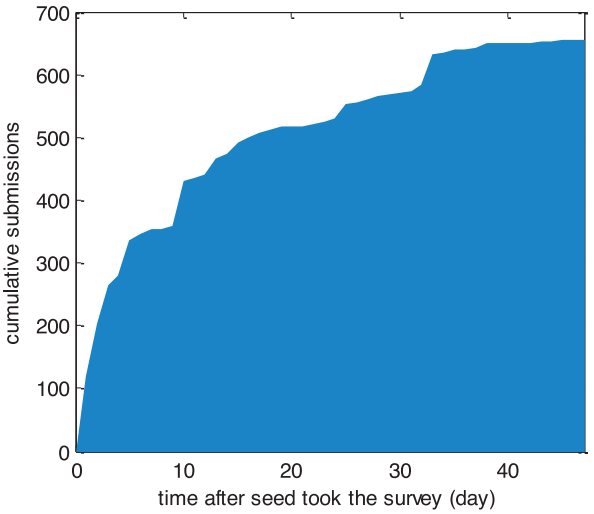


Figure 2. Cumulative number of survey submissions over time.
doi:10.1371/journal.pone.0049417.g002

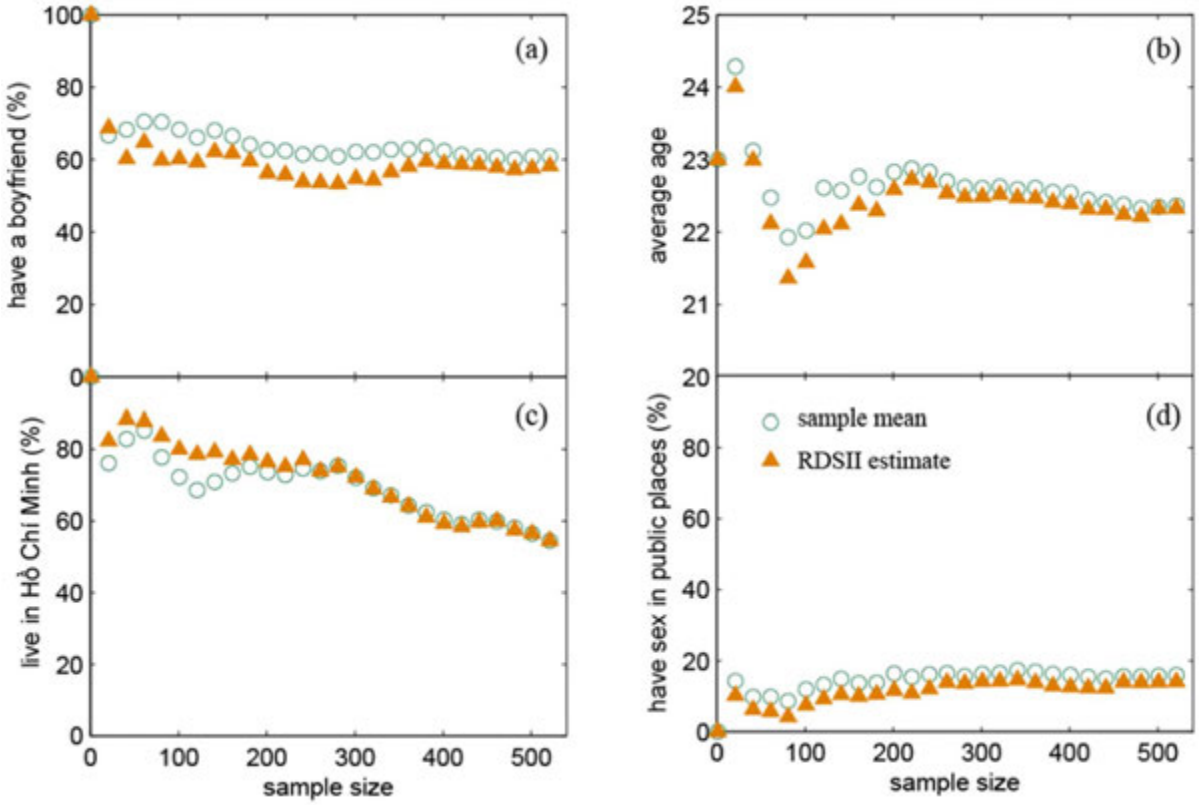


Figure 3. Change in sample composition with increasing sample size (not adjusted for network size): a) proportion that currently has a romantic relationship; b) average age; c) proportion living in Ho Chi Minh City (the only variable that did not stabilize); and d) proportion who had sex in a public place during the past six months. All variables available in the supplemental material.
doi:10.1371/journal.pone.0049417.g003

outside the large metropolitan areas with 32 provinces represented out of 63 (Fig. 4a-c).

