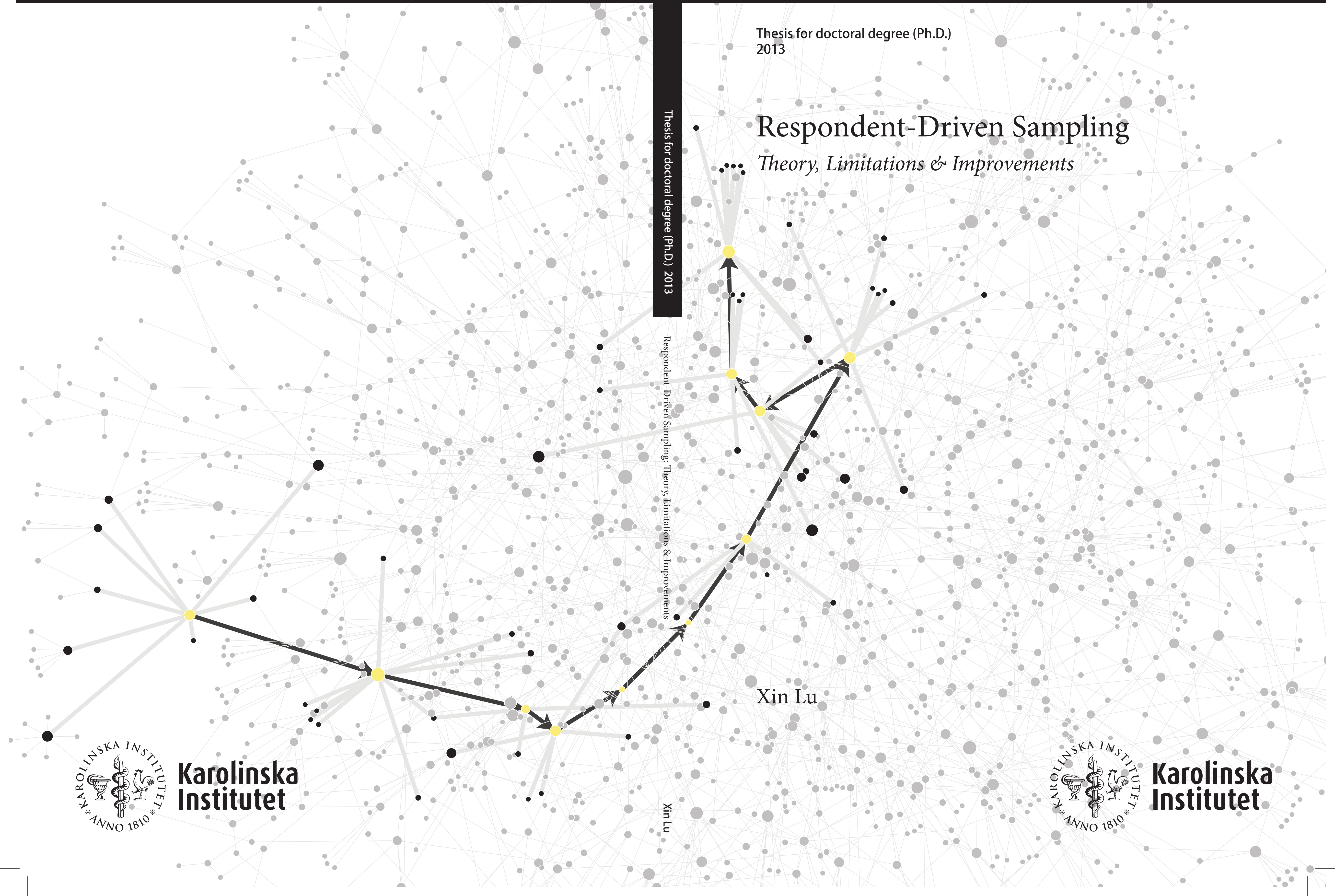


Thesis for doctoral degree (Ph.D.) 2013

Respondent-Driven Sampling: Theory, Limitations & Improvements

Xin Lu

Thesis for doctoral degree (Ph.D.)
2013

Respondent-Driven Sampling

Theory, Limitations & Improvements

Xin Lu

From Department of Public Health Sciences, Karolinska Institutet,
Stockholm, Sweden

Respondent-Driven Sampling

Theory, Limitations & Improvements

Xin Lu

Stockholm 2013

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet. Printed by US-AB, Stockholm.

ISBN 978-91-7549-047-2

Abstract

Background: The key purpose of sampling is to gain knowledge about a population using a small, affordable subset of selected individuals. This goal is often approached by choosing a representative sample with each individual's selection probability determined by a full list of individuals from the target population. However, for many populations central to the public health sciences, such as men who have sex with men (MSM), injecting drug users (IDUs), etc., the selection probability of individuals cannot be determined ahead of time because the list of all individuals is not available, impairing the generalization of results from the sample to the population. Respondent-driven sampling (RDS) was developed to generate representative samples of such hard-to-reach populations with improved accessibility. It provides an automated self-growing sampling design as well as asymptotically unbiased population estimates, making it the state-of-the-art sampling method for studying HIV-related key populations at risk in the past years. However, the availability of RDS estimates relies on many assumptions that are often not satisfied in real practice.

Aims: To assess the effect of violating assumptions on the performance of RDS estimators and to improve both the implementation and methodology of RDS for hard-to-reach populations of relevance to the public health sciences.

Contributions: The performance of RDS estimators is evaluated under various conditions. Results indicate that long chains initiated by diverse seeds are highly beneficial, while estimate bias is large if the network is directed or if respondents' participation behavior (such as preferential recruitment) depends on characteristics that are correlated with study outcomes. An Internet-based RDS (WebRDS) recruiting system is developed to circumvent the limitation of physical interview-based implementations. The system shows its ability to recruit sustaining location-free respondents in a study of MSM in Vietnam. Statistical methods are developed to generalize the RDS method from undirected networks to directed networks. The new method can function as a sensitivity test tool to account for the uncertainties of network directedness and error in self-reported degree data. Lastly, by integrating traditional RDS chain data with self-reported ego network data, a new estimator was developed to improve the reliability and validity of RDS. The new estimator shows not only improved precision, but also strong robustness to the preference of peer recruitment and variations in network structural properties.

Conclusions: Violations of assumptions are inevitable and should be investigated thoroughly in RDS practice. Due to the relatively high variance and vulnerability to certain harmful conditions, such as directedness, preferential recruitment, etc., results from RDS studies should be interpreted with caution. Researchers are encouraged to collect ego network data through the implementation of RDS to improve the precision of population estimates. In spite of its limited ability to generate close-enough population estimates, RDS is easily implementable and it offers a method with an improved response rate, providing an alternative to gain access/venue to the understanding of hard-to-access population.

Keywords: social networks, directed networks, ego networks, sampling, nonprobability sampling, respondent-driven sampling, Internet, estimator, bias, variance, public health, HIV, hidden population, differential recruitment, reporting error

List of publications

I. **Lu X**, Bengtsson L, Britton T, Camitz M, Kim BJ, Thorson A, Liljeros F. “The sensitivity of respondent-driven sampling”. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 2012, 175: 191-216.

II. Bengtsson L, **Lu X**, Nguyen QC, Camitz M, Hoang NL, Liljeros F, Thorson A et al. “Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam”. *PLOS ONE*, 2012, 7 (11): e49417.

III. **Lu X**, Malmros J, Liljeros F, Britton T. “Respondent-driven Sampling on Directed Networks”. *Electronic Journal of Statistics*, 2013, 7: 292-322.

IV. **Lu X**. “Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-driven Sampling”. 2012, *arXiv*: 1205.1971v2. http://arxiv .org/abs/1205.1971v2. (submitted)

These articles will be referred to in the text by their Roman numerals (I-IV).

OTHER PAPERS BY THE AUTHOR:

Lu X, Bengtsson L, Holme P. “Predictability of population displacement after the 2010 Haiti earthquake”. *PNAS*, 2012, 109 (29): 11576-11581.

Bengtsson L, **Lu X**, Thorson A, Garfield R, von Schreeb J. “Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti”. *PLOS Medicine*, 2011, 8 (8): e1001083.

Lu X, Camitz M. “Finding the shortest paths by node combination”. *Applied Mathematics and Computation*, 2011, 217 (13): 6401-6408.

LIST OF ABBREVIATIONS

AIDS	*Acquired immunodeficiency syndrome*
AE	*Average estimate*
BFS	*Breath-first-search*
CDC	*Centers for Disease Control and Prevention*
CI	*Confidence interval*
DE	*Design effect*
GCC	*Giant connected component*
GIN	*Giant in-component*
GOUT	*Giant out-component*
GSCC	*Giant strongly connected component*
GWCC	*Giant weakly connected component*
HIV	*Human immunodeficiency virus*
HRH	*High-risk heterosexual*
iSEE	*Institute for Studies of Society, Economy and Environment*
IDU	*Injecting drug users*
LGBT	*Lesbian, gay, bisexual, and transgender*
MAE	*Mean absolute error*
MCMC	*Markov Chain Monte Carlo*
MSM	*Men who have sex with men*
NAM	*Nodal attribute model*
NEM	*Network evolution models*
PPS	*Probability-proportional-to-size*
RMSE	*Root mean square error*
RDS	*Respondent-driven sampling*
SD	*Standard deviation*
SRS	*Simple random sampling*
SW	*Sex worker*
SWOR	*Sampling without replacement*
SWR	*Sampling with replacement*
TLS	*Time-location sampling*
TS	*Targeted sampling*
UNAIDS	*The Joint United Nations Program on HIV/AIDS*
URL	*Uniform resource locator*
VDT	*Venue-day-time*
WebRDS	*Web-based respondent-driven sampling*
WWW	*World Wide Web*

Table of contents

Preface **9**

1 Introdution to Sampling Methods **11**

1.1 Sampling methods 11

1.1.1 Probability sampling methods *12*

1.1.2 Nonprobability sampling methods *14*

1.2 Population inference 17

1.2.1 Equal inclusion probability and sample representativeness17

1.2.2 Unequal inclusion probability and sample weight *17*

1.2.3 Toward a more representative nonprobability sample *19*

2 Challenge of Sampling for HIV Surveillance **21**

2.1 High-risk populations 21

2.2 Challenge of sampling hidden population 21

2.3 A note on accessibility and generalizability 23

3 Network in a Nutshell **24**

3.1 Introduction 24

3.2 Properties of networks 24

3.2.1 Network connectivity *24*

3.2.2 Degree *25*

3.2.3 Degree distribution *26*

3.2.4 Shortest path and the small-world experiment *26*

3.2.5 Clustering *27*

3.2.6 Community structure *27*

3.2.7 Mixing patterns *28*

3.3 Network models 29

3.3.1 Regular networks *29*

3.3.2 Erdős–Rényi random network *29*

3.3.3 Small-world networks *29*

3.3.4 Scale-free networks *30*

3.3.5 Social networks models *32*

4 Respondent-Driven Sampling **34**

4.1 How does RDS work? 34

4.2 Difference between RDS and snowball sampling 34

4.3 Theory of RDS: models and estimators 36

4.3.1 Modeling RDS as a Markov process *36*

4.3.2 RDS estimator: RDSII *37*

4.3.3 RDS estimator: RDSI *38*

4.3.4 Connection between RDSI and RDSII *42*

4.3.5 Variance estimation *43*

4.4 RDS around the world 43

4.5 Limitations 44

4.6 Recent development of RDS theories 46

4.6.1 Heckathorn-estimator *46*

4.6.2 SS-estimator *47*

4.6.3 GH-estimator *48*
4.6.4 Other approaches *49*
4.7 Summary 49
5 Objectives and Framework 51
5.1 Objectives 51
5.2 Framework 51
6 Materials and Methods 53
6.1 Empirical network data (Paper I, III, IV) 53
6.1.1 Data collection *53*
6.1.2 Network formation *53*
6.1.3 Network properties *54*
6.2 WebRDS (Paper II) 55
6.2.1 The WebRDS system *55*
6.2.2 Inclusion criteria and incentives *55*
6.2.3 Questionnaire *56*
6.2.4 Sampling procedure *56*
6.2.5 Piloting and early version of the system *57*
6.3 Tools for data processing and simulation 57
6.4 Ethics 58
7 Results 60
7.1 Paper I: The Sensitivity of Respondent-Driven Sampling 60
7.1.1 Summary *60*
7.1.2 Study design *60*
7.1.3 Result *61*
7.2 Paper II: Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam 62
7.2.1 Summary *62*
7.2.2 Sampling dynamics *63*
7.2.3 Duplicated submissions, data cleaning and analysis *63*
7.2.4 Equilibrium *64*
7.2.5 Sample characteristics *64*
7.3 Paper III: Respondent-Driven Sampling on Directed Networks 66
7.3.1 Summary *66*
7.3.2 Study design *66*
7.3.3 Result *69*
7.4 Paper IV: Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-Driven Sampling 70
7.4.1 Summary *70*
7.4.2 Study design *70*
7.4.3 Result *73*
8 Conclusions and Discussion 74
8.1 Key findings and contributions 74
8.1.1 Performance of RDS estimators *74*
8.1.2 Improved RDS estimate methods *75*
8.1.3 Implementation of WebRDS *76*
8.2 Limitations 77
8.2.1 Limitations of VH_m estimator *77*
8.2.2 Limitations of $RDSI^{ego}$ estimator *78*

8.2.3 Limitations of WebRDS 78
8.3 Implications for RDS practitioners 79
8.3.1 Accessing hidden populations over the Internet.................. 79
8.3.2 Seeds, coupons and sample size... 80
8.3.3 Treatment for violation of assumptions 81
8.3.4 A note on the raw sample proportion and variance estimation 83
8.4 Concluding remarks and future research 83
8.4.1 Concluding remarks 83
8.4.2 Future research 85
Acknowledgements 88
References.. 90

PREFACE

Sampling is the art of studying the whole by the part. Instead of making an exhausting, expensive, and time consuming census enumeration, an affordable subset of individuals selected according to a specified sampling design is enough to enable researchers to gain critical knowledge about the studied population.

A good sampling method at least offers ***accessibility***, that is, a proper design that allows researchers access to a set of subjects to investigate. Second and ideally, it offers ***generalizability***, which allows researchers not only to draw conclusions based on the selected individuals, but also to make inferences about the population characteristics.

Probability sampling methods combines these two characteristics and may include simple random sampling, stratified sampling and cluster sampling. However, for certain populations that are hard to locate and approach, such as homeless people, refugees, and sex workers, we would be satisfied with the first step as long as there is a method that can provide some sort of access to them, regardless of the sample's generalizability. Nonprobability sampling methods are used in such scenarios, including convenience sampling, purposive sampling, and quota sampling.

For years, researchers have been looking for a "miracle" sampling method that provides both accessibility and generalizability to hard-to-access populations. Targeted sampling and time-location sampling were among those under discussion, but their samples are generalizable only under extreme conditions and the sampling procedures are complex and resource-intensive. It was not until 1997, when Douglas D. Heckathorn, at Cornell University, published his paper on Respondent-driven Sampling (RDS) that researchers were able to accomplish this task.

RDS is a chain-referral sampling method. It works like snowball sampling but uses a dual incentive mechanism to stimulate the peer-driven recruitment process. Additionally, RDS is able to generate asymptotically unbiased population estimates from the sample.

Originally, RDS was used in 1994 to study injecting drug users as part of an AIDS prevention intervention in US. It was not used for HIV surveillance outside the US until 2003, but since then there has been a rapid increase in RDS studies, with more than a hundred empirical studies in over 80 countries targeting a wide range of hard-to-access populations, primarily HIV/AIDS-related high-risk populations such as injection drug users, men who have sex with men (MSM), sex workers and HIV infectors.

The accessibility of RDS has been proven by its successful implementations in recruiting samples of hard-to-access populations globally; however, the assumptions under which the population estimates were generated can hardly be valid in actual practice. Little attention was paid to notice the effect of violation of assumptions on the performance of RDS estimates. To address this issue, this thesis puts together our work on RDS during these years and reviews comprehensively the history and development of RDS methodology. We conducted a comprehensive evaluation on the effect of violating RDS assumptions on the performance of estimators, using an empirical MSM social network. Several methodologies have been developed to improve the validity and reliability of RDS in actual practice, including:

(i) A Web-based RDS recruitment system;

(ii) Methods for generating RDS estimates considering network directedness and degree reporting error; and

(iii) Improved RDS estimates with reported ego network data.

This thesis is organized as follows: In Chapter 1, I introduce basic statistical sampling techniques, including probability and nonprobability sampling methods, population inference, etc.; in Chapter 2, I discuss the pressing need and challenge for representative sampling of HIV/AIDS-related high-risk populations; it is suggested that readers who are not familiar with network theory refer to Chapter 3 before proceeding; in Chapter 4, I present the history and development of RDS implementation and theory; Chapter 5, 6 and 7 provide a summary and discussion of our work on the evaluation, implementation and improvement of RDS methodology.

All source codes and network visualizations in this book are available on request.

1

INTRODUCTION TO SAMPLING METHODS

Sampling is the method of studying the whole by the part. Typically, gathering information from all individuals in the population of interest by creating a census or a complete enumeration of all the values in the population is expensive, time consuming, or infeasible. For example, the 2010 US population census cost $13 billion, approximately $42 per capita [1]. Therefore, researchers often seek to select an affordable subset of individuals from the studied population, called the *sample*, to gain knowledge about the *population* and to generate estimates about population characteristics. As fewer individuals are included, the cost is lower, data collection is faster, and data accuracy and quality can be improved. For these reasons, sampling is widely used in social and medical research [2].

1.1 SAMPLING METHODS

One of the key purposes of sampling is to make predictions about the population from sample characteristics. A fair prediction is possible only when the sample is *representative* of the population, i.e., the characteristics of the sample, or the characteristics of the sample after adjustments, need to be approximately the same as the population.

To be representative, a sample must be drawn from a sampling frame in which each unit or person has a nonzero probability of being recruited. When the *selection probability* (or *inclusion probability*) of each individual in the population can be determined, the sampling design is called ***probability sampling***, which then allows researchers to adjust/reweight sample individuals according to their inclusion probability, and to generate estimates about population characteristics.

However, it is not always possible to determine the inclusion probability of all individuals from the population, particularly in the absence of a sampling frame or if there is no access to certain population members. For example, there is usually no such a list of names for all injecting drug users (IDUs) in a city. In such cases, researchers seek to select samples based on the characteristics of the population and the research purpose, such as recruiting volunteer IDUs from sites of high concentration [3], interviewing pedestrians on the street [4], etc. Such methods are called ***nonprobability sampling***. The sample characteristics are generally not representative to the population in nonprobability sampling as some members of the population may have a greater or less chance of being included in the sample. Consequently, sample results can hardly be generalized to a larger population. Nevertheless, nonprobability sampling methods provide convenient and fast access to the studied populations, when there is no sampling frame or little knowledge about the target population, or when population members are difficult to approach. In such cases, nonprobability samples can provide critical information at the early stage of research. Nonprobability sampling methods are also commonly used within qualitative research. I will in this work focus on discussing sampling methods from a quantitative research perspective within a positivistic epistemological tradition.

1.1.1 Probability sampling methods

Simple Random Sampling

In *simple random sampling* (SRS) [5-7], each individual in the target population is given an equal probability of being selected into the sample, such that each subset of individuals of the sample size has the same probability of being chosen. A random sample is usually selected with the assistance of random numbers generated by computer programs or a random numbers table.

SRS is the most common and basic probability sampling method. Because the inclusion probability is equal, each individual has the same chance of being in the sample, i.e., each individual is "equally represented" and is of equal importance, characteristics of the sample are reasonable good estimates for the population, minimizing bias and simplifying analysis of sampling results. However, the sampling frame of all individuals from the population is not always available and the SRS generally costs longer time and higher expenses. Additionally, it limits the flexibility of investigating subgroups of the population, such as ethnic minorities, residents of small districts, etc.

> Most sampling studies are conducted without replacement (*sampling without replacement*, SWOR), meaning that once an individual is taken, it is not allowed to be put back for recruitment again, i.e., each individual can appear in the sample maximum one time; contrarily, if selected subjects can be put back to "replace itself" and be ready for the next draw, the design is called *sampling with replacement* (SWR). Apparently, in SWR, each respondent can participate the study for many times. Most sampling studies are done without replacement, as it allows researchers to recruit as diversely as possible with limited resources. When the sample size is small compared to the population, SWOR is approximately the same as SWR, since the chance of recruiting a same individual twice is low. However, when the sample size is large, SWOR may affect the analysis method of sampling results fundamentally [8,9].

Systematic sampling

Simple random sampling can be cumbersome, especially when the sampling frame list is long or on-site survey is needed [10]. An easier, and perhaps more efficient alternative, namely *systematic sampling* [11-14], is to select every k^{th} individual from the population according to some ordering scheme, with a randomly selected starting individual from the first k individuals. Given that the ordering of population is reasonably homogenous, i.e., characteristics of interest are not correlated with the ordering, systematic samples are expected to function similarly to simple random samples. Given the population size N and the sample size n, the sampling interval k can be calculated by

$$k = N / n \tag{1}$$

Systematic sampling is especially useful when a good sampling frame is not available for on-site studies.

Systematic sampling is easy to perform and less susceptible to researchers' selection errors, however, it subjects to several limitations. First, it is often that the population cannot be evenly divided (suppose that you want to sample 10 out of 136 hospital staff, $k = 136/10 = 13.6$), leading to selection bias as inclusion probabilities are not equal for all individuals when k need to be an integer; second, as the sampling interval k is fixed, systematic sampling is very sensitive to cyclic patterns of the population, for example, to make a random household survey from a street by a beach, if all odd numbers are on the beach side (much more expensive), and all even numbers are on the other side, selecting every 10th no. with any random starting no. would end up with a sample of households from only on the beach side or non-beach side.

Stratified sampling

In *stratified sampling* [15,16], the population is divided into disjoint groups, i.e., *strata*, based on the characteristics of individuals (e.g., males and females). Each stratum is then sampled independently, e.g., by simple random sampling or systematic sampling. The sample size of each stratum is usually proportional to the size of the stratum in the population, or proportional to the variability within each stratum.

There are many advantages of stratified sampling. First, when simple random sampling or systematic sampling methods is used, each sample becomes a representative sample of the stratum where it comes from, enabling researchers to investigate statistic properties of subgroups that would otherwise be lost in a more generalized random sample. Second, it reduces sampling error as the population is divided into more homogenous subgroups. Third, it increases the flexibility of sampling methods used for different subpopulations. Lastly, it allows researchers to study minor groups by sampling equal number of individuals from strata of varying sizes. Comparing with simple random sampling and systematic sampling, the design and implementation of stratified sampling is more complicated and expensive. Stratified sampling is not useful when the population cannot be partitioned into exhaustively disjoint groups, or when there are no homogeneous subgroups in the population. Sometimes, it is difficult to determine the stratification variables and hard to identify appropriate strata.

Cluster sampling and multistage sampling

Expenses for sampling involving surveys to be conducted in remote areas, or creating large sampling frames can be unaffordable for simple random sampling, systematic sampling or stratified sampling. *Clustered sampling* [17-19], alternatively, reduces the cost of such studies by selecting a "random sample of groups or clusters" from the population and then sampling individuals within each of the selected groups. The two-stage process of cluster sampling is very similar to the stratified sampling method; however, they differ fundamentally in the inclusion of clusters or strata: cluster sampling draws a sample *of* groups, while stratified sampling draws samples *within* each group.

Cluster sampling is a fast, cheap and easy technique, it is especially useful when (a) a good sample list of population units is unavailable, but a list of potential clusters is easily to obtain; and (b) the cost of survey is associated with the distances of sampling units [20,21]. The major problem of cluster sampling is that the selected clusters may be very different from the general population. It is generally the case that individuals within a cluster share more common characteristics than those outside. Therefore, cluster sampling generally increases the variability of sample estimates above that of SRS and it requires a lager sample than SRS to achieve the same level of accuracy.

In some situations, cluster sampling is implemented with a *probability-proportional-to-size* (PPS) sampling design. In this method, the inclusion probability of each cluster is proportional to the size of the cluster, i.e., larger clusters have a greater probability of selection and smaller clusters have a lower probability. If the same number (c) of individuals within each cluster is randomly selected when the clusters are sampled with PPS, the inclusion probability of each individual in the population, $\Pr_i$, is identical:

$$\Pr_i \sim \frac{|C_i|}{\sum_{j=1}^{M} |C_j|} \cdot \frac{c}{|C_i|} \tag{2}$$

where $|C_j|$ is the size of cluster C_j and individual i belongs to C_i : $i \in C_i$.

When the target population is divided into clusters of more than one hierarchical level, a more complex form of cluster sampling, *multistage sampling* may be used to first sample the primary clusters (the clusters at the highest level), and then sample the secondary clusters from the sample of primary clusters, and so on, until the desired sample units (individuals) are ultimately reached.

1.1.2 Nonprobability sampling methods

When the sampling frame is not available and the inclusion probability of each individual in the population cannot be determined, or a random sample is too expensive, researchers seek samples that can maximize their knowledge about the population, regardless of the representativeness of the sample. Nonprobability sampling methods are the primary methods used for qualitative research. It is often used because the procedures used to select individuals for inclusion in a sample are much easier, quicker and cheaper when compared with probability sampling [22]. Typical nonprobability sampling methods include convenience sampling, purposive sampling and quota sampling.

Convenience sampling

The most extreme form of nonprobability sampling methods is *convenience sampling* [23-25], also called *accidental sampling* or *haphazard sampling*. In convenience sampling, researchers recruit persons who are most accessible in terms of location, time, and effort etc. Examples of convenience sampling include interviewing friends, mall intercept interviewing, visiting a sample of closest households, recruiting participants via banner survey on Websites, etc.

Facility-based sampling [26] is a form of convenience sampling which is used widely for studying HIV/AIDS-related high-risk populations. It may involve recruiting illicit drug users and commercial sex workers from correction facilities, finding injecting drug users (IDUs) from drug treatment centers or needle exchange programs, or interviewing MSM and commercial sex workers (CSW) from clinics.

Clearly, convenience sampling is the easiest method for gathering sample data (other than making up the data). It provides researchers useful information about the target population with minimum requirement of time and effort. As sample individuals are selected primarily because of their accessibility, which may often correlated with the characteristics to be examined in the study, selection bias can be high and sampling result can rarely be generalized to the population.

Purposive sampling

Purposive sampling [27,28], or *judgmental sampling*, is a “stricter” nonprobability sampling process in which the researcher selects respondents with a purpose in mind: the researcher decides what needs to be known and sets out to find people who can and are willing to provide the information from their knowledge or experience [29]. Instead of grabbing everyone who is passing by, in a purposive sampling an interviewer might ask those who seem to fall into their category, e.g., Hispanic women who look to be in their 30s to 40s, to participate in the study. The process is often consists of recruiting potential respondents, verifying inclusion criteria, and asking about willingness to participate. In most research, we sample with a “purpose”. Purposive sampling extends to a wide subcategory of sampling methods [30] such as *modal instance sampling* [31-33], *heterogeneity sampling*, *expert sampling* and *key informant sampling* [34-37]. Purposive sampling is also used to study extreme or deviant cases, such as outstanding success/notable failures, top of the class/dropouts, and extreme events/crises [38,39]. Other types of nonprobability sampling methods that involve the purpose of selecting certain groups can also be categorized as purposive sampling, such as *quota sampling*, *snowball sampling*, and the like [40]. However, due to their atypical design, I will introduce them separately in the following sections.

Purposive sampling is one of the most commonly used sampling methods for qualitative research. As individuals being recruited in the purposive sample are only those who “suit the purpose” and are mostly subjected to the researcher’s selection, it is very likely that certain subgroups are oversampled while some subgroups are excluded. Consequently, sampling results can hardly be generalized to the population.

Quota sampling

Quota sampling [41-44] resembles stratified sampling in nonprobability sampling methods. As in stratified sampling, the population is first segmented into non-overlapping groups such as males and females. What differs from stratified sampling is that while selecting sample individuals from each segment, non-random selection methods are used and the researcher can decide on the *quota* (the proportion of individuals from each segment to be sampled) deliberately, independent of the population characteristics. Unlike stratified sampling, in quota sampling, as long as the sample size of a segment reaches the desired sample size, the recruitment process will stop in this segment and move on to other unfinished segments.

Quota sampling is a relative flexible sampling process, as different sampling methods can be applied in the second stage. It also allows the researcher to recruit a quasi-representative sample which accounts for the characteristics used for generating population segments. However, like other nonprobability sampling methods, the nonrandom process of selecting sample individuals in quota sampling may result in large selection bias, study results should be interpreted only within the sample and cannot be generalized to the population.

Snowball sampling

Snowball sampling is one of the best known form of *chain referral sampling*. In snowball sampling [45,46], several initial subjects are identified and recruited as seeds. Earlier participants are then asked to refer population members they know for further recruitment, forming a "rolling snowball" sample that increases in size. Typically, the seeds are considered to be from wave 0, respondents referred by those from wave t but those that are not in wave $0 \sim t$ form wave $t+1$, i.e., the non-seeds who are named by wave 0 form wave 1, those who are named by wave 1 but neither in wave 0 nor wave 1 form wave 2, and so on. When the maximum number of referrals is allowed, snowball sampling works much like a breath-first-search (BFS) in computer science, with sample size expanding explosively with distance (wave) to initial respondents. Originally, Goodman (1961) proposed to randomly choose the initial seeds in snowball sampling [45], however, in practice this is very difficult or impossible without a sampling frame. Common practice of snowball sampling are started with a handful initial participates which are collected through convenience sampling or purposive sampling.

Snowball sampling suffers from several other drawbacks that can lead to large bias. The first source of bias is called *volunteerism* [46], which indicates that respondents ended up in the sample tend to be more cooperative and accessible. The second is *homophily*, which is a universal pattern for social interactions and it means that social affiliations (relationships) are more likely to form among individuals sharing similar characteristics. Given the existence of homophily, the composition of an earlier wave biases the subsequent wave [47,48]. The third is *differential recruitment*, a term used to represent any act of non-random referring. For various reasons, respondents may choose to refer some peers than others; examples include recruiting close friends than unfamiliar acquaintances, unrevealing certain group members to protect them from exposure, etc. The last source of bias is due to *contact hubs*, who are population members maintaining large social network sizes. Obviously, contact hubs will be overrepresented in the snowball sample, as they are more likely to be referred through their friends. If these hubs differ largely from other individuals on the studied characteristics, the sample will be largely biased from the true population. These drawbacks have been recognized by researchers from very early times [45,49,50] and snowball samples of hidden populations are generally considered as sort of convenience samples for which no claims of representativeness can be made [26,48].

Targeted sampling

Targeted sampling (TS) was initially designed by Watters and Biernaki (1989) for the purpose of efficiently identifying and recruiting injecting drug users [51]. As described by the authors, combining aspects of "street ethnography, theoretical sampling, stratified sampling, quota sampling, and chain referral sampling", targeted sampling provides a flexible procedure for sampling hidden populations in urban settings. It involves [52]: (i) extensive formative assessments and ethnographic mapping to identify places of sufficient target population concentration, such as night clubs and street corners; (ii) developing target enrollment plans (quotas) for each location; and finally (iii) sampling in those areas based on the quotas established to approximate the makeup of the population. Carlson et al (1994) enhanced targeted sampling with the addition of estimation of density of population members in the target areas and the introduction of proportional sampling quotas [53]. The similarities between cluster sampling and targeted sampling are clear: without the use of nonrandom sampling techniques, targeted sampling is merely a cluster sampling in which the enumerated locations where target population concentrates are primary clusters.

A more complex form of targeted sampling, *time-location sampling* (TLS, also called *venue-based sampling* or *time-space sampling*), is a variant of multistage sampling which is popularly used in the study of hard-to-access populations, such as men who have sex with men (MSM) or injecting drug users (IDUs) [54,55]. Similarly to TS, in TLS, a complete list of the places where the target population congregates is created through accumulated historical information or key informant interviews. The difference between TSL and TS is that the list also contains time information about when these locations are visited. After obtaining the sampling list, a random set of venue-day-time units (VDTs) are selected, e.g., a VDT can be a given location, between 10 pm and 2 am on Tuesday night. Finally, the selected locations are visited during the day and time specified, and members of the target population are either fully recruited or systematically sampled.

Given certain assumptions, such as a comprehensive list of all locations, accurate estimates on the population density and time events, etc., the inclusion probability of each individual can be calculated [54], which makes either TL or TLS a probability sampling method. However, for hidden populations, as discussed in [26,55,56], there are many difficulties for TL or TLS samples to be representative: (i) to list all locations that are frequented by the target population is very labor intensive and time consuming; (ii) it is very difficult to measure the inclusion probability of those who visit the VDTs, and it is almost impossible to estimate the probability of missing members who do not attend any of the listed locations; (iii) some locations offer little privacy and the accuracy of self-reported data is always questionable.

1.2 POPULATION INFERENCE

1.2.1 Equal inclusion probability and sample representativeness

The common feature of probability samples is that individuals in the sample are selected by means of a probability scheme such that the inclusion probability of each individual in the population, $\Pr_i$, can be determined. Table 1 lists the calculation of inclusion probability for different sampling methods, as well as relevant notations [57].

In most cases, the inclusion probability of each individual is the same, that is to say, all units in the population have the same chance of being recruited and their opportunity of being "represented" in the sample is the same. As a result, the sample itself is "representative" of the population. If we repeatedly draw samples from the same population, any difference between these samples and the population is merely due to randomness. The population mean ($\bar{X}$), variance (s^2) and proportion (p) of any certain characteristic X can then be estimated by the sample, as shown in Table 1.

1.2.2 Unequal inclusion probability and sample weight

There are, however, situations in which the inclusion probability is not the same among individuals. For example, in a stratified sampling, if the population is divided into two strata, the majority (90% of the population) and the minority (10% of the population), because sufficient samples are needed to study characteristics of the minority group, researchers might want to use SRS to recruit the same number of units from this stratum as from the majority stratum. Consequently, each individual from the minority group would be 9 times more likely to be selected into the sample.

In the above example, we will say that the minority is “*overrepresented*” or “*oversampled*”, and the majority is “*underrepresented*” or “*undersampled*” in the sample. As the sample will contain 50% each of the two groups, it does not represent the population anymore. A common practice to adjust the sample, is to assign each sample unit i a *weight*, w_i, which indicates its importance in the population, by weighting each sample units, the resulted adjusted sample would again be representative. The weight is usually the inverse of the inclusion probability:

TABLE 1 POPULATION INFERENCE OF PROBABILITY SAMPLING METHODS

	Notation	Inclusion probability	Population inference
Simple Random Sampling	N population size n sample size n_A sample units with property A	$\Pr_i = n/N$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$, $p = n_A / n.$
Systematic Sampling	$k = N/n$ sampling interval r starting unit	$\Pr_i = n/N = 1/k$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$, $p = n_A / n.$
Stratified Sampling	S number of strata N_i number of population units in stratum i n_i sample size within stratum i	$\Pr_{ij} = n_j / N_j$ inclusion probability for unit i in stratum j	$\bar{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$ *(stratum)*, $\bar{X} = \frac{1}{N}\sum_{j=1}^{S}\bar{X}_j N_j$, $s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_j)^2$ *(stratum)*, $s^2 = \frac{1}{N-1}\left[\sum_{j=1}^{S}(N_j - 1)s_j^2 + \sum_{j=1}^{S}(\bar{X}_j - \bar{X})^2 N_j\right]$, $p_j = n_{A,j} / n_j$ *(stratum)*, $p = \sum_{j=1}^{S} p_j N_j / N.$
Cluster Sampling (two-stage)	G number of clusters g number of sampled clusters N_k number of population units in cluster k n_k sample size within cluster k	$\Pr_{ik} = (g/G)(n_k / N_k)$ inclusion probability for unit i in cluster k	$\bar{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}$ *(cluster)*, $\bar{X} = \frac{G}{N}\frac{\sum_{k=1}^{g} N_k \bar{X}_k}{g}$, $s_b^2 = \frac{\sum_{i=1}^{g}(N_i \bar{X}_i - \frac{N}{G}\bar{X})^2}{g-1}$ *(between)*, $s_k^2 = \frac{\sum_{i=1}^{n_k}(x_{ik} - \bar{X}_k)^2}{n_k - 1}$ *(within)*, $s^2 = (\frac{G-g}{G})(\frac{1}{gG(\frac{N}{G})^2})s_b^2 + \frac{1}{gG(\frac{N}{G})^2}\sum_{i=1}^{g}(N_i)^2(\frac{N_i - n_i}{N_i})(\frac{s_i^2}{n_i})$, $p_k = n_{A,k} / n_k$ *(cluster)*, $p = \frac{G}{gN}\sum_{k=1}^{g} N_k p_k.$

$$w_i = \frac{1}{\Pr_i} \tag{3}$$

As individuals with a high inclusion probability will be more often recruited, it is not as important in the population as in the sample. The higher the $\Pr_i$, the lower would the w_i. Consequently, for any probability sample, the population mean of a certain characteristic X, can be estimated by the weighted sample mean:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{4}$$

where x_i is the value of sample unit i and n is the sample size.

1.2.3 Toward a more representative nonprobability sample

Nonprobability samples, as illustrated above, are not eligible for inference from the sample to the general population, due to unknown inclusion probabilities. That is to say, if we collect a convenience sample of 100 IDUs from drug treatment centers, and from the sample we can observe that 30% of them are females, the way how we collected the sample limits our ability to estimate the proportion of females among general IDUs, as females may be either overrepresented, if they are more likely to be recruited in a drug treatment center, or underrepresented, if more female IDUs are hidden to the public.

The lack of generalizability is the major drawback of nonprobability sampling methods, despite their many advantages such as ease and speed of implementation, improved access, and economic savings. In many situations, when nonprobability sampling is the only option for the study, researchers aim to increase the representativeness of the sample as much as possible, to reduce the risk of ending up with a sample that is significantly different from the population.

Representativeness can be improved by adding "randomness" to the sampling procedure, for example, when recruiting sample individuals in each quota in a quota sampling, a coin may be flipped (or something similar) to decide whether a target member should be interviewed. Another approach, which is related to randomness, is to add "diversity" to the sample, meaning that the sample should be as varied as possible. The best example of diversity in nonprobability sampling is in snowball sampling. When a random set of seeds is not available, researchers can try to recruit a diverse set of initial participants to start the sampling to ensure that the sample gets rid of homophily and community structures among the target population's social network. These procedures cannot, however, enable us to make statistical inferences about the target population, and such sampling results should always be interpreted with caution.

To summarize, Table 2 briefly lists the major advantages and disadvantages of the sampling methods introduced in this chapter [58].

TABLE 2 ADVANTAGES AND DISADVANTAGES OF SAMPLING METHODS

	Advantages	Disadvantages
Probability sampling methods		
Simple Random Sampling	Sample is highly representative; easy to implement; simplifies data interpretation and analysis.	Require a complete list of population members which may be expensive or impossible to obtain and update; cost can be high; time-scale may be too long, data/sample could change; subgroup members may be small in the sample.
Systematic Sampling	Sample is representative and can sometimes be more efficient than a SRS sample; easy to implement.	Require a sampling frame; Vulnerable to periodicities; hard to quantify data accuracy; can result unequal selection probabilities for population members.
Stratified Sampling	Can ensure that specific groups are represented, correlations and comparisons can be made between subsets; very flexible and applicable, can be combined with other sampling methods;	More complex, requires greater effort than simple random; can be expensive; strata must be carefully defined, size of each stratum may be unknown.
Cluster Sampling	Possible to select randomly when no single list of population members exists, but local lists do; can save expenditures for sampling and listing.	Bias can be high if the selected clusters are very different from the population;
Nonprobability sampling methods		
Convenience Sampling	Easy, fast and inexpensive way of recruiting respondents.	Can be highly unrepresentative.
Purposive Sampling	Provides a wide range of nonprobability sampling techniques. It is easier to get a sample of subjects with particular characteristics.	Can be highly unrepresentative.
Quota Sampling	Ensures selection of adequate numbers of subjects with appropriate characteristics.	Not representative.
Snowball Sampling	Provide access to members of groups where no lists or identifiable clusters even exist (e.g., drug abusers, criminals)	Not representative.
Targeted Sampling	Can be combined with different sampling methods; Ethnographic mapping helps researchers to gain knowledge on the target population.	Can hardly be representative. Expensive and requires a long-time for maintaining the (time/) location list.

2
CHALLENGE OF SAMPLING FOR HIV SURVEILLANCE

HIV (Human immunodeficiency virus), the virus causes AIDS, "acquired immunodeficiency syndrome", has become one of the world's most challenging health and development problems. Ever since AIDS was first recognized by the Centers for Disease Control and Prevention (CDC) in US in 1981, the epidemic has reached almost every corner of the world, with a highly disproportional distribution, raising inequalities between North and South, as well as between rich and poor, men and women, black and white, homosexuals and heterosexuals [59-61].

HIV is a leading cause of death worldwide and the top one cause of death in Sub-Saharan Africa. By the end of 2010, AIDS-related diseases had cost 30 million lives, with 1.8 million people died in 2010 alone, more than the population of Netherland or Chili [60-62]. With 2.7 million people newly infected in 2010, there are 34 million people living with HIV in the world, almost all (97%) in low- and middle-income countries, particularly in sub-Sahara Africa (68%) [60,61].

2.1 HIGH-RISK POPULATIONS

Several key populations face higher risk of HIV transmission, such as sex workers (SWs), men who have sex with men (MSM), and injecting drug users (IDUs), see Figure 1. Recent study has shown that the pooled HIV prevalence for MSM ranged from a low of 3.0% in the Middle East and North Africa region to a high of 25.4% in the Caribbean [63]. For IDUs, HIV prevalence is as high as 25% in Eastern Europe and Central Asia and 16% in the rest of Asia. In sub-Sahara Africa, unprotected paid sex, sex between men, and the use of contaminated drug-injecting equipment are estimated to account for 33% of new HIV infection in Kenya and 40% in Ghana [60,61].

2.2 CHALLENGE OF SAMPLING HIDDEN POPULATION

The role of HIV/AIDS related high risk behaviors in the evolvement of HIV sub-epidemics consequently makes the detailed information on distribution and characteristics of these behaviors critical for the deployment of HIV surveillance and prevention programs. However, the nature of such high risk behaviors prohibits the use of traditional methods for investigation.

First, the lack of sampling frames limits the use of probability sampling methods. Obviously, there is no list containing all individuals who identify themselves as SWs, MSM or IDUs. Even if it would be possible to stretch a random sample from general population, this can be highly inefficient, as the chance of meeting a person who practices a risk behavior is small, not to mention that they are also not willing to disclose their identity.