An estimated 98% (99% of the sample) preferred only men or preferred men to women as sexual partners, and 81% (81% of the sample) thought that same-sex marriage should be allowed in Vietnam. An estimated 92% (91% of the sample) had an existing relationship to their recruiter (an estimated 8% recruited by a stranger). Median number of sexual partners during the last six months was two (Fig. 5a-c).

Discussion

We developed an automatic webRDS system to sample men who have sex with men (MSM) in Vietnam, a country in which same sex relationships are highly stigmatized and can lead to severe consequences if revealed to family members or colleagues [27]. We successfully used the system to sample and survey 676 MSM on a number of sensitive issues. We evaluated the independence of the seeds from the sample by showing that sample composition stabilized very well for all variables, possibly with the exception of home province. We used a varied set of incentives to stimulate participation and recruitment, which

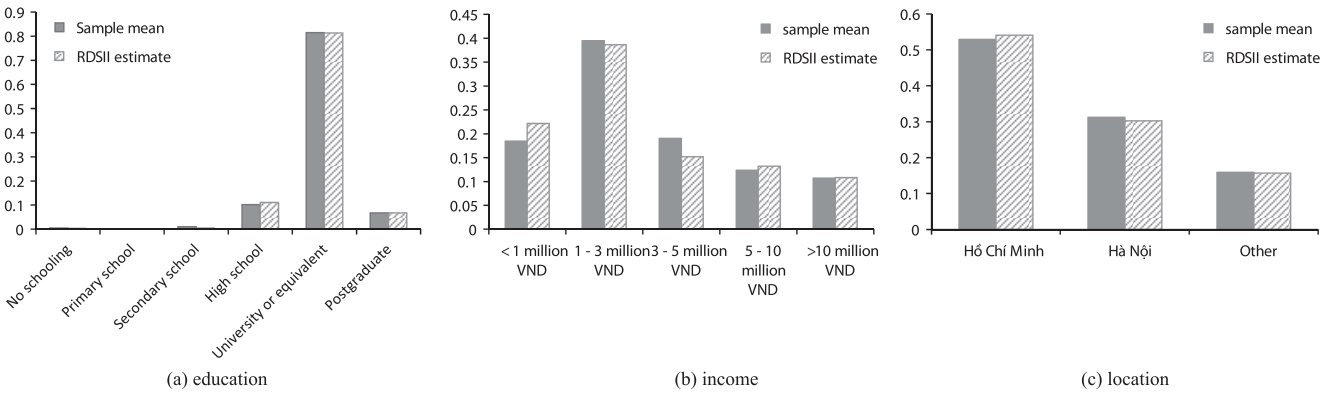


Figure 4. Sample proportions and estimated population proportions for a) education, b) income, and c) province.
doi:10.1371/journal.pone.0049417.g004

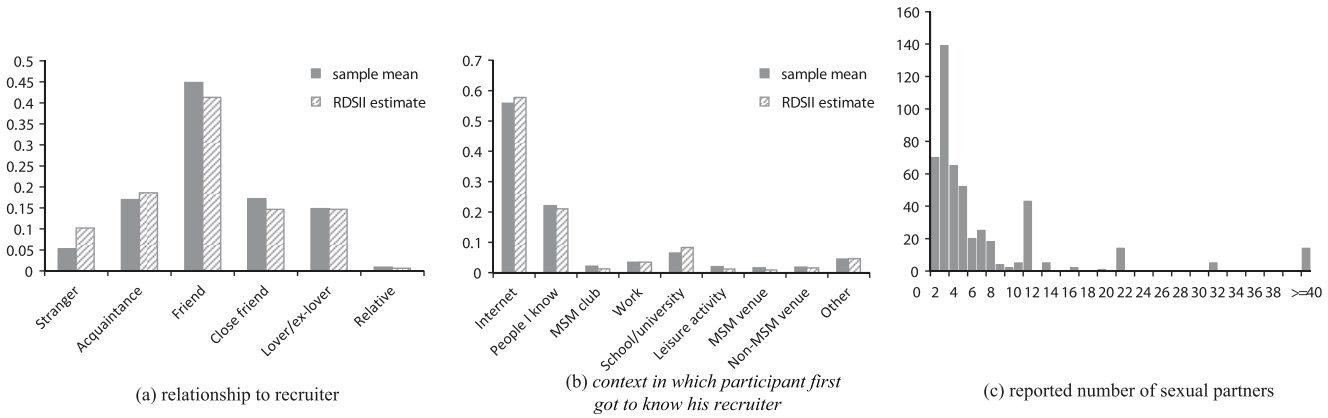


Figure 5. a) Relationship to recruiter, b) context in which participant first got to know his recruiter, and c) histogram of number of sex partners.
doi:10.1371/journal.pone.0049417.g005

became rapid and robust, reaching 24 waves. The results indicate that webRDS could potentially be implemented at a low cost among Internet-using MSM as well as potentially becoming a valuable method for sampling other Internet-using populations.