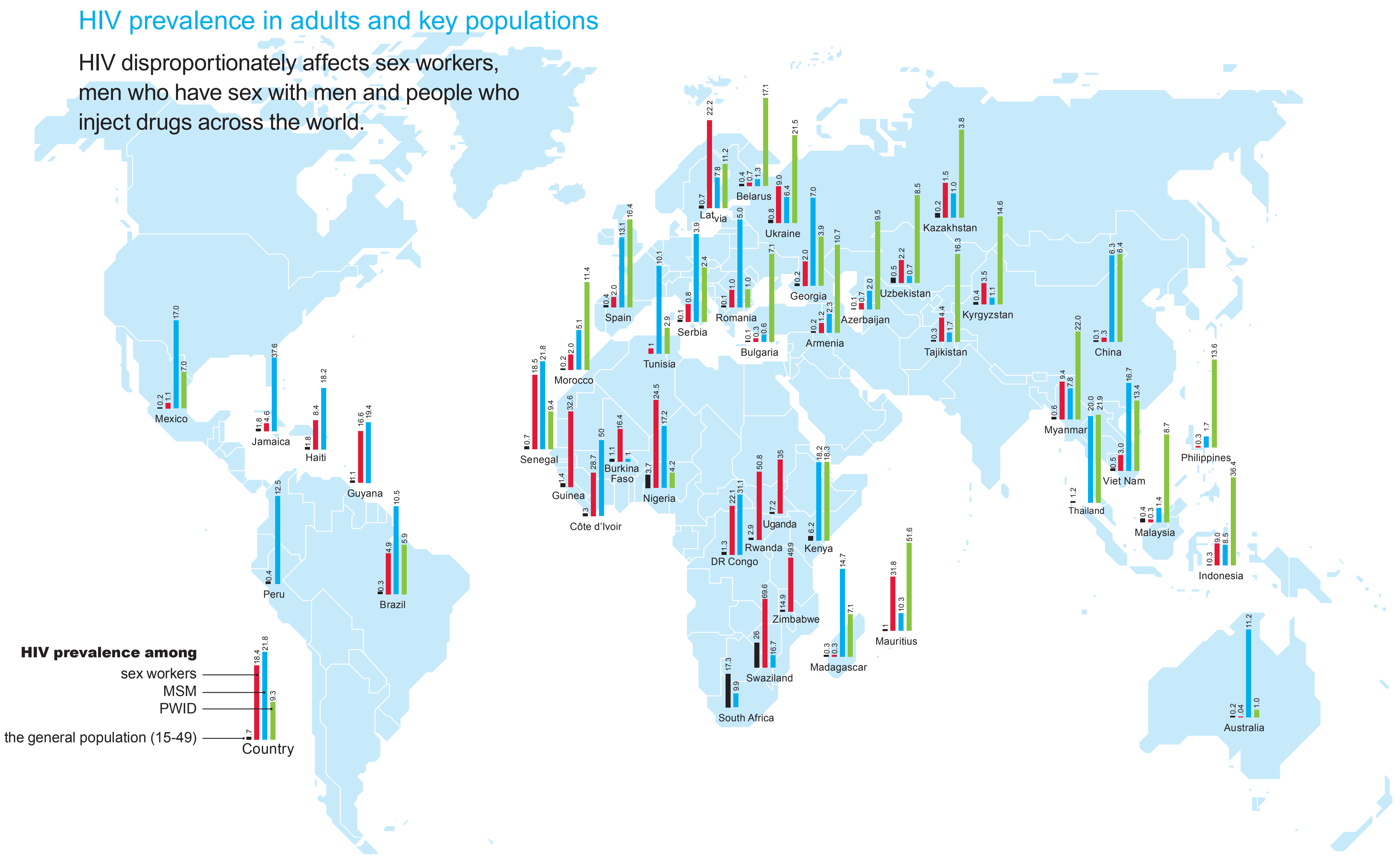


FIGURE 1 HIV PREVALENCE IN ADULTS AND KEY POPULATIONS. SOURCE: UNAIDS 2012 GLOBAL REPORT.

Secondly, these risk behaviors carry high social stigma and are considered illegal in many countries[*], making it impossible for researchers to sample them systematically. Fears of stigma and legal consequences, individuals of risk populations (particularly those known of being infected with HIV) often choose to live socially isolated from general population, or conceal their identities from friends and families. For these reasons, they are often called "*hard-to-access population*", or "*hidden population*". In public health, hidden populations of interest are primarily composed of SWs, MSM and IDUs.

Third, sampling of HIV related high risk populations often involves investigating highly sensitive issues, challenging the ethic and privacy concern of participants. HIV is a blood borne disease. Unclean needles and unprotected anal or vaginal intercourse are the risk behaviors of interest [66,67]. One particular goal with HIV surveillance and prevention, is to understand the epidemiological features of HIV infection and the evolvement of the epidemic in sub-populations and also as a pandemic [68-70]? What is the role of high risk behaviors in driving this epidemic? Which are the most effective prevention and intervention strategies? The answer to these questions requires examination of very sensitive issues, such as number of sex partners, frequency of unprotected anal/vaginal sex, drug injecting activities, etc. Collecting of such information undoubtedly increases difficulties in accessing target individuals.

With these limitations, representative samples of hidden population is extremely rare (exceptions include the US 2000 and 2010 Census where same-sex partners of household members could be reported [71,72]), and traditional methods for studying HIV/AIDS-related high-risk populations are mostly key informant sampling, targeted sampling, and snowball sampling. As discussed in Chapter 1, all these are nonprobability sampling methods and suffer from various source of bias, sampling results cannot be generalized to the population.

2.3 A NOTE ON ACCESSIBILITY AND GENERALIZABILITY

To summarize in short, HIV/AIDS-related high-risk populations are extremes from general population: the ***lack of a sampling frame*** prohibits probability sampling to draw a representative sample, additionally, the nature of these populations and the need of sensitive information for HIV surveillance studies make them exceptionally ***hard to access***.

Accessibility and generalizability, a miracle method would provide both, that allows researchers to, first, gain access to these population members, obtain reliable biological and risk behavior information with high response rate, second, be able to make population inference about the risk populations, which may guide to set up efficient prevention and intervention HIV programs.

This method is called "respondent-driven sampling" (RDS).

[*] For example, to date there are 112 countries in the world where sex work (prostitution) is deemed illegal, with penalties including fan, in prison, flogging, etc. [64][65]

3
NETWORK IN A NUTSHELL

Getting familiar with a few network concepts would be helpful for understanding the respondent-driven sampling method. This chapter gives a short introduction on network theories, including basic network types, a glossary of network properties and a review on the development of mathematical network models.

3.1 INTRODUCTION

Humans are social beings. We get involved in the society by interactions with other people: hanging out with friends, collaborate with colleagues, share emotions with families, ask passengers for directions, etc. In mathematics, all these can be depicted by networks with nodes being the actors (human individuals) and links (edges) being the interactions between them.

Let $V = \{v_1, v_2, v_3 ..., v_N\}$ be the set of nodes and $E = \{e_{ij}\} \subseteq V \times V$ be the set of links between the $|V| = N$ nodes, where $e_{ij} = 1$ represents the existence of the interaction/ relationship (e.g., friendship) between i and j and $e_{ij} = 0$ when there is none, then the network is defined as $G = (V, E)$.*

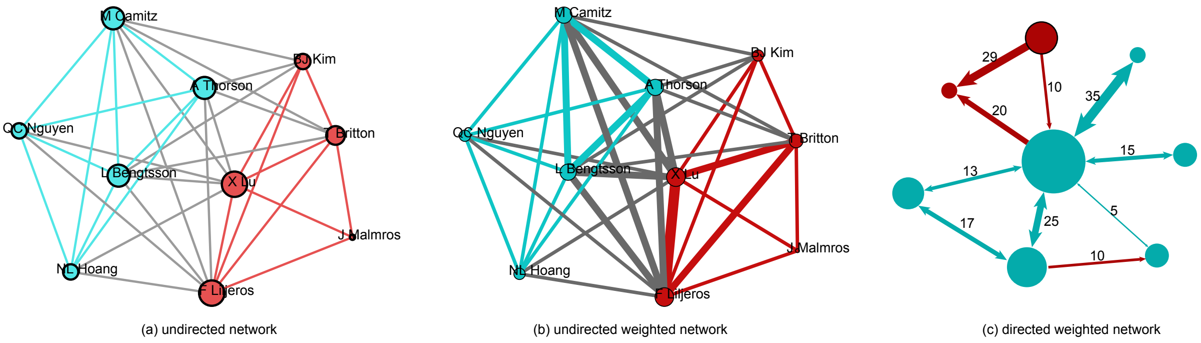
The simplest and most common form of G is an *undirected network*, in which the link is assumed to be *reciprocal*: for any $1 \le i, j \le N$, $e_{ij} = e_{ji}$. Many social interactions are undirected, such as marriage, co-authoring, neighborhoods etc. Some other relationship may only go in one direction. For example, subordination, emailing, telephone communication, etc. When there are directed links in the network, i.e., for some $1 \le i, j \le N$, $e_{ij} = 1$ but $e_{ji} = 0$, G is called an *directed network*. Links may vary not only the direction, but also the strength, such as years of marriage, the frequency of contact, number of phone calls. The strength of a link is usually represented by a value, called its weight, w_{ij} . When weight is used to model the network, G is called a *weighted undirected/directed network*. Examples of different networks are shown in Figure 2.

3.2 PROPERTIES OF NETWORKS

3.2.1 Network connectivity

A network is *connected* if there is a path between any pair of nodes, i.e., all nodes are reachable through any other nodes. In snowball sampling, a connected network ensures that all individuals in the target population can be reached from any initial seed. If not all nodes are connected together, the component which contains the largest fraction of connected nodes in the network, is called the ***Giant Connected Component*** (GCC).

* For simplicity, in this thesis we do not consider self-loops in the network.



(a) undirected network (b) undirected weighted network (c) directed weighted network

Figure 2 Basic network types

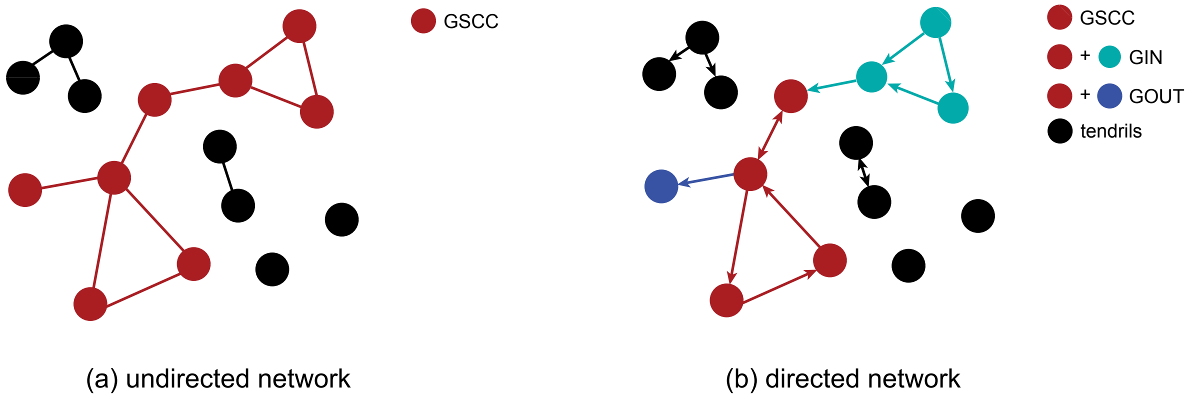
The definition of GCC is straightforward for undirected network; however, when the network is directed, there are various types of components based on the accessibility of nodes [73]:

Giant Weakly Connected Component (GWCC): GWCC is the GCC of the network when the directions of links are ignored.

Giant Strongly Connected Component (GSCC): GSCC is the largest fraction of nodes which are reachable from each other, i.e., for any pair of nodes i and j in the GSCC, there is a directed path of links from i to j. Apparently, GSCC is the guarantee of positive inclusion probability for chain-referral sampling methods, whereas in GWSC, some nodes may not be reachable from certain nodes.

Nodes that are reachable from a GSCC form the ***Giant Out-Component*** (GOUT), and nodes from which the GSCC is reachable, is called the ***Giant In-Component*** (GIN). The rest of nodes, which are not part of GWCC, form disconnected components and are called ***tendrils***. The decomposition of a disconnected directed network is (see Figure 3):

$$\text{G=GWCC+Tendrils=GSCC+(GOUT-GSCC)+(GIN-GSCC)+Tendrils} \tag{5}$$



(a) undirected network (b) directed network

Figure 3 Components in undirected network and directed network

3.2.2 Degree

Degree, also called connectivity, is the number of links a node is incident on, or say, the number of neighbors a node connects to. The degree of node i in a network is often denoted as d_i, which can be calculated by:

$$d_i = \sum_j a_{ij} \tag{6}$$

When the network is directed, the degree is further divided into *outdegree* and *indegree*, representing the number of links initiated by a node, or the number of links a node is headed to, respectively.

$$d_i^{out} = \sum\nolimits_j a_{ij} \tag{7}$$

$$d_i^{in} = \sum\nolimits_j a_{ji} \tag{8}$$

Degree provides the most basic information about the network property of a node and is often the first measurement I check for a network analysis. In some studies, degree is a measurement of node importance or position as comparing to other nodes. For example, in an undirected friendship network, degree represents the number of friends a node has, larger degree would indicates the individual is very socially active and is well known by others. In a human sexual contact network, a higher degree means that the person has many sex partners, a signal for potentially being at higher risk of STD (sexual transmitted diseases) infection and for being the hubs for disease transmission.

3.2.3 Degree distribution

The proportion of nodes with a given degree is characterized as the *degree distribution* (frequency distribution of degrees, or, the histogram of degrees):

$$P(k) = \frac{n_k}{N} \tag{9}$$

where n_k is the number of nodes with degree k in the network.

Degree distribution is one of the most fundamental characteristics of a network and one of the driving forces for the blooming of network science research in the past decade [74-78], see the following introduction on network models.

3.2.4 Shortest path and the small-world experiment

A *shortest path* from i to j is the path with the minimum sum of the weights on the links [79]. There may be more than one shortest path between two nodes. When the network is not weighted, the distance of a shortest path corresponds to the minimum number of intermediate links between the two nodes. The maximum distance between any two nodes in network, given the network is connected, is called the *diameter*. A small average shortest path length in a friendship network would indicate that a node can get to know any remote stranger by a limited number of introductions through his friends, the friends of his friends, and so on.

In 1967, Milgram et al [80,81] conducted a famous experiment on examining the average path length of social network of people living in US: randomly selected individuals were asked to send a letter to a target contact person in Boston through their acquaintance network: if the recipient knew the target person he could send the letter directly, otherwise he was instructed to send the letter to one person he know and who he thought is most likely to know the target person. Eventually, 64 out of 296 letters did reach the destination. The experimental result showed that the average path length between two randomly picked Americans was 5.2. This phenomenon, later called “six degrees of separation” [82], was also found in many other societies [83,84], revealing the fact that social networks are much better connected than previously assumed.

3.2.5 Clustering

Clustering, or network *transitivity*, is a property that many networks have in common: if j and k are neighbors with i, it is very likely that j and k are also neighbors of one another, i.e., the network has a larger probability of forming triangles. This phenomenon is very often observed in social networks and is interpreted as "the friend of your friend is also likely to be your friend", indicating that individuals tend to cluster together when we look into social relationships. Clustering is quantified by the clustering coefficient:

$$\begin{aligned} C &= \frac{3\times \text{number of triangles in the network}}{\text{number of connected triples}} \\ &= \frac{6\times \text{number of triangles in the network}}{\text{number of paths of length two}}, \end{aligned} \tag{10}$$

where a path of length two is a directed path starting from a specific node. For a fully connected network, $C=1$, and for many real-world networks, C ranges between 0.1 and 0.5 [74,85,86].

There is a direct side effect of clustering for chain-referral sampling methods. Suppose that a respondent-driven sampling (see Chapter 4) is implemented on a network with large C, when a participant i has passed coupons to his friends (e.g., j, k) and all his friends have also attended the interview and have received more coupons, at the next step, all these friends (j and k) are asked to distribute their coupons to their friends, it is very likely that those who receive coupons at this stage are the same friends with i, who have been recruited by their mutual friend (i) previously. Such recruitments cycling in triangles would consequently lead to low response rate for SWOR and inaccurate estimates if the method is assumed to be SWR.

Saturation is another side effect for sampling on networks with clustering. As friends are clustered together, the recruitment can rarely get out of their friendship clusters and has a lower chance of penetrating into other parts of the network, the distribution of new coupons will end up with those who participated previously and will thus risk the sampling waves to stop early.

3.2.6 Community structure

For many real-world networks, friends of friends are likely to become friends, as the clustering coefficient measures; on a more macro level, groups of individuals may interact more often than others, forming various types of community structures on networks. Typically, a group of nodes are said to form a community if there is a higher density of links within the group and a lower density of links between groups [74,85]. A clear identification of communities in a network would undoubtedly be beneficial for us to understand and investigate the network more effectively.

Finding of communities on a network may be nature and intuitive [87], for example, friendship networks can be divided into groups based on the age, interests, occupation, or ethnics, scientific citation network can be divided into groups based on research areas, transportation networks can be divided into groups based on locations, etc. It is however not always easy to find an obvious division of community structures for a network due to the unknown number of communities to be determined and the unequal size and density of communities [88-91]. Various community detection algorithms have been developed during the past years, such as hierarchical clustering [91], modularity maximization [91-93], spectral partitioning [91], and random walk mapping [87,88].

3.2.7 Mixing patterns

Most of the previous discussion focuses on the network topological structure. Moving forward, a more complicated way of thinking is to bring together also the characteristics of nodes, such as gender, nation of birth, marriage status. When we try to investigate both the network position (e.g., degree) and nodes' properties, a good start is to summarize the frequency of links connecting between different types of nodes.

Suppose we are studying a marriage network in which each link represents a married couple, and each node is associated with gender and ethnicity, then the frequency of type of links represents the likelihood for a person to choose a partner within or outside his/her own ethnicity. In a study of 1,958 couples in San Francisco, California, Catania et al. found that participants appeared to choose their partners preferentially from their own race, see [94,95]. This phenomenon of associating preferentially with people who are similar to themselves is found to be another common phenomenon in social networks, and it is called ***assortative mixing*** or ***homophily***.

Homophily can be quantified by the probability that nodes connect with neighbors who are similar to themselves with respect to the studied property A rather than that they connect randomly [48,96-98]. Let h_A be the homophily for nodes with property A, it holds that [99]

$$S_{AA}^{*} = h_A + (1 - h_A)P_A^{*}, \tag{11}$$

where S_{AA}^{*} is the proportion of type $A \rightarrow A$ links among all links originating from type A nodes, and P_A^{*} is the proportion of type A nodes in the network. Consequently, when $h_A = 1$, we have $S_{AA}^{*} = 1$, meaning that all type A nodes only connect with type A nodes themselves and there is no cross-group connection between type A nodes and type B nodes; when $h_A = 0$, we have $S_{AA}^{*} = P_A^{*}$, meaning that type A nodes connect to other nodes proportional to their proportions in the population, there is no preference of link formation regarding property A.

Sometimes it is of interest to check whether nodes with a lot of connections prefer to connect with others that are also highly connected, a special case of assortative mixing when the degree of nodes is considered as the type of nodes. This is often measured by the ***degree correlation***, or ***assortativity ratio*** [100-102]:

$$\gamma = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}, \tag{12}$$

Where M is the number of links in the network, and j_i, k_i are the degrees of nodes at the end of the i^{th} link, $i = 1,...,M$.

3.3 NETWORK MODELS

3.3.1 Regular networks

One of the simplest network models is the so called “regular” network, in which nodes are associated with exactly the same number of links, such as a square lattice where each node is connected to its nearest neighbors in the four directions, or a ring on which each node is connected to the same number of nodes on each side. See Figure 4.

3.3.2 Erdős–Rényi random network

If links between nodes are formed in a purely random way (the Erdős–Rényi network model [103,104]), most of nodes will have a similar degree, while a few will be connected with either too few or too many nodes, resulting a “bell-shaped” degree distribution[*]. Given p the probability of establishing a link between any pair of nodes, the probability of a node having degree k is:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{z^k e^{-z}}{k!}, \tag{13}$$

where $z = p(N-1)$ is the average degree of the network.

The ER model (see Figure 7) captures one important characteristic of real-world networks, i.e., the average shortest distance. As links are allowed to randomly connect with any others in the network, ER networks have small average shortest path lengths. It can be proven that, given the network is connected and the average degree $\overline{k}$ is fixed, the average shortest path length of an ER network scales with the logarithm of its size [75]:

$$\ell_{ER} \sim \frac{\ln N}{\ln \overline{k}}. \tag{14}$$

However, ER model fails in creating networks with high clustering coefficients. The scaling of clustering coefficient follows [75]

$$C_{ER} \sim \frac{\overline{k}}{N}. \tag{15}$$

When the network size is large, C_{ER} approximates to zero.

3.3.3 Small-world networks

In an attempt to generate networks with small shortest path lengths, as well as high clustering coefficients capturing real-world network property, Watts and Strogatz (1998) proposed the WS model that interpolates between a ring lattice and a random graph [105]. The model starts with a ring lattice where nodes are placed on the ring and each node is connected to its first $\overline{k}$ neighbors ($\overline{k}/2$ on each side). Then links of each node from the clockwise side (or counterclockwise side) are *rewired* to randomly chosen nodes with probability p, self-loops and duplicate links are excluded.

Apparently, by varying p, the WS model forms networks between extremely regular $(p=0)$, and random $(p=1)$. Let $\ell(p)$ and $C(p)$ be the expected average shortest path length and clustering coefficient for a WS model with rewiring probability p. It is

[*] Strictly speaking, the degree distribution of an ER random network is Binomial, which can be approximated by a Poisson distribution for large N and constant Np.

not hard to find out that when $p=0$, $\ell(0) \simeq N/2\bar{k} \gg 1$ and $C(0) \simeq 3/4$, that is, ℓ scales linearly with the network size, and C is a large constant. When $p \to 1$, $\ell(1) \simeq \ln N / \ln \bar{k}$ and $C(1) \simeq \bar{k}/N$, that is, ℓ scales logarithmically with N, and C decreases with N. It has shown that for a large range of $p \in (0,1)$, WS network holds both small shortest path length and large clustering coefficient, a property with which the networks are called "***small-world networks***" [74,105-107]. Actually, even with very small rewiring probability p, a few rewired links would be enough to create "short cuts" in the ring lattice to decrease ℓ significantly, with little effect on C.

Instead of rewiring, adding a few random short-cuts to the existing ring lattice would also produce small-world networks [74,108,109].

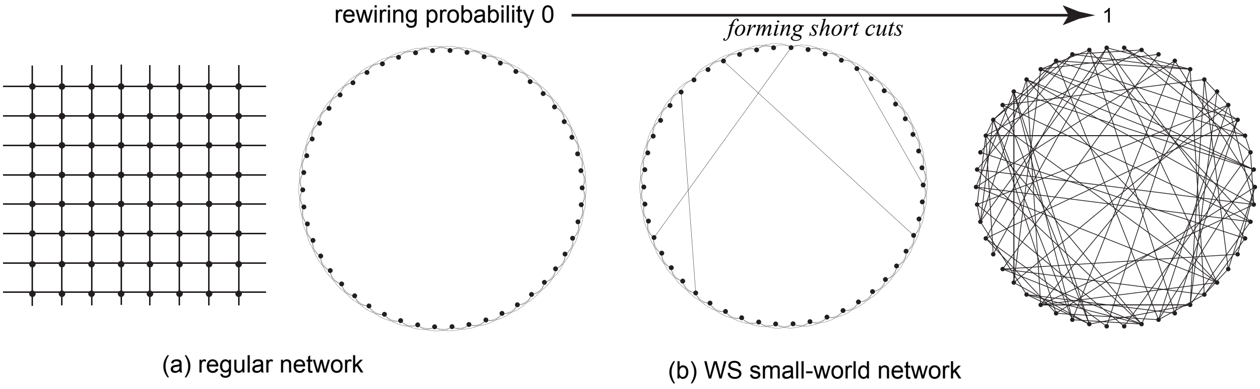


FIGURE 4 REGULAR NETWORKS AND WS SMALL-WORLD NETWORKS

3.3.4 Scale-free networks

The WS model successfully characterizes the small-world effect of most real-world networks, however, its degree distribution is similar to a random graph, i.e., the topology of the network is relative homogenous, all nodes having approximately similar number of links. On the contrary, in most large-scale real-world networks, nodes are rather heterogeneous, that is, most nodes have very few connections but a small number of particular nodes have excessively many connections. More precisely, the degrees of nodes in large complex networks follows a "power-law": the proportion of nodes with degree k decreases dramatically with k:

$$P(k) \sim k^{-\alpha}, \tag{16}$$

where α is positive and typically ranges in $2 < \alpha < 3$ [74,77,110,111]. Networks with power-law degree distributions are called "scale-free" networks since the function form $P(k)$ remains unchanged to within a multiplicative factor under a rescaling of the independent variable k [74]: $P(ak) \sim (ak)^{-\alpha} = bk^{-\alpha} \propto P(k)$. The cumulative probability distribution of a power-law is also a power-law, with a less than one scaling exponent: $P(x \geq k) \sim k^{-(\alpha-1)}$.

The most common way to visualize a power-law degree distribution is to plot it on an axis with *x*-axis being the logarithm of degree k, $\log_{10} k$ and *y*-axis being the logarithm of $P(k)$, $\log_{10} P(k) \sim -\alpha \log_{10} k$, the plots is a straight line with a negative slop $-\alpha$. Many real-world networks, including human sex networks, mobile communication networks, World Wide Web links, etc., are found to have power-law (or similar to power-law) degree distributions, see Figure 5.

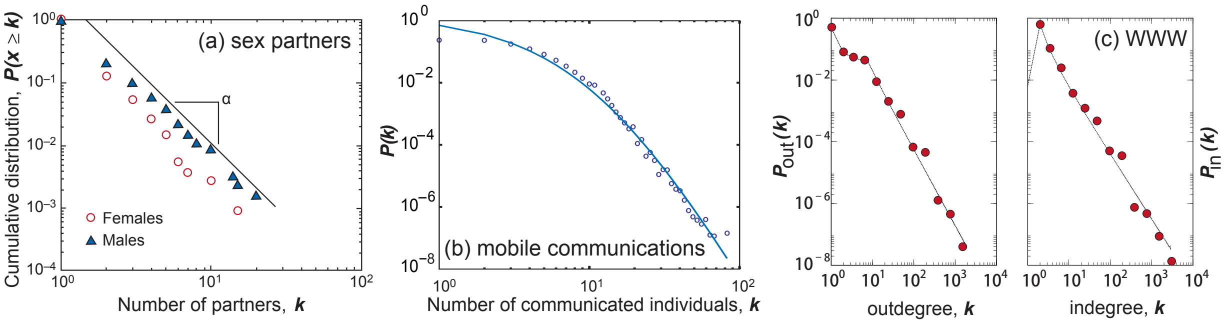


FIGURE 5 EXAMPLES OF SKEWED DEGREE DISTRIBUTION FROM REAL-WORLD NETWORKS. (A) NUMBER OF SEXUAL PARTNERS [112]; (B) NUMBER OF CONTACTS WITH MOBILE PHONE COMMUNICATION [113]; (C) OUTGOING AND INCOMING LINKS ON THE WORLD WIDE WEB, I.E., URLS FOUND ON HTML DOCUMENTS [76].

Mechanisms forming power-law degree distributions were investigated probably at the earliest by Herbert Simon in the 1950s. He showed that power-law arises when "the rich get richer", a phenomenon also referred to as the *Matthew effect* after the biblical edict[*]. Price explained the power-law of indegrees and outdegrees of scientific citation network with a later called "Price's model", in which nodes with varying outdegrees enter into the network consecutively and connect to pre-existing nodes with probability proportional to their indegree: $\Pr_{i \to j} \sim k_0 + k_j^{in}$, i.e., papers which have been cited many times would be more likely to be cited by new papers. Price called the mechanism in his model "*cumulative advantage*" [114,115].

A more well-known model for scale-free networks is the "BA model" developed by Barabási and Albert (see Figure 7) [116]: (1) *Growth*: Start with a small number (m_0) of nodes, each with at least one connection. At each step, add a new node in the network; (2) *Preferential attachment*: Connect the new node with $m \leq m_0$ links to m different nodes that are already in the network. The probability of choosing a node to connect is $\prod(k_i) = k_i / \sum_J k_j$, that is, nodes with higher degree will be more likely to be connected with new nodes. After t time steps this procedure generates a network with $N = t + m_0$ nodes and mt links. The average degree of a BA model is thus approximately

$$\bar{k} \simeq \frac{2mt}{t + m_0} \simeq 2m. \tag{17}$$

The degree distribution of a BA network follows [117]:

$$p(k) \sim 2m^2 k^{-3}. \tag{18}$$

BA networks also give short average path lengths: $\ell_{BA} \sim \ln N / \ln \ln N$, however, the clustering coefficients of BA networks are rather small, scaling with the network size and following approximately a power law: $C_{BA} \sim N^{0.75}$. Consequently, when the network grows large, the clustering coefficient approaches zero quickly.

Despite the limitation on capturing clustering for practical networks, the ability to resemble the growth and preferential attachment phenomenon, together with the ability of producing a power-law degree distribution, make the BA model one of the most studied and applied network models for complex networks studies in the last decade

[*] Matthew 25:29, King James Version: *For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath.*

[77,110]. The “scale-free” property of degree distribution has critical effects on many network dynamics. For example, the connectivity of these networks is extremely vulnerable to the removal of highly connected nodes [118-121], the critical threshold[*] for the spreading of diseases no longer exists [122,123], etc.

3.3.5 Social networks models

Social networks, like human sex networks, friendship networks and scientific collaboration networks, are found to be fundamentally different from non-social networks such as World Wide Web, power grids and airline networks. *Many studies on empirical networks have revealed that, in addition to common properties of real-world networks such as the **small-world effect** and **skewed degree distribution**, social networks often demonstrate positive degree correlation, i.e., **assortativity mixing**, and clear **community structures*** [124]. Large efforts have been made in recent years for designing models that can generate networks with these features [125-132].

Based on the process of how a network is generated, the current social network models can be classified into three categories [133]: network evolution models (NEMs), nodal attribute models (NAMs) and exponential random graph models (ERGMs), see Figure 6.

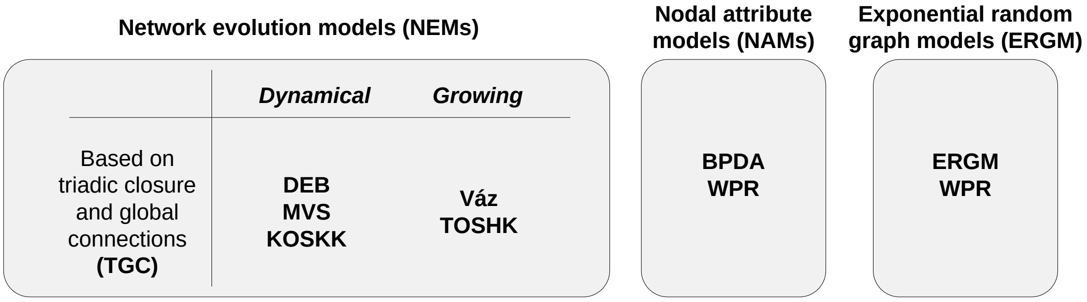
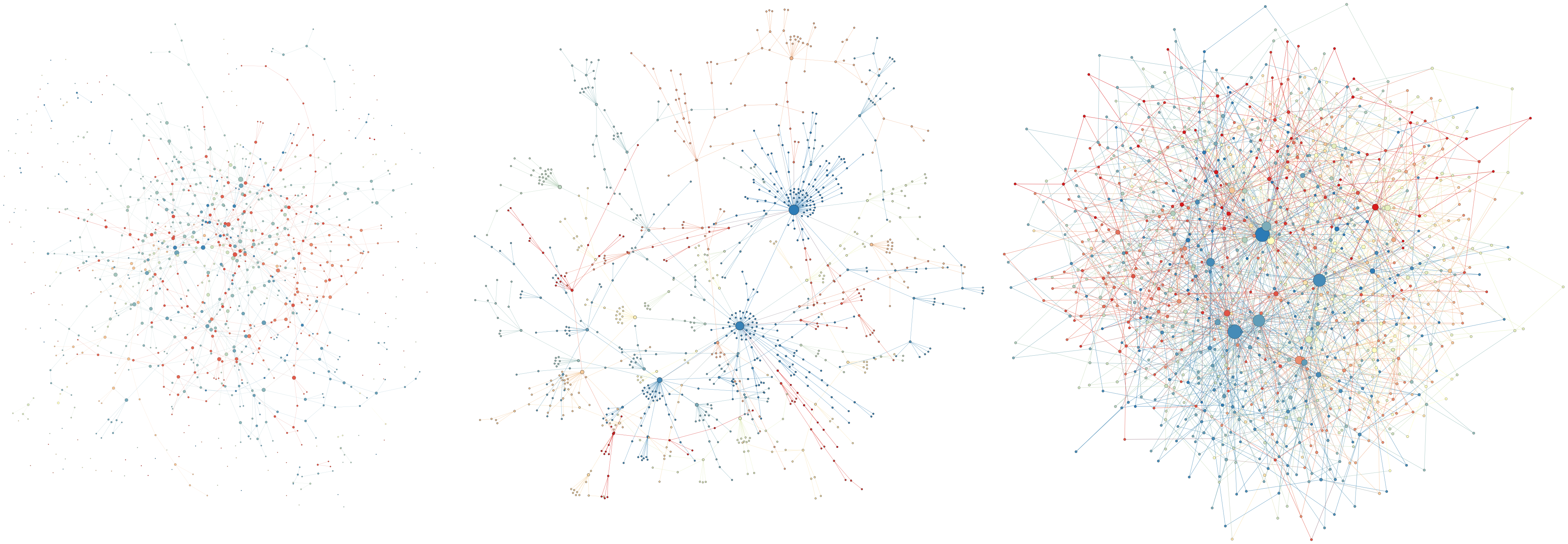

Network evolution models (NEMs)
Dynamical
Growing
Based on triadic closure and global connections (TGC)
DEB
MVS
KOSKK
Váz
TOSHK
Nodal attribute models (NAMs)
BPDA
WPR
Exponential random graph models (ERGM)
ERGM
WPR

FIGURE 6 CATEGORIES OF SOCIAL NETWORK MODELS. SOURCE: [133].

[*] A number (λ_c) calculated by the disease infectiousness and network structure parameter. The final state of an infectious disease spreading on networks with $\lambda < \lambda_c$ will die out.



(a) E-R random network, average degree =2.

(b) BA scale-free network, average degree =2.

(c) BA scale-free network, average degree =4.

Figure 7 ER random network and BA scale-free network. Network size is 1000, the size of each node is proportional to its degree.

4
RESPONDENT-DRIVEN SAMPLING

Respondent Driven Sampling (RDS), proposed by Douglas D. Heckathorn (1997), was first used in 1994 in the Eastern Connecticut Health Outreach (ECHO) project for the study of IDUs as part of an AIDS prevention intervention in US [134]. It was not used for HIV surveillance until 2003 outside the US [135,136], but since then there has been a rapid increase of RDS studies, with more than a hundred empirical studies in over 80 countries (Figure 10) targeting a wide range of hidden populations, such as injection drug users, men who have sex with men, sex workers and HIV infectors.

4.1 HOW DOES RDS WORK?

RDS begins with the selection of several initial respondents, which are called the "seeds". The seed is then given a number of "coupons" to distribute to friends and acquaintances. When interviewed, the new respondent is in turn given coupons to distribute. Everyone is rewarded both for completing the interview, and for recruiting their peers into the research. If recruitment chains are sufficiently long, the sample composition would stabilize and become independent of seeds. Additionally, the recruitment information about who recruit whom and each respondent's personal network size are recorded to be used for adjusting the sample composition. An illustration is presented in Figure 8.

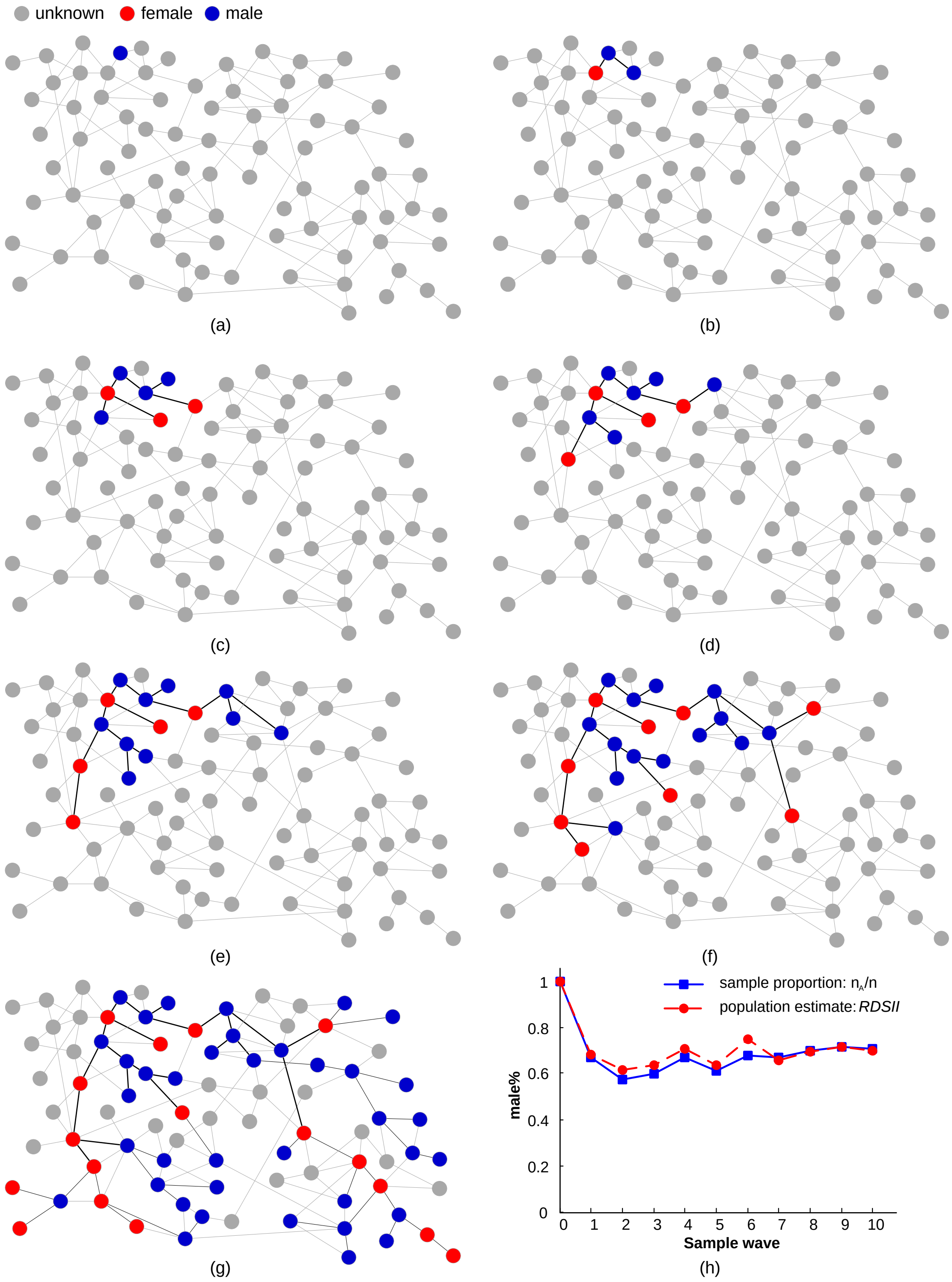
4.2 DIFFERENCE BETWEEN RDS AND SNOWBALL SAMPLING

Apparently, RDS is a form of "chain-referring" sampling strategy and are similar to the snowball sampling method introduced in 1.1.2. However, RDS differs from snowball sampling in several ways:

First, it uses a dual incentive mechanism to impulse the recruitment efficiency. Since each respondent is rewarded not only for the participation of him/her-self, but also the participation of peers he/she recommends, response rates are generally much higher than snowball sampling.

Second, rather than asking participants to name and reveal contact details of their friends, RDS let respondents recruit peers by themselves. Recruiting respondents by population members themselves instead of researchers who are from outside avoids the sensitivity and privacy concerns when hidden populations are approached. The peer-recruitment mechanism also reduces work load for researchers and allows the sample to grow automatically.

Third, the number of coupons is limited in RDS, i.e., each participant is allowed to recruit only a certain number of others. On the one hand, the restricted number of distributable coupons for each participant forces the sample recruitment chain penetrates into the inner most of the social network to reach the desired sample size, generating samples with improved representativeness; on the other hand, the limited nature of coupons makes recruiters consider them as valuable rights when recruiting peers, improving success rate of recruitment as they will try to recruit those they know and are more likely to participate to reward themselves. This difference is also important for developing models which can generate estimators for population characteristics, see 4.3.


● unknown ● female ● male
sample proportion: n_A/n
population estimate: *RDSII*
male%
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8 9 10
Sample wave

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 8 Illustration of an RDS process

Lastly, chain-referral sampling methods (including snowball sampling and RDS) generally have a critical source of bias due to the oversampling of population members with large social network sizes (*contact hubs*), if these hubs differ largely from other individuals on certain characteristics, the sample will be largely biased from the true population. To overcome this problem, in most RDS studies, respondents are asked to report their personal network sizes. This information is critical for the derivation of RDS estimators, which can then be used to generate asymptotically unbiased population estimates under several assumptions, see also 4.3.

4.3 THEORY OF RDS: MODELS AND ESTIMATORS

4.3.1 Modeling RDS as a Markov process

After collection, the properties of the nodes (respondents), information about who recruit whom (recruitment matrix), and the personal network sizes of respondents (degree) form the basis for generating inferences about the population characteristics.

Due to the non-random manner RDS samples are collected, an RDS sample is not sufficiently representative for the population as it suffers from various sources of biases, such as the underlying social network structure on which the recruitment takes place and the heterogeneity of personal network sizes. For example, if individuals in the population with a certain property (e.g., males) have more personal connections (i.e., degree) than those without this property (females), they would be more likely to be recruited by respondents, resulting uneven inclusion probabilities in the sample. Consequently, RDS will oversample those with more personal connections and can hardly be “representative” for the target population.

However, it is possible to build mathematical models to weight the sample to compensate for the fact that the sample is collected in a non-random way. The models are based on the following assumptions* [48,137-139]:

i. *Connectedness*: the network on which the recruitment takes place is connected, i.e., all individuals in the target population are connected, thus everyone can be accessed through her/his personal contacts.

ii. *Reciprocity*: all network links are undirected, i.e., the friendship/acquaintance relationships between individuals are reciprocal: if i can recruit j, j can recruit i, too.

iii. Sampling is *with replacement* (SWR): each individual can participate the study as long as he/she receives a valid coupon, no matter whether he/she has participated before.

iv. *Degree*: respondents can accurately report their personal network sizes.

v. *Random recruitment*: peer recruitment is a random selection from the respondent’s personal network, i.e., all friends in a recruiter’s personal network have the same probability of receiving a coupon.

vi. Only *one coupon* is used in the sampling procedure, i.e., each participant recruits a single peer.

* Note that these are assumptions required to build the mathematical model for developing RDS estimators. Most of these assumptions are merely theoretical and are not valid in real RDS deployments, I will discuss this issue in 4.5.