Comparing national statistics and other published research data to our estimates show interesting similarities and dissimilarities that may reflect sampling bias, variability between data collection instruments and systematic differences between the sexually active Internet-using MSM population and the general population.

Using the RDSII estimator, 97% of the MSM population under study was estimated to be below 30 years of age and the sample mean and median ages were 22 years. By comparison, 43% of the adult male population in Vietnam is between 18 and 29 [33]. The lower mean age of sampled MSM compared to the national age distribution for men is consistent with an offline RDS study of MSM in Khanh Hoa, Vietnam, which reported a median ages of 24 years [34] and an RDS in Hanoi with median age of 20–24 years [31]. One online survey among visitors to Vietnamese MSM websites has been published and had a median age of 23 years with 18% stating an age above 30 years [31]. Income distribution (Fig. 4b) is broadly consistent with the national average monthly per capita income for urban areas (2,130,000 VND, 2010 [35]). It is also comparable to data from the online survey among visitors to Vietnamese MSM websites [31] and the offline RDS in Hanoi 2008 [31], although inflation, economic growth and differential categorization of income levels precludes an exact comparison. An estimated 88% had some type of post-secondary education, including vocational training. This can be compared with 68% in the offline RDS in Hanoi [31] and 79% in the survey among visitors to Vietnamese MSM websites [31]. The sample was heavily concentrated to the two large metropolitan areas of Ho Chi Minh City and Hanoi, with a population estimate of 84% for these cities combined. Ho Chi Minh City and Hanoi constitute approximately 55% of the urban population in Vietnam and about 16% of the national population [36,37]. This is similar to the online banner survey on Vietnamese MSM websites where 74% came from Hanoi and HCMC [31]. Explanation for the observed differences compared with national statistics may include migration of young MSM to the large cities, urban-rural differences in prevalence of male-male sex and different levels of access to the Internet. We did not find evidence that the men's social networks formed geographically isolated groups, which otherwise would have been a source of bias. The recruitment chains in our sample frequently crossed over between provinces. In total, 30% of all recruitment events took place between persons in different provinces. Additionally, like other social networks, MSM networks in Vietnam are most likely small-world networks [38], with short numbers of steps between provinces.

One percent stated that they preferred only women or preferred women to men as sexual partners. The banner survey on MSM sites [31] and an offline RDS in Hanoi with similar question [31] recorded 15% and 1.9% respectively for the same responses. A middle option ("Prefer women and men equally") was available in these studies in contrast to our study, with 14% and 8% of answers respectively.

In summary, the webRDS reached a varied sample of largely young men concentrated in the two major cities of Vietnam, with an education higher than the average for the country. Whether these results indicate sampling bias or reflect differences between the general Vietnamese population and the Internet-using MSM population is difficult to assess.

There are limitations to this study. We excluded 13% (n = 85) of the submissions because of duplicated personal information or an age below 18 years. While this shows that the recruitment system did not work perfectly, it also shows the potential for eliminating duplicate submissions.

17.5% of completed surveys (115 surveys) included an IP number that had previously been registered in the system, which may signal duplicated submissions. However, it is important to note that IP-numbers are shared by all users at an Internet cafe and often by all users within a neighborhood. An array of non-Vietnamese IP-numbers is also used in Vietnam to access restricted sites like Facebook. We checked whether the final estimates were sensitive to exclusion of these submissions as well as of submission with very fast completion time, and did not find that this was the case. Similar protocols for quality check as those used in this study have been employed in other Internet-based surveys among MSM [39,40]. Because sincere and insincere participants are likely to interact differently with web survey pages, analyzes of online behavioral data gathered during surveys may in the future provide a way to improve these protocols.

We opted for removal of ineligible participants after the study was concluded in order to observe recruitment behavior without outside involvement and to avoid running the risk of inadvertently stopping the survey. This procedure should not produce bias in the RDSII estimation if removal of ineligibles is made in a correct way. However, there will surely always remain questions as to the extent to which ineligibles have been fully removed or not. If

future studies show that ineligibles can safely be removed without stopping global recruitment this is preferable.