Given the above assumptions, if individual i is selected in sample wave t, the probability of each node to be selected in wave $t+1$ is

$$\Pr_{i \to j} = \begin{cases} 1/d_i & \text{if there is a link between } i \text{ and } j \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

and the RDS can be modeled as a Markov process with the following transition matrix:

$$T = \begin{bmatrix} 0 & e_{12}/d_1 & \cdots & e_{1N}/d_1 \\ e_{21}/d_2 & 0 & \cdots & e_{2N}/d_2 \\ \vdots & \vdots & \ddots & \vdots \\ e_{N1}/d_N & e_{N2}/d_N & \cdots & 0 \end{bmatrix}, \tag{20}$$

where $e_{ij} = 1$ if there is a link from individual i to individual j, and $e_{ij} = 0$ otherwise, and d_i is the degree of i. The equilibrium state distribution for this process is a vector $X = \{x_1, x_2, ..., x_N\}^T$ such that

$$X^T T = X^T. \tag{21}$$

Since the network is undirected, we have $e_{ij} = e_{ji}$. It can be verified that (21) has a unique solution

$$X = \{\frac{d_1}{\sum_{j=1}^{N} d_j}, \frac{d_2}{\sum_{j=1}^{N} d_j}, ..., \frac{d_N}{\sum_{j=1}^{N} d_j}\}^T \tag{22}$$

such that $\sum_{i=1}^{N} x_i = 1$.

(22) indicates that when an RDS sample reaches equilibrium, **the probability that each node to be included in the sample is proportional to its degree**:

$$\Pr_i = \frac{d_i}{\sum_{j=1}^{N} d_j}. \tag{23}$$

4.3.2 RDS estimator: RDSII

The conclusion of (23) is crucial, as it implies that, even collected in a non-random manner, we can treat the RDS sample as a probability sample such that the inclusion probability of each subject in the sample can be approximated by its degree, which can be used as the sampling weight to generate population estimates (see Figure 9).

Specifically, for a given sample $U = \{v_1, v_2, \ldots, v_n\}$, with n_A being the number of respondents in the sample with property A (e.g., HIV-positive) and $n_B = n - n_A$ being the rest. Let $\{d_1, d_2, \ldots, d_n\}$ be the respondents' degree. Then $\Pr_i$ can be used to obtain the *Hansen-Hurwitz estimator* [140-142] in which observations are weighted by the inverse of the sampling probability; the proportion of individuals belonging to group A (we consider a binary property such that each individual belongs to either group A or group B) in the population can be estimated by [139]:

$$\hat{P}_A = \frac{\sum_{i \in A \cap U} d_i^{-1}}{\sum_{j \in U} d_j^{-1}}. \tag{24}$$

(24) is called the RDSII estimator (or *VH* estimator), as there is another so-called "RDSI" estimator (or *SH* estimator) which appeared earlier in literature, see next section.

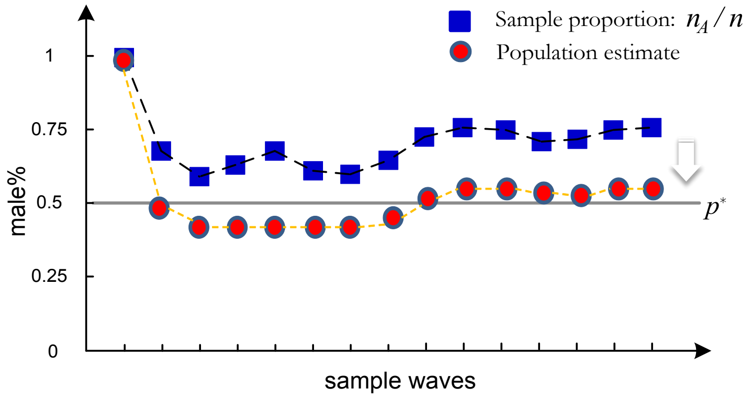


FIGURE 9 ILLUSTRATION ON THE FUNCTION OF RDS ESTIMATORS

4.3.3 RDS estimator: RDSI

4.3.3.1 The reciprocal model

The RDSI estimator has a more complicated form than RDSII, it was developed based on the reciprocal model [143]. When the network is undirected, the number of cross-group links from A to B should equal the number of links from B to A. Let

$$S^* = \begin{bmatrix} s_{AA}^* & s_{AB}^* \\ s_{BA}^* & s_{BB}^* \end{bmatrix} \tag{25}$$

be the recruitment matrix in the population, where s_{XY}^* is the proportion of links from group X to group Y ($X, Y \in \{A, B\}$), such that $s_{XX}^* + s_{XY}^* = 1$, then

$$N_A \bar{D}_A^* s_{AB}^* = N_B \bar{D}_B^* s_{BA}^*, \tag{26}$$

where $N_A = N - N_B$ is the number of individuals of group A in the population, and $\bar{D}_A^*, \bar{D}_B^*$ are average degrees for the two groups.

(26) can be rewritten as:

$$P_A^* \bar{D}_A^* s_{AB}^* = (1 - P_A^*) \bar{D}_B^* s_{BA}^*, \tag{27}$$

where P_A^* is the proportion of individuals in group A in the population.

To find a possible estimator for P_A^*, both $\bar{D}_A^*, \bar{D}_B^*$ and S^* need to be estimated from the sample data.

4.3.3.2 Estimate of average degree

Given the degree distribution of group A in the network, $p_A(d)$, the sample degree distribution, $q_A(d)$, is [144]

$$q_A(d) = \frac{d \cdot p_A(d)}{\sum_{d=1}^{\max(d)} d \cdot p_A(d)}, \tag{28}$$

where $p_A(d)$ is the population degree distribution and $\sum_{d=1}^{\max(d)} d \cdot p_A(d)$ is a normalizing constant to ensure that $q_A(d)$ sums to 1.

Note that $d \cdot p_A(d)$ is proportional to $q_A(d)$, it is also the case that $p_A(d)$ is proportional to $\frac{1}{d} \cdot q_A(d)$. So, if a sample has a degree distribution, $q_A(d)$, then the population degree distribution, $p_A(d)$, can be estimated as

$$\hat{p}_A(d) = \frac{\frac{1}{d} \cdot q_A(d)}{\sum_{d=1}^{\max(d)} \frac{1}{d} \cdot q_A(d)}. \tag{29}$$

Then the average degree of members in group A can be estimated as

$$\hat{\bar{D}}_A = \sum_{d=1}^{\max(d)} d \cdot \hat{p}_A(d). \tag{30}$$

This can also be written as

$$\hat{\bar{D}}_A = \frac{n_A}{\sum_{i=1}^{n_A} d_i^{-1}}. \tag{31}$$

Another way to estimate the average degree is to use a ratio of two *Hansen-Hurwitz* estimators [143]: the estimated number of links from group A, and the estimated number of individuals in group A:

$$\hat{\bar{D}}_A = \frac{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{\Pr_i} d_i}{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{\Pr_i}}. \tag{32}$$

Replace $\Pr_i$ with (23), we have

$$\hat{\bar{D}}_A = \frac{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{d_i} d_i}{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{d_i}} = \frac{n_A}{\sum_{i=1}^{n_A} d_i^{-1}}. \tag{33}$$

Similarly, the average degree of group B can be estimated by

$$\hat{\bar{D}}_B = \frac{n_B}{\sum_{i=1}^{n_B} d_i^{-1}}. \tag{34}$$

4.3.3.3 Estimate of recruitment matrix

When nodes in the network are selected proportional to their degrees, the selection probability of each link $e_{i \to j}$, can be written as

$$\Pr_{i \to j} = \Pr_i \cdot \frac{1}{d_i}. \tag{35}$$

The first term indicates the probability of selecting individual i, and the second term indicates each link from i has the same probability to be chosen to pass a coupon, i.e., assumption v.

Replace $\Pr_i$ with (23), we have

$$\Pr_{i \to j} = \frac{d_i}{\sum_{j=1}^{N} d_j} \cdot \frac{1}{d_i} = \frac{1}{\sum_{j=1}^{N} d_j}. \tag{36}$$

Note $\sum_{j=1}^{N} d_j$ is a constant for any network, (36) indicates that when the RDS sample reaches equilibrium, each link in the network has the same probability to be selected. Consequently, the recruitment links observed from the RDS sample, form a random sample of all links from the underlying social network. Let

$$S = \begin{bmatrix} s_{AA} & s_{AB} \\ s_{BA} & s_{BB} \end{bmatrix} \tag{37}$$

be the raw recruitment matrix observed from the sample, where s_{XY} is the proportion of all individuals recruited by members of group X who are members of group Y (X, $Y \in \{A, B\}$), such that $s_{XX} + s_{XY} = 1$, S then is an unbiased estimate for S^*.

4.3.3.4 RDSI estimator

With $\hat{\bar{D}}_A = n_A / \sum_{i=1}^{n_A} d_i^{-1}$, $\hat{\bar{D}}_B = n_B / \sum_{i=1}^{n_B} d_i^{-1}$ being the estimators for average degrees of group A, B, and S being the estimator for population recruitment matrix S^*, we can then solve (27) and obtain the RDSI estimator (or *SH* estimator):

$$\hat{P}_A = \frac{s_{BA}\hat{\bar{D}}_B}{s_{AB}\hat{\bar{D}}_A + s_{BA}\hat{\bar{D}}_B}. \tag{38}$$

4.3.3.5 Data smoothing

When there are more than two disjoint groups in the population, the reciprocal model will generate a set of overdetermined equations, i.e., the number of unknown parameters is less than the number of equations. For example, if there are three different groups in the population, the reciprocal model becomes:

$$\begin{aligned}
1 &= \hat{P}_1 + \hat{P}_2 + \hat{P}_3 \\
\hat{P}_1 \cdot \hat{\bar{D}}_1^* \cdot s_{12} &= \hat{P}_2 \cdot \hat{\bar{D}}_2^* \cdot s_{21} \\
\hat{P}_1 \cdot \hat{\bar{D}}_1^* \cdot s_{13} &= \hat{P}_3 \cdot \hat{\bar{D}}_3^* \cdot s_{31} \\
\hat{P}_2 \cdot \hat{\bar{D}}_2^* \cdot s_{23} &= \hat{P}_3 \cdot \hat{\bar{D}}_3^* \cdot s_{32},
\end{aligned} \tag{39}$$

where the population size parameter is canceled out and $\hat{P}_1, \hat{P}_2$, and $\hat{P}_3$ are estimated population proportions of the three groups.

Linear least squares may be applied to solve the system; alternately, Heckathorn proposed an approach called *data smoothing* [48,145]. The basic idea of data smoothing is that if links in the network are reciprocal, if all groups recruit with equal effectiveness (i.e., for any group X, the number of respondents recruited by X is equal to the number of recruitments of group X, $RB_X = R_{XX} + R_{XY} + \cdots + R_{XN} = R_{XX} + R_{YX} + \cdots + R_{NX} = RO_X$), and if recruitments from personal networks are random, then cross groups recruitments will be equal for each pair of groups, i.e., for any groups X and Y, $R_{XY} = R_{YX}$.

In the data smoothing process, first, each element R_{XY} is transformed to $s_{XY}\hat{E}_X RB$, where s_{XY} is the transition probability from the sample recruitment matrix, $\hat{E}_X$ is the Markov equilibrium given the transition matrix S, and RB is the total number of recruitments in the sample. The purpose of such a transformation is to make the transformed recruitment matrix keep the original selection proportions between groups and equal the row and column sums. The next step is then to use the mean of these counts, to yield a smoothed recruitment matrix R^{**} as follows:

$$R^{**} = \begin{bmatrix}
s_{11}\hat{E}_1 RB & \dfrac{s_{12}\hat{E}_1 RB + s_{21}\hat{E}_2 RB}{2} & \cdots & \dfrac{s_{1M}\hat{E}_1 RB + s_{M1}\hat{E}_M RB}{2} \\
\dfrac{s_{12}\hat{E}_1 RB + \hat{s}_{21}\hat{E}_2 RB}{2} & \hat{s}_{22}\hat{E}_2 RB & \cdots & \dfrac{s_{2M}\hat{E}_2 RB + s_{M2}\hat{E}_M RB}{2} \\
\vdots & \vdots & \ddots & \vdots \\
\dfrac{s_{1M}\hat{E}_1 RB + s_{M1}\hat{E}_M RB}{2} & \dfrac{s_{2M}\hat{E}_2 RB + s_{M2}\hat{E}_M RB}{2} & \cdots & s_{MM}\hat{E}_M RB
\end{bmatrix}$$

$$= \begin{bmatrix}
R_{11}^{**} & R_{12}^{**} & \cdots & R_{1M}^{**} \\
R_{21}^{**} & R_{22}^{**} & \cdots & R_{2M}^{**} \\
\vdots & \vdots & \ddots & \vdots \\
R_{M1}^{**} & R_{M2}^{**} & \cdots & R_{MM}^{**}
\end{bmatrix}. \tag{40}$$

Based on R^{**}, the selection probabilities are recalculated in (39), and the excess equations that cause the problem of over determination become redundant. For example, based on the smoothed selection proportions and the estimated degrees, the smoothed population estimate is calculated as follows in a system with M groups:

$$
\begin{aligned}
&1 = \hat{P}_1 + \hat{P}_2 + \hat{P}_3 + \cdots \hat{P}_M \\
&\hat{P}_1 \cdot \hat{\hat{D}}_1^* \cdot \hat{s}_{12} = \hat{P}_2 \cdot \hat{\hat{D}}_2^* \cdot \hat{s}_{21} \\
&\hat{P}_1 \cdot \hat{\hat{D}}_1^* \cdot \hat{s}_{13} = \hat{P}_3 \cdot \hat{\hat{D}}_3^* \cdot \hat{s}_{31} \\
&\cdots \\
&\hat{P}_1 \cdot \hat{\hat{D}}_1^* \cdot \hat{s}_{1M} = \hat{P}_M \cdot \hat{\hat{D}}_M^* \cdot \hat{s}_{M1},
\end{aligned}
\tag{41}
$$

where $\hat{s}_{XY} = R^{**}_{XY} / \sum_K R^{**}_{XK}$. Note that the data smoothing method doesn't alter the average degree.

When there are only two groups in the population, it can be verified that the data smoothing will affect neither the estimation of recruitment matrix S nor the RDSI estimator.

4.3.4 Connection between RDSI and RDSII

Both RDSI and RDSII estimator are asymptotically unbiased [48,137-139]. From (41), we can see that for any group X ,

$$
\hat{P}_X = \frac{\hat{P}_A \hat{\hat{D}}_A \hat{s}_{AX}}{\hat{\hat{D}}_X \hat{s}_{XA}} = \frac{\hat{P}_A \hat{\hat{D}}_A \dfrac{R^{**}_{AX}}{\sum_K R^{**}_{AK}}}{\hat{\hat{D}}_X \dfrac{R^{**}_{XA}}{\sum_K R^{**}_{XK}}},
\tag{42}
$$

from which it follows that

$$
\begin{aligned}
\sum_X \hat{P}_X &= \sum_X \frac{\hat{P}_A \hat{\hat{D}}_A \dfrac{R^{**}_{AX}}{\sum_K R^{**}_{AK}}}{\hat{\hat{D}}_X \dfrac{R^{**}_{XA}}{\sum_K R^{**}_{XK}}} = \frac{\hat{P}_A \hat{\hat{D}}_A}{\sum_K R^{**}_{AK}} \sum_X \frac{R^{**}_{AX} \sum_K R^{**}_{XK}}{\hat{\hat{D}}_X R^{**}_{XA}} \\
&= \frac{\hat{P}_A \hat{\hat{D}}_A}{\sum_K R^{**}_{AK}} \sum_X \frac{\sum_K R^{**}_{XK}}{\hat{\hat{D}}_X} = 1.
\end{aligned}
\tag{43}
$$

When seeds are excluded, the number of type X participants recruited into the study is the same as the number of type X participants in the sample, i.e, $\sum_K R^{**}_{XK} = n_X$. Solving (43) for $\hat{P}_A$ we have

$$
\begin{aligned}
\hat{P}_A &= \frac{n_A}{\hat{\hat{D}}_A} \left(\sum_X \frac{n_X}{\hat{\hat{D}}_X} \right)^{-1} = \frac{n_A}{\hat{\hat{D}}_A} \left(\sum_X \frac{n_X}{n_X \big/ \sum_{i \in U \cap X} d_i^{-1}} \right) \\
&= \frac{n_A}{\hat{\hat{D}}_A} \left(\sum_X \sum_{i \in U \cap X} d_i^{-1} \right) = \frac{n_A}{\hat{\hat{D}}_A} \cdot \frac{n}{\sum_{i \in U} d_i^{-1}}.
\end{aligned}
\tag{44}
$$

The right-hand side can be re-written as

$$
\hat{P}_A = \frac{n_A}{\hat{\hat{D}}_A} \cdot \frac{n}{\sum_{i \in U} d_i^{-1}} = \frac{n_A}{n_A / \sum_{i \in A \cap U} d_i^{-1}} \cdot \frac{n}{\sum_{i \in U} d_i^{-1}} = \frac{\sum\limits_{i \in A \cap U} d_i^{-1}}{\sum\limits_{i \in U} d_i^{-1}},
\tag{45}
$$

yielding the exact form of (24), i.e., the RDSII estimator. Therefore, as long as data smoothing is used, RDSI and RDSII will coincide.

4.3.5 Variance estimation

4.3.5.1 Bootstrap method

The precision of a sample estimate is usually enhanced by providing a confidence interval (CI), which gives a range within which the true population is expected to be found with some level of certainty. Due to the complex sample design of RDS, simple random sampling based CIs are generally narrower than expected [48,146,147]. Consequently, bootstrap methods are used to construct CIs around RDS estimates.

Salganik (2006) proposed a later widely used bootstrap procedure for RDS estimates to generate CIs. The procedure is as follows [147,148]:

(i) Divide the sample respondents into two groups based on the property of their recruiters, that is, those who are recruited by type A nodes (A_{rec}) and those who are recruited by type B nodes (B_{rec});

(ii) Randomly select a respondent from the sample, if the respondent has property A, then the next respondent is randomly picked from A_{rec}, otherwise the next respondent is randomly picked from B_{rec}. Continue to draw a new respondent until the original sample size is reached.

(iii) Calculate RDS estimate based on the replicated sample.

(iv) Repeat step (ii) and (iii) until R bootstrapped estimates are calculated.

(v) The middle 90%/95% estimates from the ordered R bootstrapped estimates are then used as the estimated CI.

4.3.5.2 MCMC-based variance estimation

To account for the non-uniform selection probabilities and the MCMC structure of the RDS sample, Volz and Heckathorn developed an estimator for the variance of $\hat{P}_A$ [139]:

$$\hat{V}_{P_A} = \frac{1}{n(n-1)} \sum_U (Z_i - \hat{P}_A)^2 + \frac{\hat{P}_A^{\ 2}}{n} \left((1-n) + \frac{2}{n_A} \sum_{i=2}^{n} \sum_{n=1}^{i-1} (S^{|i-j|})_{AA} \right) \tag{46}$$

where $Z_i = nd_i^{-1} / \sum_{j \in U} d_j^{-1}$ if $i \in A$ and $Z_i = 0$ otherwise, and $|i - j|$ indicates the distance of sampling waves between respondent i and j. Details of the derivations can be found in [139].

4.4 RDS AROUND THE WORLD

There are two significant improvements in RDS compared to other non-random methods when sampling hidden population. First, it uses dual incentives to impulse the respondents to recruit more persons into the research, improving response rate. Second, unbiased estimates can be obtained by RDS estimators, enabling researchers to draw conclusions for the entire studied population from the RDS sample.

The efficiency and effectivity of RDS have been proven by the wide practices of RDS studies around the world. It has become the state-of-the-art sampling method for studying hard-to-access populations [136,149]. For example, the US Centers for Disease Control and Prevention (CDC), whose decisions often influence global public health standards, have selected RDS for a 25-city study of injection drug users that is part of the National HIV Behavioral Surveillance System [150,151]. It has also been used by Family Health International, the largest non-profit agency in international public health, in more than a dozen countries, including Bangladesh, Burma, Cambodia, Egypt, Honduras, India, Kosovo, Mexico, Nepal, Vietnam, Pakistan, Papua New Guinea and Russia to study MSM, IDUs and SWs [152].

In a review of RDS studies used for HIV biological and/or behavioral surveillance, Malekinejad et al (2008) and Johnston et al (2008) identified that, from 2003 through October 1, 2007, there were 128 RDS studies conducted in 28 countries outside US, with over 32,000 IDUs, MSM, SWs and high-risk heterosexual (HRH) men being surveyed.

In addition to HIV/AIDS-related high-risk populations, RDS has been applied to study a variety of other populations, such as jazz musicians [153,154], visual artist [155], regular nightlife users [156], young people [157-159], homeless people [160], university students [161], migrant worker [162-164], refugees [165], immigrants [166,167].

I have made a recent literature search. By January 03, 2013, there have been more than 80 countries that had at least one RDS implemented worldwide, see Figure 10.