We lack information as to the true participation rate as we do not have information on the proportion of invitation messages forwarded by the participants. However, we get some additional information by considering that on average, in order for recruitment to be sustained, each participant needs to recruit a minimum of one new participant. Persons who participated in the formative research for the study may be part of the sample, but it seems unlikely that this should have created important bias. The network size question did not exclude MSM living outside Vietnam and those of ages under 18 years. Potentially this may have underestimated e.g. the proportion of young persons. Eight percent of participants were recruited by a stranger. Other RDS studies among MSM have recorded similar proportions [41,42,43]. We do not think this caused serious bias in this study (see e.g [20]) but the issue should be monitored in future webRDS studies.

Although this study aimed to sample Internet-using MSM, access to the Internet, including literacy, will always be a limiting factor for representative sampling of MSM in general. Bio-markers will obviously also be challenging to collect with webRDS.

Although more than 600 submissions were received within five weeks, recruitment eventually died out despite being far from the total size of the Internet-using MSM population in Vietnam (at least 10,000 persons). One explanation may be local as opposed to global saturation. It is very common in acquaintance networks that an individual's neighbors are connected with each other. The risk that people will try to recruit acquaintances who have already been recruited will therefore increase over time, decreasing the effective reproductive number and could result in a curve shape similar to the one in Fig. 2 [44].

WebRDS may have several advantages over standard offline RDS and other Internet-based sampling methods for hidden groups. In comparison to standard RDS it may allow for representative sampling of hidden groups without geographical limits and can potentially generate larger samples than standard RDS. This would also enable valuable data on variables for which design effects are high [14]. Online networks may also cross social boundaries more often than offline social networks (decreased homophily) and are likely to generally produce larger average personal network sizes than offline networks, both of which can decrease design effects [15]. Individuals who for various reasons prefer not to access an RDS survey office physically may additionally be willing to take part in an anonymous web survey.

Web-RDS will in most cases entail a lower costs than a standard RDS study. Cost for monetary incentives were in our study on average 5.9 USD per participant in the cleaned sample (3353 USD in total). Staff hours to interact with seeds, deliver incentives, monitoring invalid submissions etc, totaled one month full-time equivalents (FTE). Adjustment of the site to appeal to the local target group is technically easy but requires formative research. For comparison, an offline RDS would have shared similar costs for incentives and formative research about the study population (see e.g. [45]) but would also require a survey office and at least five months staffing (conservative FTE estimate). We are currently improving the ease of use of the software and researchers interested in the survey system are welcome to contact the authors.

Current online recruitment of hidden groups is based on self-selected samples of persons who access certain Internet sites and click banner ads for a study. These surveys often have participation rates of a few in a thousand to a few in a hundred [23,24]. As compared with such online sampling, successful webRDS is likely to achieve considerably reduced self-selection bias, because sustained recruitment is likely to be highly correlated to high participations rates.

In summary, we developed a webRDS system to sample men who have sex with men in Vietnam and showed that is was possible to survey participants on a range of sensitive issues, including sexual behavior, while sustaining recruitment and achieving equilibrium. The results indicate that the method could potentially be implemented at low cost among Internet-using MSM. With further evaluation and among suitable population groups, Internet-based RDS could become a promising method for representative sampling online.

Supporting Information

Figure S1 Sample proportions and RDSII estimates with increased sample size. Cleaned sample used. The curves for sample groups with other criteria are similar.
(EPS)

Table S1 RDSII estimates for samples with progressively stricter inclusion criteria.
(DOCX)

Author Contributions

Conceived and designed the experiments: LB XL QCN MC FL AT. Performed the experiments: LB XL QCN MC NLH TAN FL AT. Analyzed the data: XL LB. Wrote the paper: LB XL MC FL AT.

References

1. Baral S, Sifakis F, Cleghorn F, Beyrer C (2007) Elevated risk for HIV infection among men who have sex with men in low- and middle-income countries 2000–2006: a systematic review. PLoS Med 4: e339.
2. Smith AD, Tapsoba P, Peshu N, Sanders EJ, Jaffe HW (2009) Men who have sex with men and HIV/AIDS in sub-Saharan Africa. Lancet 374: 416–422.
3. Mumtaz G, Hilmi N, McFarland W, Kaplan RL, Akala FA, et al. (2011) Are HIV Epidemics among Men Who Have Sex with Men Emerging in the Middle East and North Africa?: A Systematic Review and Data Synthesis. PLoS Med 8: e1000444.
4. The Commission on AIDS in Asia (2008) Redefining AIDS in Asia. Crafting an Effective Response. New Delhi. http://data.unaids.org/pub/Report/2008/20080326_report_commission_aids_en.pdf.
5. HIV/AIDS Survey Indicators Database (2011). http://hivdata.measuredhs.com/start.cfm.
6. Heckathorn DD (1997) Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems 44: 174–199.
7. Heckathorn DD (2002) Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. Social Problems 49: 11–34.
8. Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology, 2004, Vol 34 34: 193–239.
9. Salganik MJ (2006) Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health-Bulletin of the New York Academy of Medicine 83: I98–I112.
10. Volz E, Heckathorn DD (2008) Probability Based Estimation Theory for Respondent Driven Sampling. Journal of Official Statistics 24: 79–97.
11. Wejnert C, Heckathorn DD (2011) Respondent-Driven Sampling: Operational Procedures, Evolution of Estimators, and Topics for Future Research. In: Williams M, Vogt WP, editors. The SAGE Handbook of Innovation in Social Research Methods. Los Angeles, London,: SAGE.
12. Goodman LA (1961) Snowball Sampling. Annals of Mathematical Statistics 32: 148–170.
13. Goel S, Salganik MJ (2009) Respondent-driven sampling as Markov chain Monte Carlo. Statistics in Medicine 28: 2202–2229.
14. Goel S, Salganik MJ (2010) Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences of the United States of America 107: 6743–6747.

15. Lu X, Bengtsson L, Britton T, Camitz M, Kim BJ, et al. (2012) The sensitivity of respondent-driven sampling. Journal of the Royal Statistical Society Series a-Statistics in Society 175: 191–216.
16. Wejnert C, Heckathorn DD (2008) Web-based network sampling - Efficiency and efficacy of respondent-driven sampling for online research. Sociological Methods & Research 37: 105–134.
17. Bengtsson L, Thorson A (2010) Global HIV surveillance among MSM: is risk behavior seriously underestimated? AIDS 24: 2301–2303.
18. Johnston LG, Malekinejad M, Kendall C, Iuppa IM, Rutherford GW (2008) Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: Field experiences in international settings. Aids and Behavior 12: S131–S141.
19. Lu X (2012) Improving Sample Estimate Reliability and Validity with Linked Ego Networks. arXiv: 12051971v1.
20. Lu X, Malmros J, Liljeros F, Britton T (2012) Respondent-driven Sampling on Directed Networks. arXiv: 12011927v2.
21. Magnani R, Sabin K, Saidel T, Heckathorn D (2005) Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS 19 Suppl 2: S67–72.
22. Berk R (1983) Introduction to Sample Selection Bias in Sociological Data. American Sociological Review 48: 386 to 398.
23. Zhang D, Bi P, Lv F, Tang H, Zhang J, et al. (2007) Internet use and risk behaviours: an online survey of visitors to three gay websites in China. Sexually Transmitted Infections 83: 571–576.
24. Jakopanec I, Schimmer B, Grjibovski AM, Klouman E, Aavitsland P (2010) Self-reported sexually transmitted infections and their correlates among men who have sex with men in Norway: an Internet-based cross-sectional survey. Bmc Infectious Diseases 10.
25. 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.
26. Bauermeister JA, Zimmerman MA, Johns MM, Glowacki P, Stoddard S, et al. (2012) Innovative Recruitment Using Online Networks: Lessons Learned From an Online Study of Alcohol and Other Drug Use Utilizing a Web-Based, Respondent-Driven Sampling (webRDS) Strategy. J Stud Alcohol Drugs 73: 834–838.
27. Vu ML, Le Thi MP, Nguyen TV, Doan KT, Tran QT, et al. (2009) MSM in Viet Nam- Social Stigma and Consequences. Hanoi: SHAPC. http://www.google.se/url?sa=t&rct=j&q=stigma%20vietnam%20msm&source=web&cd=2&ved=0CCwQFjAB&url=http%3A%2F%2Fwww.unaids.org.vn%2Fsitee%2Fimages%2Fstories%2Fresearch_report-eng.pdf&ei=VaumTq24COLO4QT8kb0K&usg=AFQjCNFS6GksrggZ9HzVoVf1DI4fYkreEQ&sig2=FVysds8P_5KW3hlIZzw3Kw&cad=rja.
28. Ministry of Health Vietnam (2009) Integrated Biological and Behavioral Surveillance. Hanoi.
29. Cimigo (2010) Urban Vietnam Internet Penetration Hits 50%. http://www.cimigo.vn/en-US/PressCoverage/vietnam-internet-research/urban-vietnam-internet-penetration-hits-50.aspx.
30. Ngo DA, Ross MW, Phan H, Ratliff EA, Trinh T, et al. (2009) Male Homosexual Identities, Relationships, and Practices among Young Men Who Have Sex with Men in Vietnam: Implications for Hiv Prevention. Aids Education and Prevention 21: 251–265.
31. Nguyen QC (2010) Sexual risk behaviors of men who have sex with men in Viet Nam. Chapel Hill: North Carolina State University. http://gradworks.umi.com/34/18/3418727.html.
32. Heckathorn D, Semaan S, Broadhead R, Hughes J (2002) Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18–25. Aids and Behavior 6: 55–67.
33. U.S. Census Bureau's International database (2012) Vietnam.
34. Colby D, Minh TT, Toan TT (2008) Down on the farm: homosexual behaviour, HIV risk and HIV prevalence in rural communities in Khanh Hoa province, Vietnam. Sex Transm Infect 84: 439–443.
35. General Statistics Office of Vietnam (2010) Result of the Vietnam Household living standards survey 2010. Hanoi. http://www.gso.gov.vn/default_en.aspx?tabid=515&idmid=5&ItemID=12426.
36. World Bank (2010) Open data: Urban Population. http://data.worldbank.org/indicator/SP.URB.TOTL.
37. General Statistics Office of Vietnam (2009) The 2009 Vietnam Population and Housing census: Completed results. Hanoi. http://www.gso.gov.vn/default_en.aspx?tabid=515&idmid=5&ItemID=10799.
38. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
39. Konstan JA, Rosser BRS, Ross MW, Stanton J, Edwards WM (2005) The story of subject naught: A cautionary but optimistic tale of Internet survey research. Journal of Computer-Mediated Communication 10: article 11.
40. Bauermeister J, Pingel E, Zimmerman M, Couper M, Carballo- Dieguez A, et al. (2012) Data Quality in HIV/AIDS Web-Based Surveys Handling Invalid and Suspicious Data. Field Methods 24: 272–291.
41. Iguchi MY, Ober AJ, Berry SH, Fain T, Heckathorn DD, et al. (2009) Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications. Journal of Urban Health-Bulletin of the New York Academy of Medicine 86: S5–S31.
42. Ma XY, Zhang QY, He X, Sun WD, Yue H, et al. (2007) Trends in prevalence of HIV, syphilis, hepatitis C, hepatitis B, and sexual risk behavior among men who have sex with men - Results of 3 consecutive respondent-driven sampling surveys in Beijing, 2004 through 2006. Jaids-Journal of Acquired Immune Deficiency Syndromes 45: 581–587.
43. Frost SD, Brouwer KC, Firestone Cruz MA, Ramos R, Ramos ME, et al. (2006) Respondent-driven sampling of injection drug users in two U.S.-Mexico border cities: recruitment dynamics and impact on estimates of HIV and syphilis prevalence. J Urban Health 83: i83–97.
44. Liljeros F (2009) Human sexual networks. Encyclopedia of Complexity and System Science: Springer Science and Business Media.
45. Johnston LG, Whitehead S, Simic-Lawson M, Kendall C (2010) Formative research to optimize respondent-driven sampling surveys among hard-to-reach populations in HIV behavioral and biological surveillance: lessons learned from four case studies. Aids Care-Psychological and Socio-Medical Aspects of Aids/Hiv 22: 784–792.