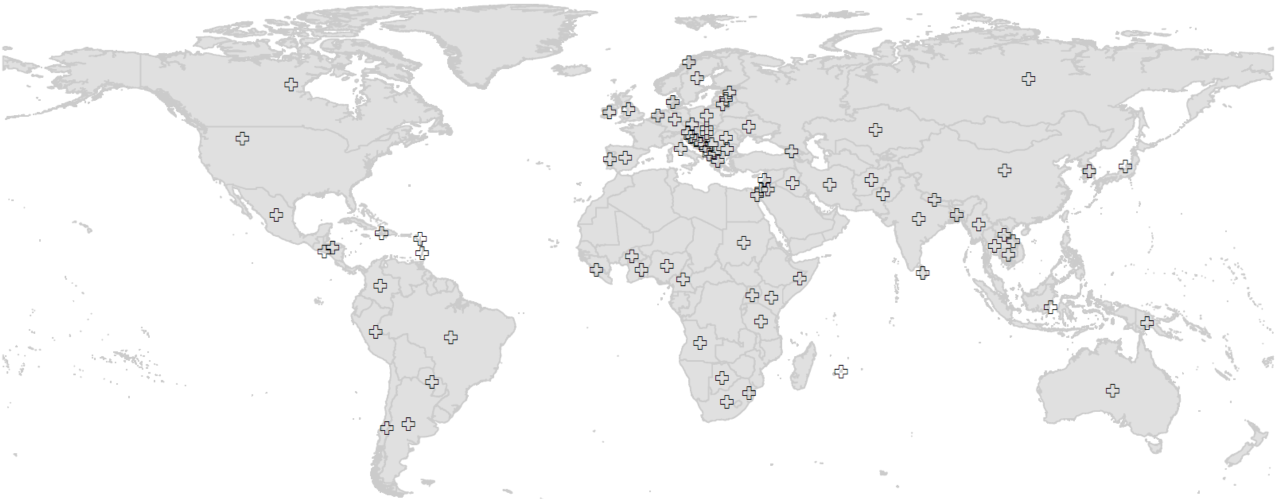


FIGURE 10 WORLD MAP OF COUNTRIES WITH AT LEAST ONE RDS STUDY IMPLEMENTED

4.5 LIMITATIONS

It has been shown that the RDS estimators are asymptotically unbiased when all the assumptions are fulfilled [139]. However, almost all of these assumptions are not met in real life [168]:

First, RDS assumes all relationships are reciprocal, however, most social networks contain directed links, or links that do not have the same strength in both directions [169-171]. For example, if i and j can participate the study and receive coupons to distribute, i considers j as the first candidate to pass a coupon, but j may not recruit i since he has other more favorite friends.

Second, RDS assumes SWR, however, to prevent participants from colluding to recruit each other back and forth to gain rewards, and to maximize cost-efficiency, real life RDS studies sample without replacement (SWOR), meaning that respondents can participate only once.

Third, RDS estimators need to use the degree data from the sample, however, it is difficult for respondents to report their personal network sizes accurately [172].

Fourth, participants usually pass their coupons to peers with whom they have a close rather than a more distant relationship, which is not a random selection [168,173,174].

Fifth, to avoid recruitment chains stopping too early, researchers often use more than one coupon in the RDS study [136,149].

Apparently, given violation of, part of or all of, these assumptions, the validity and reliability of RDS estimators become questionable. In parallel with our work of study I, in which the effect of violation of RDS assumptions were thoroughly evaluated, Gile and Handcock (2010) found a potential bias caused by preferential selection of peers and SWOR and addressed the possibility of a reduction of bias by discarding early waves [175]. However, the numbers of seeds, coupons and waves were fixed and many other assumptions that might affect the RDS estimates, such as directness of networks, recruitment failures and degree reporting error, were not simulated. This study was also subjected to the limit that the simulated population was only 1000 and the tested sample sizes range between 500~950, occupying 50%~95% of the entire population.

As traditional RDS evaluations were mostly based on synthetic networks and ideally fulfilled assumptions, the precision of RDS estimates have long been overestimated. As a consequence, the sample size of RDS was usually determined based on the presumption that RDS has the same variance as simple random sampling. In a later study where Salganik (2006) developed the bootstrapping method for variance estimates, he recommended to use a sample size as twice as for SRS. However, Salganik’s recommendation was based also on simulated RDS on synthetic networks with ideally fulfilled assumptions.

It was not until recently for researchers to find that the variance in RDS might have been severely underestimated. By simulating RDS on empirical networks (one high-risk heterosexual network focusing on sex workers and drug injectors and their sexual and drug partners in Colorado Spring; and 84 middle and high school friendship networks from the National Longitudinal Study of Adolescent Health in US), Goel and Salganik (2010) found that the variance of RDS estimates were as high as 5.7~58.3 times higher than SRS when the sample size is 500, indicating a serious overestimation on the precision of RDS. They also found that the bootstrapping method tend to produce misleadingly narrow CIs, masking the effects of inadequate sample sizes.

All the above evaluations were made by simulated RDS process on networks with known characteristics, McCreesh et al (2012), on the contrary, conducted an RDS study in an empirical setting, where the RDS method was used to recruit household heads in rural Uganda where the true population data was known [176,177]. They found that only one-third of RDS estimates outperformed the raw proportions in the sample, and only 50%-74% of RDS 95% CIs (based on the bootstrapping method) included the true population proportion. The narrower than expected CIs produced by the bootstrapping method was also found by Wejnert et al (2008, 2009), who had tested the RDS method by recruiting college students in 2004 and 2008 [161,168,178]. They had also tested the performance of the second variance estimator, the MCMC-based method, and they found that it tended to overestimate variance.

4.6 RECENT DEVELOPMENT OF RDS THEORIES

The recent discovery on the limitations of RDS has led to an intensive research effort on the development of new RDS estimators.

4.6.1 Heckathorn-estimator

Aiming to analyze continuous variables and control for differential recruitment, Heckathorn proposed a variant of RDSI estimator (H-estimator) in 2007. The H-estimator was developed by partitioning the sample into contiguous degree groups and model the RDS as a Markov chain on these degree groups. It has a similar form of RDSI [145]:

$$\hat{P}_A = \frac{s_{BA}\widehat{AD}_B}{s_{AB}\widehat{AD}_A + s_{BA}\widehat{AD}_B}, \tag{47}$$

where $\widehat{AD}_A$ is the adjusted average degree estimate for members of group A and can be calculated by

$$\widehat{AD}_A = \frac{\sum_{i\in A\cap U}\hat{E}_{g(i)} / p_{g(i)}}{\sum_{i\in A\cap U}(\hat{E}_{g(i)} / p_{g(i)})d_i^{-1}}, \tag{48}$$

with $p_g = n_g / n$ being the sample proportion of respondents in degree group g, and $\hat{E}_g$ being the proportion of respondents in degree group g when the Markov chain on degree groups reaches equilibrium. Consequently, if the RDS sample starts from the equilibrium state, which is very unlikely to happen, the H-estimator estimator becomes the same as RDSI since $\widehat{AD}_A = n_A / \sum_{i\in A\cap U} d_i^{-1}$. Heckathorn also recommended to divide the sample into $c = \sqrt{n / n_c}$ groups of approximately equal size. In the standard software for RDS data analysis, RDSAT, the default value for n_c is 12 [179].

It has been found that the H-estimator is almost identical to RDSI under various simulation settings [180], including the presence of differential recruitment*, non-response and non-recruitment. The difference exists only under very unlikely scenarios, e.g., when all seeds have extremely low (or high) degree, and there is a big difference between the average degrees of different groups.

* Note that even the original motivation of the H-estimator was to overcome the problem of differential recruitment, in the evaluation of Tomas (2011), no evidence was found that the H-estimator adjusts for differential recruitment or non-response.

4.6.2 SS-estimator

Other studies seek to use some priori information to improve the performance of RDS estimates. For example, **based on known population size**, Gile (2011) developed a successive-sampling-based estimator (*SS-estimator*) to adjust the SWOR feature of empirical RDS [181]. The estimator has a similar form of RDSII, instead of the degree, it approximates the inclusion probability of each node by a series of simulated successive sampling* samples from an estimated population degree distribution. The basic procedure of calculating SS-estimator is as follows:

i. Initialize a unit size to inclusion probability mapping function $f^0(k): k \to \pi$,

$$f^0(k) = \frac{k}{N}\sum_l \frac{v_l}{l}, \tag{49}$$

where v_l is the number of respondents with degree l in the sample. The initialization ensures that $f^0(k)$ is proportional to k.

ii. Iteratively estimate population distribution of degrees. For $i = 1, ..., r$:

   a. Estimate the number of individuals with degree k in the population:

   $$N_k^i = N \cdot \frac{v_k}{f^{i-1}(k)} \Big/ f^0(k) = \frac{k}{N}\sum_l \frac{v_l}{f^{i-1}(l)}. \tag{50}$$

   This procedure uses the population size N as a known parameter and are very similar to the degree estimation introduced in equation (29).

   b. Estimate the inclusion probabilities for nodes from the population of $\{N_k^i\}$. This is achieved by simulating M SS-samples of size n from $\{N_k^i\}$, and the inclusion probability for a node with degree k can be estimated by

   $$f^i(k) \approx \frac{U_k + 1}{MN_k^i + 1}, \tag{51}$$

   where U_k is the total number of observed units with degree k from the M SS-samples.

iii. After r iterations, $f^r(k)$ is then used as an approximation of inclusion probability for nodes of degree k, i.e., $\Pr(k) \propto f^r(k)$. Substituting d_i with $f^r(d_i)$ in the RDSII-estimator, the population estimate then becomes:

$$\hat{P}_A = \frac{\sum_{i \in A \cap U} f^r(d_i)^{-1}}{\sum_{j \in U} f^r(d_j)^{-1}}. \tag{52}$$

It is recommended to use $M = 2000$ and $r = 3$ [181]. The SS-estimator has shown superior performance with simulations on networks of 1000 nodes with large fraction of sample sizes (over 50% of the network size), when there is a big difference between SWR and SWOR. However, in an evaluation where more complex simulation settings were used, e.g., when RDS was implemented with differential recruitment and non-response rates, SS-estimator failed to outperform other estimators under many situations [180].

* Successive sampling (SS) is also called probability proportional to size without replacement sampling (PPSWOR). In SS, each unit is selected into the sample with probability proportional to unit size from among the remaining units.

Note that the *SS-estimator* is dependent on the knowledge of the true population size, which is usually not known for hidden populations. A compromise would be to use it as a sensitivity test method to check the variation of estimate given a range of population sizes.

4.6.3 GH-estimator

In Gile and Hancock (2011), the SS-estimator was extended to adjust for the bias induced by the selection of seeds. Instead of drawing SS samples from a population degree distribution, simulated RDS samples (WOR) with replicated features (e.g., sample size, number of seeds, off spring distributions[*]) of the observed sample were drawn from networks generated by ERGM models. The new estimator (*GH-estimator*) **requires knowledge about both the population size and the property of neighbors among each participant's personal networks** [182]:

i. Initialize $f^0(i)=\frac{d_i}{N}\sum_{j=1}^{N}\frac{S_j}{d_j}$, where $S_i=1$ if person i is sampled, otherwise $S_i=0.$

ii. For $i=1,...,r$:

   a. Estimate the number of individuals with degree k and property $X=\{0,1\}$ in the population:

   $$N^i_{k,X}=\frac{1}{N}\sum_{j=1}^{N}\frac{S_j}{f^{i-1}(j)}\mathrm{I}(d_j=k,z_j=X), \tag{53}$$

   where $z_i=1$ if $i\in A$, otherwise $z_i=0$.

   b. Estimate $\tilde{g}(G,\Theta)^i=\sum_{j=1}^{N}\frac{S_j(\theta_j(1-z_j)+(d_j-\theta_j)z_j)}{2f^{i-1}(d_j)}$ for the ERGM model

   $$P_{\eta,g(G,\Theta)}(Y=G)=\frac{\exp(\eta g(G,\Theta))}{C} \tag{54}$$

   where θ_j is the number of nodes with property A among j's personal network, η is the model parameters and C the normalizing function. The model is then fitted to compute η, denote by $\hat{\eta}^r$, based on $\{N^i_{k,X}\}$ and $\tilde{g}(G,\Theta)^i$.

   c. Simulate M_1 networks according to the distribution given by $\hat{\eta}^i$, $\{N^i_k\}$, and $\tilde{g}(G,\Theta)^i$. For each of these network, simulate M_2 RDS samples according to the sampling parameter n, N^{seeds}, p^s_c. The inclusion probability for any node of degree d_j and type X can be estimated by:

   $$f^i(j)=\frac{U^i_{d_j,z_j}+1}{MN^i_{d_j,z_j}+1},\ j\in U. \tag{55}$$

   where $U^i_{d_j,z_j}$ is the total number of observed units with degree d_j and property z_j from the $M_1\times M_2$ samples.

iii. Estimate the population proportion by:

[*] The offspring distribution $\{p^s_c, c=1,..,\text{maximum number of coupons}\}$ is merely the distribution of proportions of number of succeed recruitments for respondents in the sample.

$$\hat{P}_A = \frac{\sum_{i \in A \cap U} f^r(i)^{-1}}{\sum_{j \in U} f^r(j)^{-1}}. \tag{56}$$

Gile and Hancock (2011) have also developed a bootstrap approach for constructing CIs. In this method, RDS processes are repeatedly simulated on the ERGM networks generated in (54), CIs are then calculated based on estimates (the middle 95%/90% ordered estimates by (56)) from simulated RDS samples. They have shown that the new estimator is able to generate estimates with minimum bias and it is robust to selection bias of seeds.

4.6.4 Other approaches

A few researchers focus on developing innovative methods for analyzing the RDS sample data. For example, Poon et al (2009) modeled the tree-like structure of RDS as a multitype branching process (MBP) based on stochastic context-free grammars (SCFGs). The new method allowed them to find latent variability in the recruitment process of an RDS study for IDU in Tijuana, Mexico, that IDUs tended to emulate the recruitment behavior of their recruiter, and the recruitment of a peer of their own type was dependent on the number of recruits [183].

In a recent unpublished work, Handcock et al (2012) developed a Bayesian inference approach for estimating the population size based on RDS sample data. With adequate prior information on the population degree distribution and population size distribution, it has been shown in their case studies that the new approach is able to generate estimates compatible with UNAIDS guideline estimates as well as capture/recapture estimates of population sizes [184].

4.7 SUMMARY

From more than a decade RDS has proven its ability in efficiently accessing hidden populations. The power of generating unbiased population estimates, however, has been less applauded as more and more researchers recognize that violation of assumptions in empirical studies is common, and that even on simulated networks with ideal recruitment, RDS tends to generate estimate with large variance.

Evaluation and improvement are consequently critical for the continuing popularization of RDS in the study of hidden populations. The current literature, however, is limited in the following aspects:

(i) *Lack of a systematical overview on the effect of violation of RDS assumptions.* RDS estimators are based on six assumptions, with almost all of them are violated in practices; however, most evaluation studies focus only on part of these assumptions, such as SWOR, seed selection bias and recruitment behavior. There is a lack of knowledge on the performance of RDS estimators regarding network directedness, degree reporting error, etc.

(ii) *Excessively large sample fraction in population of limited size.* As I mentioned before, many RDS evaluation studies extensively used the tested network size of 1000, with sample sizes ranges from 500 to 950. The purpose of such a setting would undoubtedly help to identify the problem of SWOR, as opposed to SWR in the RDS assumption; however, sampling 50%~95% of population individuals would prohibit the generalization of results to more typical scenarios where the sample fraction is much smaller.

(iii) *Lack of a test network resembling a hidden population*. The performance of RDS estimators are assessed on either synthetic networks or networks from non-hidden population, and often the effect of network structure such as degree distribution and communities which are considered of central importance is ignored in these studies.

(iv) *New estimators are not applicable* when the network is directed, or when the prior information about the population size is difficult to obtain.

As an attempt to overcome these limitations, this thesis focuses on the **evaluation** and **improvement** of RDS estimate methods, see Chapter 5.

5

OBJECTIVES AND FRAMEWORK

5.1 OBJECTIVES

This thesis aims a comprehensive study on the *evaluation* and *improvement* of respondent-driven sampling, empirically and theoretically. We try to find answers for the following questions:

i. What is the effect of violation of *any* assumptions on the performance of RDS estimators?

ii. Is it possible to implement RDS through Internet for hidden population?

iii. What is the benefit and challenge of implementing Web-based RDS?

iv. How to develop more robust RDS estimators such that the estimate bias is not subject to as many violations of assumptions as RDSI/RDSII?

5.2 FRAMEWORK

The four studies included in this thesis focus on the two key words in the objectives, as illustrated in Figure 11.

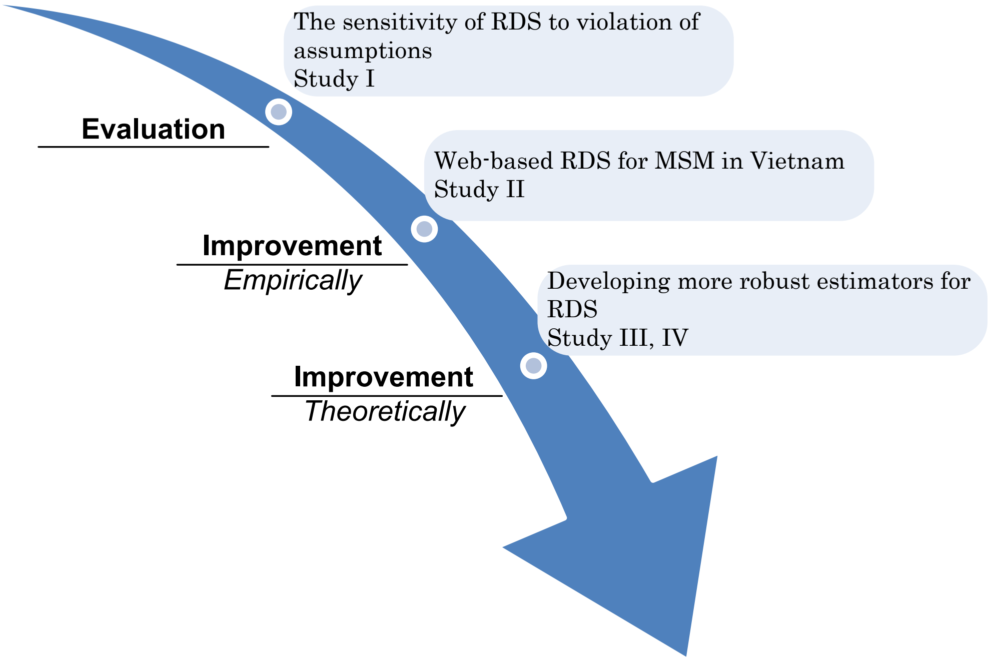

The sensitivity of RDS to violation of assumptions
Study I
Evaluation
Web-based RDS for MSM in Vietnam
Study II
Improvement
Empirically
Developing more robust estimators for RDS
Study III, IV
Improvement
Theoretically

FIGURE 11 ILLUSTRATION OF THE THESIS FRAMEWORK.

Evaluation: In study I, we exam the potential bias of RDS estimators by simulating RDS with violation of assumptions, one by one, based on an empirical social network of online LGBT (lesbian, gay, bisexual, and transgender) community with known population characteristics. The results of such a thorough evaluation thereby provide

RDS practitioners a useful manual on accessing the severity of violations in sample data as well as cautiousness needed to interpret RDS estimates.

Improvement: With widely existing evidence of violations of RDS assumptions in practice, and sever bias such violations may effect traditional RDS estimators revealed in study I, we identified a pressing need for advanced methodologies to be applied to improve RDS estimators to be more robust to violation of assumptions. We improve RDS from two aspects:

Empirically: Location-based face-to-face interviews usually barrier potential respondents from long distance traveling to reach the study site; Internet-based surveys, on the contrary, provide easy access to participation as well as covering for sensitive conversations. In Study II, we implemented a Web-based RDS study for the study of MSM in Vietnam. The study aimed to evaluate the feasibility and challenge of implementing a Web-based RDS towards hidden population.

Theoretically: Study III developed a method which generalizes RDS method from undirected network to directed network; Study IV proposed an RDS estimator which does not require population value as prior information and has superior performance over traditional RDS estimators. The new estimator also exhibits strong robustness to violation of the random recruitment assumption.

6
MATERIALS AND METHODS

6.1 EMPIRICAL NETWORK DATA (PAPER I, III, IV)

6.1.1 Data collection

An anonymized online social MSM network was used to evaluate the performance of existing and newly developed RDS estimators. The network came from the Nordic region's largest and most active Web community for homosexual, bisexual, transgender and queer persons (www.qruiser.com) [185]. Contacts between members on the Web site were maintained mainly by a "favorites list", on which each member could add any other member *without approval* from that member. Members could attend clubs (Web pages with specific topics) and sent messages to each other.

We collected information on personal profiles registered on this Website as well as all messages that were sent within the Web community from November 15th, 2005, to January 18th, 2006*. During the 65 days of the data collection period, 12,590,911 messages were recorded and 184,819 distinct members were registered on the Web site.

6.1.2 Network formation

On the basis of the membership profiles, we extracted a network in which each node represents a member registered as homosexual male, and each link represents the relationship that a member added another member on his favorite list. Note that approval is not needed from whom was added, the link is directed. If a pair of members added each other, the link is reciprocal; if there is only one directed link between them, the link is irreciprocal.

To make sure each node could be recruited with simulated RDS, only members of the giant connected component (GCC) from the network with only reciprocal links were kept as nodes in all following variants of networks (16082 active, gay men).

Undirected network (G1): when only reciprocal links are kept, the 16082 gay men and the links between them forms the fundamental undirected MSM network for our test. It was examined in Study I and Study IV.

Directed network (G2): if we add previously excluded irreciprocal links to the undirected network, we obtain a directed network, with larger link density but the same number of nodes. It was examined in Study I and Study III.

Weighted network (G_{max} and G_{min}): we weighted each reciprocal link in the undirected network, by either the maximum number or minimum number of messages sent in any one direction, to test the effect of nonrandom recruitment, see Study I.

Variants of the undirected network ($G1_{add}$ and $G1_{rand}$): to avoid misleading conclusions resulting from the effects of network structure and link density, variations of the undirected network were created in Study I by randomly adding links or rewiring links

* There is a typo in Paper I, in which the date was written as "from December 15th, 2005, to January 18th, 2006".

according to certain criteria. For each population property examined, the link addition and rewiring process were specially designed such that the homophily remained unchanged in the obtained denser or rewired network.

Variants of the directed network: in Study III, the directed network was used as the basis for generating networks with different levels of indegree correlation, for each studied population property.