Implementation of Web-Based Respondent-Driven Sampling among Men who Have Sex with Men in Vietnam

Supporting Information

1. RDS estimates for sample data with progressively stricter criteria

We checked all surveys for other signs of duplication or invalid submission. We flagged surveys containing a repeated IP number, deviating answers, or short completion times. We analyzed the sensitivity of the estimates to inclusion and exclusion of these flagged submissions. Specifically, we compared the RDSII estimates generated from the full sample of non-seed submissions with valid age with the RDSII estimates generated from groups with progressively stricter inclusion criteria according to the following:

Non-Strict:

All non-seed respondents with valid age (≥18);

Cleaned sample[1]:

All non-seed respondents with valid age (≥18);
Exclude submissions with repeated email, Yahoo! Chat ID or telephone number;

Strict:

All non-seed respondents with valid age (≥18);
Exclude submissions with repeated email, Yahoo! Chat ID, telephone number;
Exclude submissions with repeated IP address;

Very strict:

All non-seed respondents with valid age (≥18);
Exclude submissions with repeated email, Yahoo! Chat ID, telephone number;
Exclude submissions with repeated IP address;
Exclude submissions with short completion times (< 3 minutes), submissions stating no education (rare in Vietnam), or submission stating six-month partner numbers of above 1,000.

RDSII estimates of all 17 questions for the above sample groups with progressively stricter inclusion criteria are listed in Table 1. The differences between groups are small. The average absolute differences in proportional estimates when comparing the full sample (non-strict criteria) to the other groups, is less than 0.64% (maximum difference 6.6%), and the average absolute differences in numeric estimates is 0.12 (maximum difference 0.30), see Table 1.

[1] This is the cleaned sample discussed in the paper.