A simple illustration on the relationship of these networks is presented in Figure 12. Details of the generation processes for the above networks can be found in Paper I, III and IV.

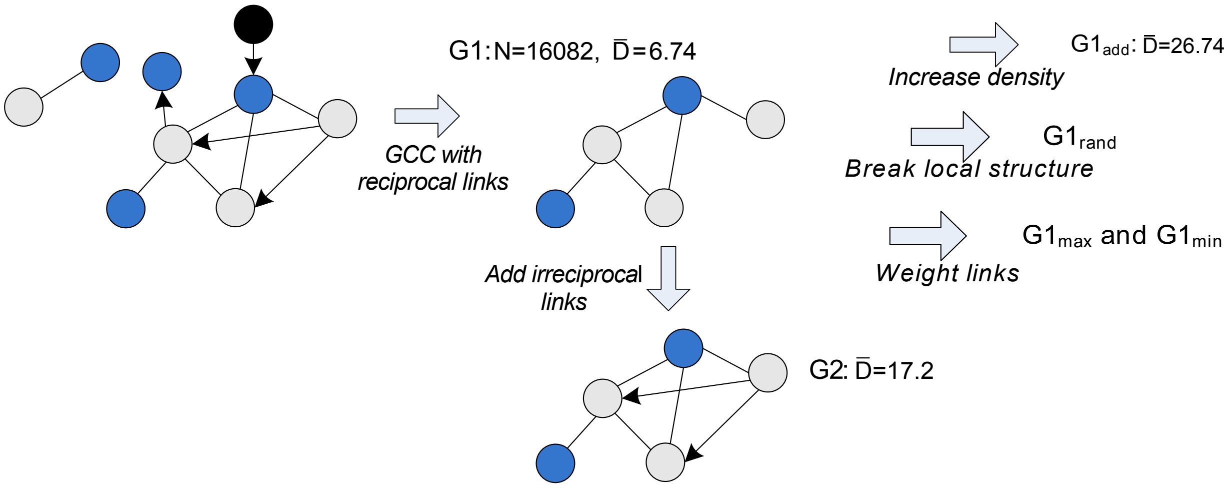


FIGURE 12 ILLUSTRATION OF THE NETWORK GENERATION PROCESS

6.1.3 Network properties

The average degree was 6.74 for the undirected network. With irreciprocal links added, it increased to 17.2 for the directed network. Both the undirected and directed network had very skewed degree distributions, for example, half of the nodes in the directed network had no more than 10 outgoing links, while a small proportion of members had a large number of outgoing links, see Figure 13.

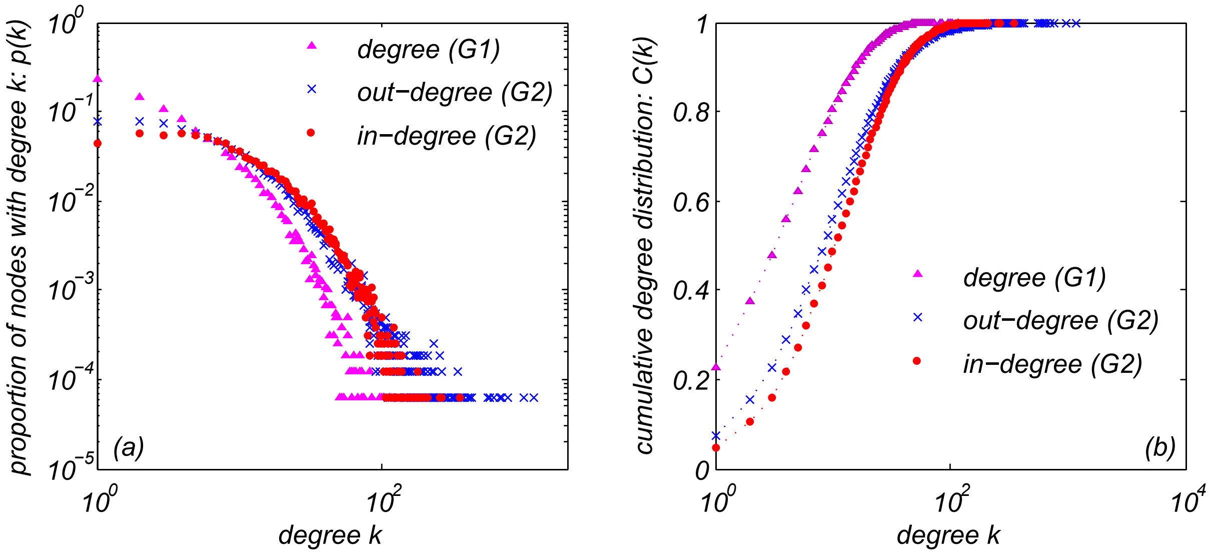


FIGURE 13 (A) DEGREE DISTRIBUTION AND (B) CUMULATIVE DEGREE DISTRIBUTION OF THE MSM NETWORK. SOURCE: [137]

The performance of RDS estimators was evaluated by comparing estimates from simulated RDS samples with the true population value, for four selected dichotomous properties extracted from users' profiles: *age* (born before 1980), county (live in Stockholm, *ct*), civil status (married, *cs*), and profession (employed, *pf*).

These four properties covered a wide range of population proportion, cross-group link probability, homophily and activity ratio[*], forming a rich test base for the evaluation of RDS estimators (see Table 3). Take homophily for an example, the homophily for the county was 0.50, which means that members who live in Stockholm formed links with members who also live in Stockholm 50% of the time, while they formed links randomly with members from among all cities (including Stockholm) the remaining 50% of the time. The civil status had a very low level of homophily, indicating that links were formed as if randomly among other members, regardless of their marital status.

TABLE 3 POPULATION PROPORTIONS p^*, HOMOPHILIES H AND ACTIVITY RATIO w OF THE STUDIED VARIABLES IN THE MSM NETWORKS

	Age		**County**		**Civil status**		**Profession**	
	Before 1980	*others*	*Stockholm*	*others*	*Single*	*Others*	*Employed*	*Others*
p^*	77.77	22.23	38.79	61.21	40.39	59.61	38.19	61.81
H G1	0.4	0.37	0.5	0.4	0.05	0.08	0.13	−0.05
H G2	0.23	0.34	0.5	0.28	0.03	0.07	0.06	0.02
w G1	1.05	0.95	1.22	0.82	0.97	1.03	1.21	0.83
w G2	1.22	-0.95	0.82	-1.05	1.15	-1.32	0.87	-0.76

6.2 WEBRDS (PAPER II)

6.2.1 The WebRDS system

An automated WebRDS surveying system was developed to recruit MSM Internet users in Vietnam. With this system, a participant logs into the Website with a password (coupon). After completing the survey, the system automatically generates four additional passwords for the participant to distribute. Participants can choose either forwarding these coupons by themselves or through the system by providing their email or Yahoo! Messenger addresses (popular for Internet communication in Vietnam).

6.2.2 Inclusion criteria and incentives

This survey was cross-sectional, performed online with the WebRDS system and carried out between February 18 and April 12, 2011. Eligible participants were adult men (≥18 years) 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.

In order to simulate the recruitment, we offered each participant 1) 50,000 VND (2.45 USD) as credit on the participant's SIM card and the same amount for each successful recruitment of an MSM friend; 2) the option of donating the monetary reward to a MSM community organization chosen by the participant; 3) a lottery with the

[*] Activity ratio, is the ratio of mean degree for group A to group B, $w = \bar{D}_A / \bar{D}_B$.

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).

6.2.3 Questionnaire

The questionnaire contained 17 questions (see Appendix), 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)*
- *duration of the respondent's longest relationship*
- *opinion on legalizing same-sex marriage in Vietnam*
- *frequency of Internet-use*
- *sociodemographic characteristics*
- *network size*
- *relationship between the participant and his recruiter*
- *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.

6.2.4 Sampling procedure

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 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

[*] A line of tablet computers designed and marketed by Apple Inc. http://www.apple.com/ipad/

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.

6.2.5 Piloting and early version 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).

6.3 TOOLS FOR DATA PROCESSING AND SIMULATION

Database software, Microsoft SQL server and MySQL, were used to store and process RDS sample data and empirical network data. The official analytical tool for RDS sample data is RDSAT with the latest version 7.0. However due to the flexibilities required by our analysis, I used self-coded programs in Microsoft Visual Studio C#.net and Matlab for data processing and simulation. Network visualization was made with Gephi, Pajek, Netdraw and Adobe Illustrator.

Useful links

Microsoft SQL:

http://www.microsoft.com/sqlserver/en/us/default.aspx

MySQL:

http://www.mysql.com/

RDSAT:

http://www.respondentdrivensampling.org/main.htm

Microsoft Visual Studio:

http://www.microsoft.com/visualstudio/

Matlab:

www.mathworks.com/products/matlab/

Gephi:

https://gephi.org/

Pajek:

http://pajek.imfm.si/doku.php?id=pajek

Netdraw:

http://www.analytictech.com/downloadnd.htm

Adobe Illustrator:

http://www.adobe.com/products/illustrator.html

6.4 ETHICS

As discussed in Chapter 2, ethical considerations are especially important when studying HIV/AIDS-related high-risk populations.

The empirical network data studied in paper I, III and IV was de-identified and extracted by the Website Administrator with the approval from the Regional Ethical Review Board in Stockholm (EPN). We are particularly concerned with data privacy and confidentiality, and possible risks of diluting data anonymity [186,187]. For example, a user like “XY3769” maybe meaningless to other people but people who are active online are as well known by their usernames as their traditional names. Revealing of MSM identity in Nordic countries may be a less sensitive issue comparing to other worlds like Asia, however, in addition to the risk of bringing stigmatization and harms to personal reputation to individual users, the Website owner/company, also need to risk losing customers and violating data policies once harm has been done to their users based on this data. For these reasons, all usernames have been replaced with identifiers that provide no link to the actual participant when the data is fully de-identified. We did not store any information that can used to reveal user identities from the Website, such as email, IP address, or message content.

The use of the empirical network data has several important outcomes: first, it helps us to understand the social network structure of hidden populations; second, it provides a rich test base for the evaluation of RDS method, the outcome of which is critical to guide the implementation and data analysis of RDS in other countries; third, improvements of RDS methods developed based on this data would be critical for future RDS applications and will help researchers and policy makers gain a better knowledge about hidden population. As time goes by, the increased clarity of hidden population societies will strengthen our understanding and decrease the stigmatization around them.

Unlike Nordica countries, in Vietnam as elsewhere in Asia, identities of MSM are heavily stigmatized though they are not illegal. Most men get married to follow the culture and norms even they perceive themselves as homosexual. Consequently, study for MSM in Vietnam is highly sensitive and challenging. MSM are unwilling to reveal their identities to friends and families and are often afraid of being discriminated from the public. The WebRDS recruiting system avoids the sensitivity and privacy concerns raised during physically-based face-to-face interviews. However, we do aware of that there is a possibility of identifying an individual even from Internet. In paper II, we put a lot of effort into making the site and recruitment system safe and confidential for participants, including: 1) only individual with an authorized coupon could log into the system; 2) once logged in, the participant was given the consent page to choose to either participate or leave. All information about the survey was available on all survey pages; 3) participants could choose to leave the system at any time; 4) when a participant clicked the log-out button, the browser was automatically directed to Google.com and a detailed instruction on how to delete browser history was given; 5) communication between the users and the server was encrypted, the original IP address was coded using the one-way encryption algorithm MD5 and was deleted after the encryption; 6) all visiting information was emptied and the user needed to log in again if he did not have any activity on the survey for 5 minutes. This study was approved by the Hanoi Medical University Review Board for Bio-Medical Research and EPN. All data was analyzed in fully de-identified form.

The successful implementation of WebRDS system in recruiting more than 600 respondents reveals that, with minimized sensitivity and privacy concern, WebRDS is a useful tool for sampling MSM Internet users in Vietnam. This study is a first attempt to studying the characteristics of demography and risk behaviors of online MSM population with a representative sample, and would undoubtedly contribute to the understanding of hidden populations and to the setting up of HIV surveillance and prevention programs in Vietnam.

7
RESULTS

7.1 PAPER I: The Sensitivity of Respondent-Driven Sampling

7.1.1 Summary

In Paper I we use the empirical MSM social network and its variants as the test base, to run simulated RDS processes with violation of assumptions, one by one, and compare RDS estimates with true population values, to assess the sensitivity of RDS methods to different violation of assumptions.

7.1.2 Study design

We ran simulated RDS processes on the test networks in various settings. After each simulation, the RDS estimates for the four properties, age, county, civil status, and profession, were then compared with the population values. Average estimates (AE), bias, standard deviation (SD), mean absolute error (MAE) and design effect (DE)[*], were used to assess the performance of RDS estimators[†].

Six scenarios were used to simulate RDS in ideal or real-life settings:

(i) *Ideal scenario*: We ran RDS on the undirected MSM social network (G1) with all assumptions specified in 4.3.1 fulfilled.

(ii) *Violation of the reciprocal assumption*: We ran RDS on the directed MSM social network (G2).

(iii) *Violation of the SWR assumption*: We ran RDS with SWOR, i.e., each individual can only participate once. G1 and its variants, i.e., the link-added denser networks ($G1_{add}$) and link-rewired random networks ($G1_{rand}$) were tested.

(iv) *Violation of the degree assumption*: We allowed participants to reject invitations and let participants ignore (miscount) peers when inviting. We simulated the rejection and ignoring behavior both independently and dependently of the characteristics of participants. G1, $G1_{add}$ and $G1_{rand}$ were tested.

(v) *Violation of the random recruitment assumpt*ion: We allowed respondents to be more likely to recruit friends with whom they communicate more often. The weighted networks were tested ($G1_{max}$ and $G1_{min}$).

(vi) *Violation of the one coupon assumption*: We simulated RDS with different selection method of seeds and with varied number of seeds and coupons. G1, $G1_{add}$ and $G1_{rand}$ were tested.

Note that in the above settings, some of them are actually combinations of violation of assumptions, such as when participants were allowed to reject invitation and ignore peers, this could be seen as, first, a violation of degree assumption which requires participants to report degree accurately, and second, a violation of the one coupon

[*] The variance of the RDS estimates divided by the variance of SRS with the same sample size.
[†] We chose RDSII estimator as it is equivalent to RDSI when the population is composed of two disjoint groups in the population and it has shown improved analytical power in literature.

assumption which states that all participants use their only coupon to make one successful recruitment.

7.1.3 Result

RDS under ideal assumptions

When all assumptions are fulfilled, the RDSII estimates converge to the true population proportions very quickly. When sample sizes are between 500 and 1000, the SD is around 0.05, and the MAE is around 0.04. The design effects are around 13 and 10, for age and county respectively, and 5 for both civil status and profession (see Figure 14).

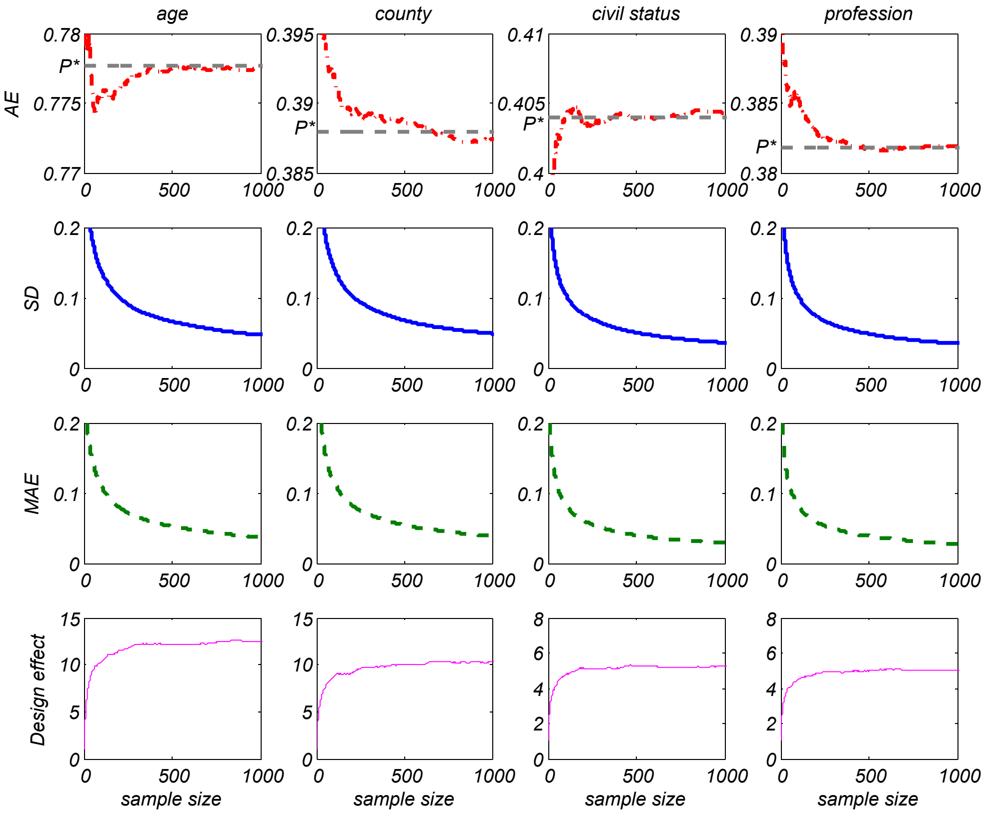


Figure 14 RDSII estimations on the undirected network G1. (a) age; (b) county; (c) civil status; (d) profession. Source: [137]

Violation 1: RDS on networks with irreciprocal links

Estimates are biased for all variables. Biases for age and county can be as high as 0.06, whereas for variables with less homophily (civil status and profession), biases are lower, at 0.005 and 0.022 respectively. The SDs are similar for all four groups (and very similar to the SD of the undirected networks). However, the MAE is much higher than that of the undirected networks for age and county (0.07–0.08).

Violation 2: sampling without replacement[*]

SWOR generates bias in different directions when the RDS sample occupies a large fraction of population. However, when sample size is less than 1000, the bias of SWOR is negligible and sometimes even less than the bias of SWR, and that the SD, MAE and DE are always smaller than those for SWR. These results indicate that in practical RDS

[*] Correction: In Paper I, fig 6 and fig 7, the number of seeds should be 10, instead of 1.

implementations where SWOR is used, as long as the sample fraction is small, this violation of assumption will actually be beneficial to the performance of RDS estimates.

Violation 3: RDS with rejection rates and miscounted personal networks

When the probabilities of rejection and miscounting do not depend on the outcome variables, i.e., all nodes exhibit the same recruitment behavior, regardless of their characteristics, the bias is small-to-moderate on G1 and $G1_{rand}$, and negligible on the dense network, $G1_{add}$. Both SD and MAE decrease with increased probability of rejection and miscounting.

By contrast, when the rejecting and miscounting behavior is dependent on the characteristics of individuals, large bias (and MAE) may be generated. The absolute worst-case scenario occurs when individuals of the two disjoint groups behave in opposite ways. For example, when members who were born before 1980 reject half of the invitations that were given to them and the members who were born after 1980 do not reject any invitations (no miscounting of personal networks), the bias is over 0.3 for age.

Violation 4: RDS with non-random recruitments

RDS estimates are biased for all four variables when the probability of distributing coupons to peers proportional to the contact frequency (amount of messages sent) between each pair of nodes. The bias is, however, not subjected to the homophily of variables: biases for age, county, civil status, and profession on G_{max} are 0.01, 0.02, 0.04, and 0.03 respectively. The non-random recruitment also result in higher SD and MAE than ideal conditions.

Violation 5: RDS with non-randomly selected seeds and increased number of coupons

We simulated RDS by choosing nodes as seeds either uniformly or proportional to their degree; however, the differences in biases between the two methods are minute. The SD and MAE generated by these two methods are in essence the same when the sample size is 500.

The number of seeds and coupons, on the other hand, has a clear effect on SD and MAE of RDS estimates: both the SD and the MAE increased when the samplings used more coupons, especially combined with limited number of seeds.

7.2 **PAPER II:** Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam

7.2.1 Summary

Paper II is an attempt to improve the RDS implementation by providing a system for respondent to participate and recruit with Internet-based surveys (WebRDS). The use of the Internet enables respondents to participate in the study easily and avoids the sensitivity issues that arise during face-to-face interviews by answering Web surveys anonymously.

7.2.2 Sampling dynamics

676 submissions were recorded. 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 (Figure 15).

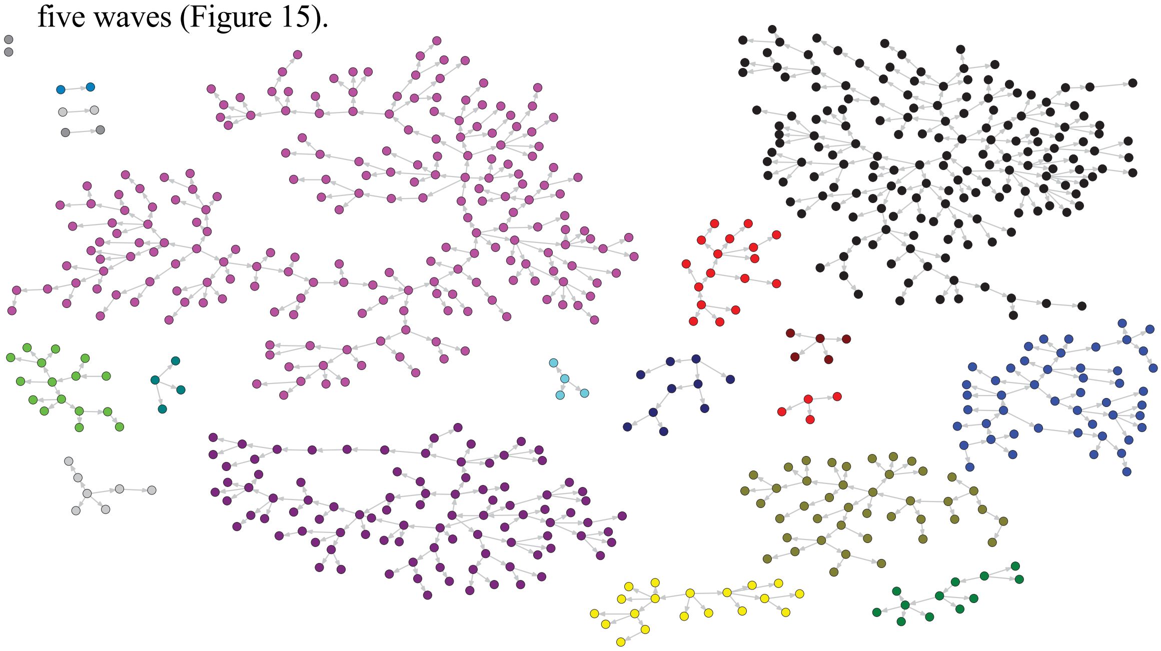


FIGURE 15 RECRUITMENT CHAINS OF SUBMITTED SURVEYS. SOURCE: [188]

Five seeds were added 14 days after the first group. If we backdate the start date of these five seeds by 14 days so that all seeds could be considered to have started on the same day, the site received around 500 submissions from the activation of seeds to two weeks later. 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.