Variable	Non-Strict (n=634)	Cleaned sample (n=571)	Strict (n=490)	Very Strict (446)
Proportional estimates (%)				
1. Have a boyfriend now	57.5	56.6	58.0	57.4
2. Longest relationship with men ≥ 6 months	52.0	52.5	51.4	51.3
3. Prefer "good looking" when looking for someone for sex	47.6	48.5	49.4	48.6
4. Prefer "faithful" when looking for someone for long-term relationship	39.2	41.1	41.5	39.5
5. Support same sex marriage	77.3	79.0	80.0	81.2
6. Prefer only men as sexual partners	68.0	68.3	67.3	66.2
7. Had sex in public places during past 6 months	14.2	13.4	11.8	12.2
8. Have some education after high school (vocational training, college or university)	87.6	88.1	87.2	86.5
9. Monthly income ≥ 5 million VND	24.0	24.0	25.7	25.9
10. Live in Hồ Chí Minh city	53.4	54.1	55.3	55.9
11. Use the Internet everyday	57.0	59.6	63.6	62.2
12. Recruited by friend	40.9	41.3	42.9	42.3
13.First go to know recruiter through friends, lovers or relatives	21.4	21.1	20.0	20.3
Numeric estimates				
14. Number of men had sex with during past 6 months	4.03	4.07	3.96	4.05
15. Average age	22.23	22.22	22.07	22.06
16. Number of MSM friends	6.74	6.76	6.91	7.04
17. Number of MSM friends who use Internet	5.23	5.24	5.43	5.47

*For categorical questions one answer per question is shown

2. Equilibrium curves for all variables surveyed in study

To get an overview of whether the sample reached equilibrium, we plot both the sample proportion and the RDSII estimates along with the increased sample size for all variables surveyed in this study (see Figure 1).

To measure the change during the last part of the sampling process when the sampling compositions should have stabilized, we calculate the changes in the sample compositions comparing the full sample and the full sample excluding the last 200th respondents:

$$\Delta p = \left| p_n - p_{n-200} \right| \quad (1.1)$$

$$\Delta \hat{p}^{RDSII} = \hat{p}_n^{RDSII} - \hat{p}_{n-200}^{RDSII} \quad (1.2)$$

Or

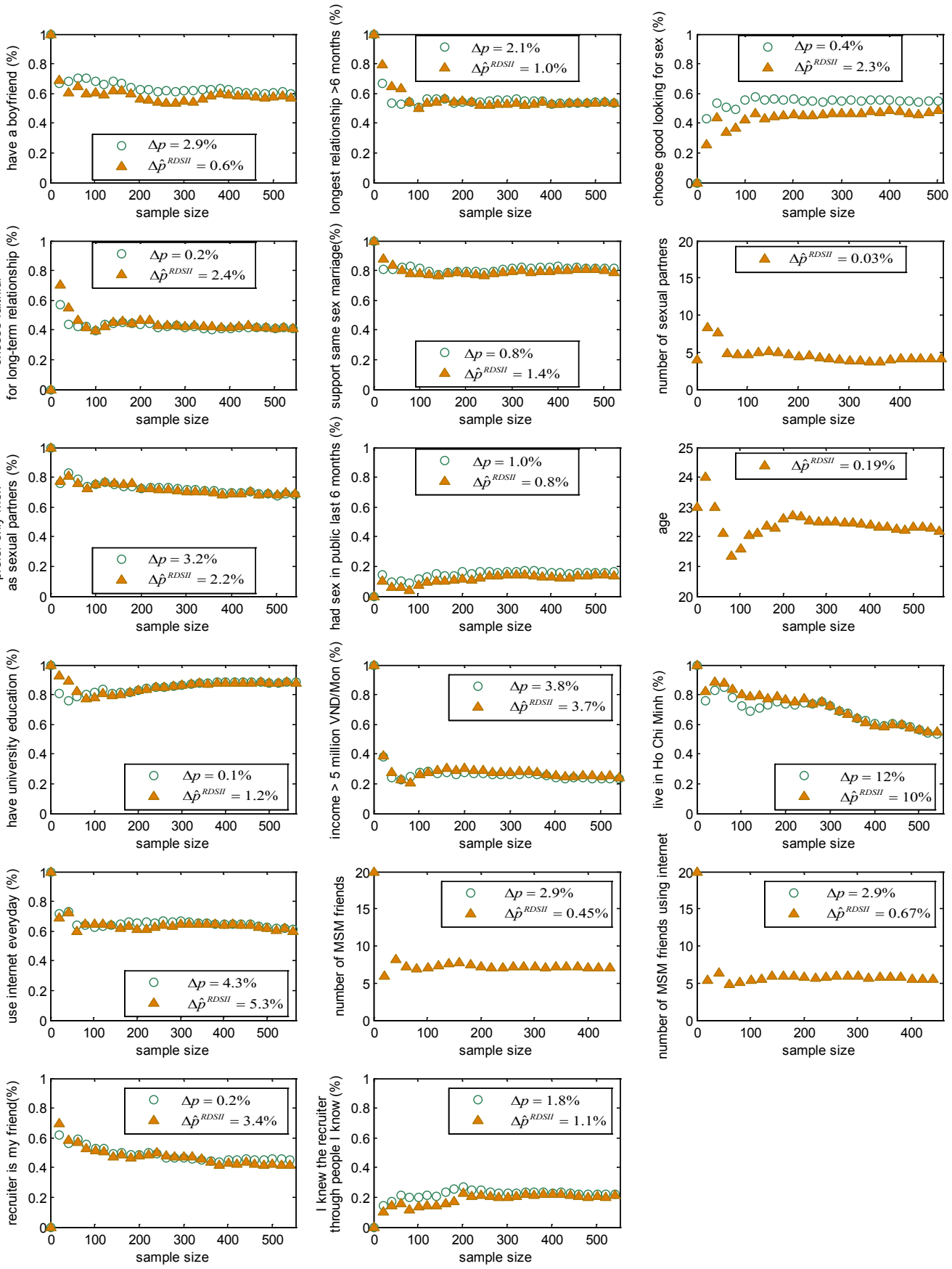
$$\Delta \hat{y}^{RDSII} = \hat{y}_n^{RDSII} - \hat{y}_{n-200}^{RDSII} \quad (1.3)$$

We can see that except for answers for province of residence, the sample proportion and RDS estimates for all other variables became quite stable after 100~200 submissions, the maximum absolute difference between last 200 respondents, is 5.3% for raw sample proportions, 4.3% for RDSII estimated proportions, and 0.67 for the RDSII estimated average numerical variables.

The decreased proportions of respondents from Hồ Chí Minh City, indicates that the recruitment chains on average spread from Hồ Chí Minh City to other provinces. The stable estimates for variables not related to place of residence suggest that there is little difference between provinces regarding the social and sexual behaviors/opinions studied here.

Four seeds were selected from Hoa Binh and as can be expected from the RDS Markov process that should create independence between the seeds and the final sample, the proportion from the small province of Hoa Binh was only an estimated 1.6%.

There may be a number of possible reasons for not reaching stable proportions for place of residence in the present study sample. Internet-using persons from countryside may have less access to computers in the home and thus take longer time before completing surveys. The network may also be less developed between the large cities and the countryside and social networks between any two rural provinces may be less developed than between rural and urban provinces causing recruitments between rural provinces to go through urban hubs.


have a boyfriend (%)
$\Delta p = 2.9\%$
$\Delta \hat{p}^{RDSII} = 0.6\%$
longest relationship >6 months (%)
$\Delta p = 2.1\%$
$\Delta \hat{p}^{RDSII} = 1.0\%$
choose good looking for sex (%)
$\Delta p = 0.4\%$
$\Delta \hat{p}^{RDSII} = 2.3\%$
choose faithful for long-term relationship (%)
$\Delta p = 0.2\%$
$\Delta \hat{p}^{RDSII} = 2.4\%$
support same sex marriage(%)
$\Delta p = 0.8\%$
$\Delta \hat{p}^{RDSII} = 1.4\%$
number of sexual partners
$\Delta \hat{p}^{RDSII} = 0.03\%$
prefer only men as sexual partners (%)
$\Delta p = 3.2\%$
$\Delta \hat{p}^{RDSII} = 2.2\%$
had sex in public last 6 months (%)
$\Delta p = 1.0\%$
$\Delta \hat{p}^{RDSII} = 0.8\%$
age
$\Delta \hat{p}^{RDSII} = 0.19\%$
have university education (%)
$\Delta p = 0.1\%$
$\Delta \hat{p}^{RDSII} = 1.2\%$
income > 5 million VND/Mon (%)
$\Delta p = 3.8\%$
$\Delta \hat{p}^{RDSII} = 3.7\%$
live in Ho Chi Minh (%)
$\Delta p = 12\%$
$\Delta \hat{p}^{RDSII} = 10\%$
use internet everyday (%)
$\Delta p = 4.3\%$
$\Delta \hat{p}^{RDSII} = 5.3\%$
number of MSM friends
$\Delta p = 2.9\%$
$\Delta \hat{p}^{RDSII} = 0.45\%$
number of MSM friends using internet
$\Delta p = 2.9\%$
$\Delta \hat{p}^{RDSII} = 0.67\%$
recruiter is my friend(%)
$\Delta p = 0.2\%$
$\Delta \hat{p}^{RDSII} = 3.4\%$
I knew the recruiter through people I know (%)
$\Delta p = 1.8\%$
$\Delta \hat{p}^{RDSII} = 1.1\%$
sample size