7.2.3 Duplicated submissions, data cleaning and analysis

9.6 percent 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 together with the aforementioned invalid submissions to produce a cleaned sample (571 respondents). From this sample we estimated population proportions using RDSII. 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 include or exclude these flagged submissions. Specifically we compared the estimates generated from the full sample of non-seed submissions with valid age with the 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 (<3 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%.

7.2.4 Equilibrium

Using the standard criteria in the literature [189], 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 (see supplementary material of Paper II). 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 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).

7.2.5 Sample characteristics

The personal network size used for RDSII adjustment is defined as the number of persons the participant believed used the Internet and had interacted with in anyway during the past seven days (including on the phone, Internet, or in person). The average network size was 5.5 persons. Adjusted by the reported personal network sizes, 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 outside the large metropolitan areas with 32 provinces represented out of 63.

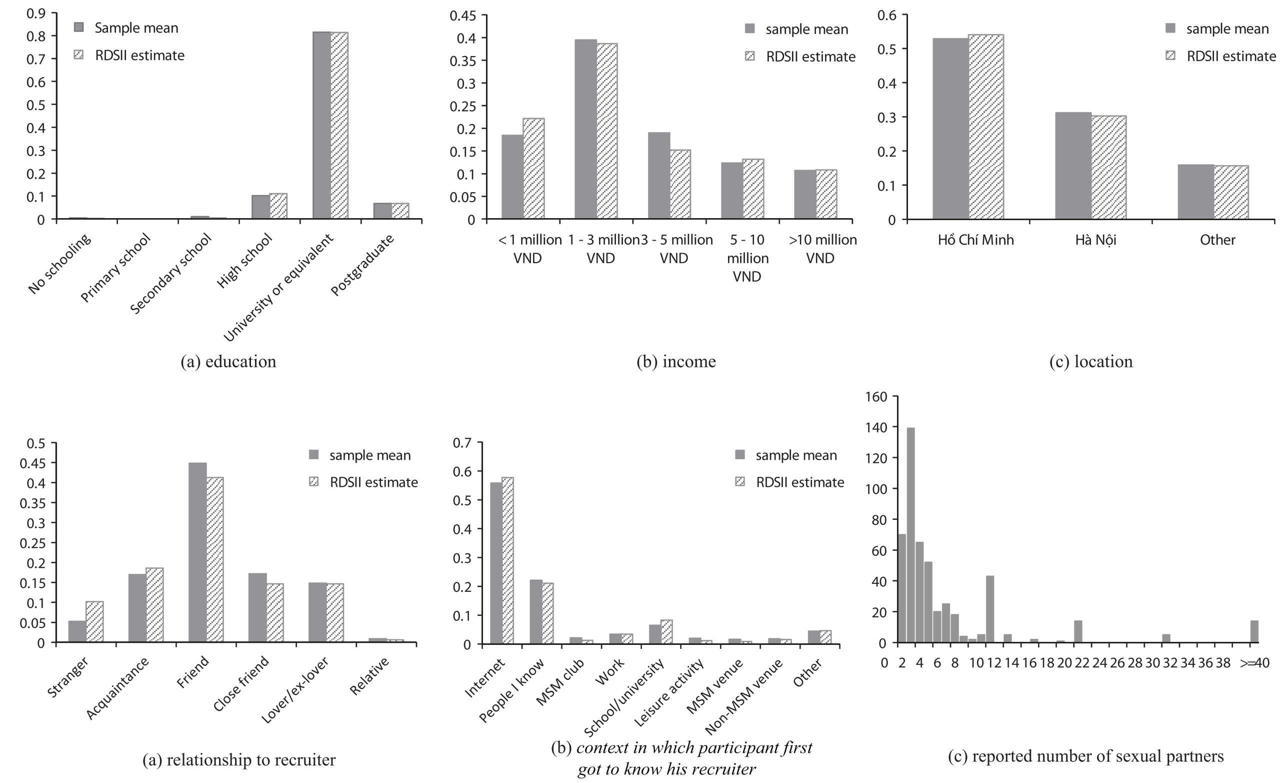
An estimated 98% (99% of the sample) preferred only men or preferred men to women as sexual partners, and 81% (81% of the sample) thought that same-sex marriage should be allowed in Vietnam. An estimated 92% (91% of the sample) had an existing relationship to their recruiter (an estimated 8% were recruited by a stranger). Median number of sexual partners during the last six months was two. Figure 16 presents the sample proportions and estimates of selected variables.

Comparison with existing statistics

Comparing national statistics and other published research data with our estimates shows 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.

Age: 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 [190]. 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 [191] and an RDS in Hanoi with median age of 20–24 years. 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 [191].



(a) education

(b) income

(c) location

(a) relationship to recruiter

(b) *context in which participant first got to know his recruiter*

(c) reported number of sexual partners

FIGURE 16 SAMPLE PROPORTIONS AND ESTIMATED POPULATION PROPORTIONS FOR SELECTED VARIABLES. SOURCE: [188]

Income: Income distribution is broadly consistent with the national average monthly per capita income for urban areas (2,130,000 VND, 2010 [192]). It is also comparable to data from the online survey among visitors to Vietnamese MSM Websites [191] and the offline RDS in Hanoi 2008 [191], although inflation, economic growth and differential categorization of income levels precludes an exact comparison.

Education: 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 [191] and 79% in the survey among visitors to Vietnamese MSM Websites [191].

Location: The sample was heavily concentrated on 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 [193,194]. This is similar to the online banner survey on Vietnamese MSM Websites where 74% came from Hanoi and HCMC [191]. 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 [105], with short numbers of steps between provinces.

Sexual partner preference: One percent stated that they preferred only women or preferred women to men as sexual partners. The banner survey on MSM sites [191] and an offline RDS in Hanoi with a similar question [191] 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.

7.3 **PAPER III:** Respondent-Driven Sampling on Directed Networks

7.3.1 Summary

In Paper I, we showed that one of the most harmful violations of assumptions is that the underlying network over which the coupons are distributed contains irreciprocal relationships, i.e., the network is directed. Unfortunately, this violation of assumption occurred quite often in RDS practices. Paper III aims to improve the RDS methodology by developing new estimators that allows RDS samples collected from directed networks to be generalized to the population.

7.3.2 Study design

Extension of RDSII (VH$_{out}$) estimator to directed network

When the network is directed but strongly connected, i.e., all nodes can be reached from any initial node, given that all other assumptions are fulfilled, the RDS process can be modeled as a Markov process with a transition matrix $R = \{a_{ij} = e_{ij} / d_i^{out}, 1 \le i, j \le N\}$ where d_i^{out} is the outdegree of node i. This process has a unique equilibrium distribution $\pi = [\pi_1 \cdots \pi_N]$ satisfying $R^T \pi^T = \pi^T$, indicating that π is the eigenvector corresponding to eigenvalue 1 for R^T. Consequently, π_i can be used to obtain the Hansen-Hurwitz estimator where observations are weighted by the inverse of the sampling probability:

$$\hat{p}_A^{Eig} = \frac{\sum\limits_{i \in U \cap A} \pi_i^{-1}}{\sum\limits_{j \in U} \pi_j^{-1}}. \qquad (57)$$

Unfortunately, no analytical solution for π is available for a general directed network. However, note that under the above assumptions, the RDS process is merely a random walk on the network, for which we can easily adopt the mean field approach in [195] to derive an approximation of π (see analytical details in Paper 3).

When there is no degree-degree correlation in the network, we have proven that the inclusion probability for any node i is approximately proportional to its indegree d_i^{in}, i.e., the RDS sample can be weighted by respondents' indegrees to estimate population proportions:

$$p_A^{VH_{in}} = \frac{\sum\limits_{i \in U \cap A} (d_i^{in})^{-1}}{\sum\limits_{j \in U} (d_j^{in})^{-1}}. \qquad (58)$$

Extension of RDSI (SH_{out}) estimator to directed network

In a directed network, the sum of nodes' indegrees in a group equals the total number of links pointing to nodes in that group:

$$\begin{cases} N_A \bar{D}_A^{out} S_{AA}^* + N_B \bar{D}_B^{out} S_{BA}^* = N_A \bar{D}_A^{in} \\ N_A \bar{D}_A^{out} S_{AB}^* + N_B \bar{D}_B^{out} S_{BB}^* = N_B \bar{D}_B^{in} \end{cases}, \tag{59}$$

where e.g., $\bar{D}_A^{out}$ is the average outdegree in group A and S_{AB}^* is the proportion of links originating in group A which end in group B in the network.

Solving (59) yields a generalization of the SH_{out} estimator:

$$\hat{p}_A^{SH_{in}} = \frac{\hat{\phi}}{1+\hat{\phi}}, \tag{60}$$

where $\phi = N_A / N_B$ is the relative group size proportion and can be calculated by

$$\phi = \frac{w^* S_{AA}^* - m^* S_{BB}^*}{2m^* w^* S_{AB}^*} + \sqrt{\frac{S_{BA}^*}{m^* w^* S_{AB}^*} + (\frac{m^* S_{BB}^* - w^* S_{AA}^*}{2m^* w^* S_{AB}^*})^2}, \tag{61}$$

in which $m^* = \bar{D}_A^{in} \big/ \bar{D}_B^{in}$ and $w^* = \bar{D}_A^{out} \big/ \bar{D}_B^{out}$ are the average indegree and outdegree ratio of the two groups of nodes in the network. Consequently, given the estimates for m^*, w^* and S^*, we can estimate population characteristics with (60).

In Paper III, m^*, w^* and S^* are estimated by

$$\hat{m}^* = \frac{n_A / \sum_{i \in U \cap A} (d_i^{in})^{-1}}{n_B / \sum_{i \in U \cap B} (d_i^{in})^{-1}}, \tag{62}$$

$$\hat{w}^* = \frac{n_A / \sum_{i \in U \cap A} (d_i^{out})^{-1}}{n_B / \sum_{i \in U \cap B} (d_i^{out})^{-1}}, \tag{63}$$

$$\hat{S}^* = \begin{bmatrix} s_{AA} & s_{AB} \\ s_{BA} & s_{BB} \end{bmatrix}. \tag{64}$$

where e.g., s_{AB} is the observed proportion of all individuals recruited by members of group A who are members of group B.

The factor w^* was named the *activity ratio* in literature [175], since it quantifies how active nodes in different groups are in building their personal networks. Following this, we henceforth refer to m^* as the *attractivity ratio*, as it reflects how "attractive" nodes in different groups are, or to which group of nodes links are inclined to connect to.

Use VH_{in} and SH_{in} as a sensitivity test method

With the notation of attractivity ratio m^*, we can rewrite the VH_{in} estimator as

$$\hat{p}_A^{VH_{in}} = \frac{\sum_{i \in U \cap A} (d_i^{in})^{-1}}{\sum_{j \in U} (d_j^{in})^{-1}} = \frac{n_A / \hat{\bar{D}}_A^{in}}{n_A / \hat{\bar{D}}_A^{in} + n_B / \hat{\bar{D}}_B^{in}} = \frac{n_A / n_B}{n_A / n_B + \hat{\bar{D}}_A^{in} / \hat{\bar{D}}_B^{in}} = \frac{n_A / n_B}{n_A / n_B + \hat{m}^*}. \tag{65}$$

As indegree is not collected in RDS studies, $\hat{m}^*$ is an unknown parameter for both VH_{in} and SH_{in}. However, with proper prior information, we can, instead of providing a point estimate with fixed parameters, use a range of m values to generate an estimate interval for p_A^*. That is, if m^* is assumed to lie within a certain range, $[m_{min}, m_{max}]$, we get an interval of $\hat{p}_A$, $[\hat{p}_A(m_{min}), \hat{p}_A(m_{max})]$, by varying m in (60) and (65). We emphasize that this interval is not a confidence interval, but a range of point estimates of p_A reflecting the dependence on the plausible values of m^*. When tested m values are used, we denote *VH_{in}* and *SH_{in}* as *VH_m* and *SH_m*, separately.

Simulation design

Network data: In the evaluation, we consider the following parameters which are important both to directed networks and RDS estimation: *Directedness* (λ, the proportion of irreciprocal links in the network); *indegree correlation* (γ, quantified by the indegree-based assortativity as defined in [100]); *indegree-outdegree correlation* (ρ, the Pearson correlation of indegree and outdegree); *homophily* (h_A) the activity ratio w^*, as well as the attractivity ratio m^*, are also used as network structure parameters in our assessment. For further explanations of these parameters see Paper 3.

The above network structural parameters are incorporated in the generated networks to assess the proposed new estimators (see Table 4):

- Indegree-outdegree uncorrelated networks: Net1.
- Indegree-outdegree correlated networks: Net2.
- Empirical MSM network.
- Indegree correlated networks: Net3.

TABLE 4 BASIC STATISTICS OF NET1, NET2, NET3 AND THE MSM NETWORK

	Network size (N)	Average degree ($\bar{D}$)	Directed-ness (λ)	indegree corre-lation (γ)	indegree-outdegree correlation (ρ)		Homophily (h)	Attractivity ratio (m^*)	P
Net1	$10,000$	10	$[0,1]$	$[-0.09, 0.01]^*$	≈ 0		$[-0.30, 0.22]^*$	$[0.7, 1.4]$	70%
Net2	$10,000$	10	$[0,1]$	$[-0.03, 0.14]^*$	$\approx 1-\lambda$		$[0, 0.5]$	$[0.7, 1.4]$	30%
MSM Network	$16,082$	17.2	0.61	0.03	0.39	*age*	0.23	0.95	77%
						ct	0.50	1.32	39%
						cs	0.03	0.96	40%
						pf	0.06	1.05	38%
Net3	$--^\dagger$	$--$	$[0.61, 0.91]^*$	$[0, 0.4]$	$--$		$--$	$--$	$--$

* parameter not controlled during the generation process;
† same as the MSM network.

Estimators: For each simulation, we estimate the population proportion with our suggested estimators as well as existing estimators. Then, the root mean square error (RMSE), standard deviation (SD) and bias of estimators are calculated in order to quantify the results. The estimators are divided into five categories:

(i) The naïve estimator: The raw sample composition;
(ii) Outdegree-based estimators: SH_{out} and VH_{out};
(iii) Indegree-based estimators: SH_{in} and VH_{in};
(iv) Estimators based on known population size N: SS_{out} and SS_{in};
(v) Estimators based on known parameter m^*: SH_{m^*} and VH_{m^*}.

Simulation setting: in each simulation, seeds are uniformly selected and coupons are randomly distributed to the recruiters' neighbors. To simulate RDS in real practice, we let the number of seeds be 10 and let the number of distributed coupons be 3 when shorter sample waves are desirable, and, 6 and 2 for longer sample waves. Sampling is done WOR and we choose sample size 500 for Net1 and Net2, and 1000 for the MSM network and Net3. All simulations are repeated 1000 times. In the estimation procedure of SS_{out} and SS_{in}, $M = 500$ times successive sampling samples per each of $r = 3$ iterations are used.

7.3.3 Result

Performance of RDS estimators on directed networks

Raw sample proportion, SH_{out} and VH_{out}: When the indegree and outdegree of nodes in the network are independent (no indegree-outdegree correlation, Net1), both *SH_{out}* and *VH_{out}* perform as poorly as the raw sample proportion as long as the network directedness is positive; when the indegree and outdegree are correlated (Net2), the biases of SH_{out} and VH_{out} increase with network directedness, and are smaller than bias of sample proportion. In both scenarios, bias and error increase with $\left|m^* - 1\right|$, i.e., the difference between average indegrees of groups of the studied variables.

SH_{m^*} and VH_{m^*}: When the ratio of average indegree of the studied groups is known, both SH_{m^*} and VH_{m^*} perform consistently well over all networks, and are robust to changes in the evaluated network structural properties, i.e., directedness, indegree correlation, indegree-outdegree correlation and attractivity ratio.

SS_{out}: When the size of the population is known, SS_{out} provides the minimum SD over all simulated networks, despite its uncertain biases.

SH_{in}, VH_{in} and SS_{in}: As it is impractical for researchers to collect individual indegree data, we implemented these estimators merely for theoretical purposes. When indegree of respondents is known, both SH_{in} and VH_{in} performs quite similarly to SH_{m^*} and VH_{m^*}, while SS_{in} performs similar to SS_{out}. These tests show that as long as indegree is known to researchers, it should be used instead of outdegree to approximate the inclusion probability of samples.

Application of the sensitivity testing method

From the results of sensitivity analysis on Net1 and Net2, it is shown that the performance of VH_m and SH_m is determined primarily by the attractivity ratio m^*, rather than by network directedness λ. Thus, if the network instead is assumed to be undirected, in which the ratio of indegrees is equal to the ratio of outdegrees ($m^* = w^*$), the sensitivity analysis may instead be used to assess the uncertainty of reported (out)degrees. The differential function of VH_m over m:

$$\frac{\partial VH_m}{\partial m}\Big|_{m=\hat{w}^*} = \left(\frac{n_A/n_B}{n_A/n_B + m}\right)'\Big|_{m=\hat{w}^*} = -\frac{n_A/n_B}{(n_A/n_B + \hat{w}^*)^2}, \qquad (66)$$

then provides the quantity by which the RDS estimate would change if there were any reporting error in the degree information.

Through the evaluations of VH_m and SH_m with varying m over the tested networks, it is shown that when the tested m value equals m^*, VH_m and SH_m can always generate estimates with minimum bias and error; when m departs from m^*, VH_m generate less RMSE, implying that when m^* is not known, VH_m may be a better option than SH_m in real practice.

7.4 **PAPER IV:** Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-Driven Sampling

7.4.1 Summary

Non-random recruitment (differential recruitment) is another harmful violation of RDS assumptions, according to Paper I. However, due to the chain-referral sampling design, once the sample is started from seeds, the distribution of coupons is largely out of the control of researchers, and non-random recruitment often occurs.

In order to improve the robustness of RDS estimates, we developed a new estimator, RDSI^{ego} for Paper IV by integrating traditional RDS data with ego network data reported by RDS respondents. The ego network data is collected by asking RDS respondents to report the composition of their peer networks, among which they would distribute coupons, regarding variables of interest, such as “What proportion of your friends is married?”. The new estimator shows improved reliability and validity and exhibits superior performance on the robustness to non-random recruitment, homophily, activity ratio and community structure.

7.4.2 Study design

Linked ego networks and the RDSI^{ego} estimator

When questions like “*What proportion of your IDU friends is married (is employed, is male, lives in this city, etc.)?* ” are asked in RDS interviews, the sample data can be illustrated as “linked ego networks”, in which egos are participants and alters are peers with characteristics of interests reported by their corresponding egos (see Figure 17).

For each respondent v_i in an RDS sample $U=\{v_1,v_2,\ldots,v_n\}$, let n_i^A, n_i^B be the number of v_i’s friends with property A, B, respectively. Given all RDS assumptions are fulfilled, the probability of each link $e_{i\to j}$ to be reported by “ego” v_i can be calculated as

$$\Pr(e_{i\to j}^{ego}) = \Pr(v_i) \sim \frac{d_i}{\sum_{j=1}^{N} d_j}, \qquad (66)$$

where d_i is the degree of v_i.

Based on , the proportion of type $e_{X\to Y}$ $(X,Y\in\{A,B\})$ links in the population, s_{XY}^{*}, can be estimated by the reweight proportion of type $e_{X\to Y}$ links in the sample:

$$\hat{s}_{XY}^{ego} = \frac{\hat{N}_{XY}^{ego}}{\hat{N}_{XA}^{ego}+\hat{N}_{XB}^{ego}} = \frac{\sum_{v_i\in X\cap U}\frac{n_i^Y}{d_i}}{\sum_{v_j\in X\cap U}\frac{n_j^A}{d_j}+\sum_{v_j\in X\cap U}\frac{n_j^B}{d_j}} = \frac{1}{n_X}\cdot\sum_{v_i\in X\cap U}\frac{n_i^Y}{d_i}. \qquad (66)$$

Replacing S_{XY} with $\hat{s}_{XY}^{ego}$ in (38) yields the improved RDSI estimator, in which the ego network information is integrated with personal network size information to estimate the proportion of individuals with property A in the population, P_A^{*}:

$$\hat{P}_A = \frac{\hat{s}_{BA}^{ego}\hat{\hat{D}}_B}{\hat{s}_{AB}^{ego}\hat{\hat{D}}_A+\hat{s}_{BA}^{ego}\hat{\hat{D}}_B} \text{ (RDSI}^{\text{ego}}\text{)}. \qquad (66)$$

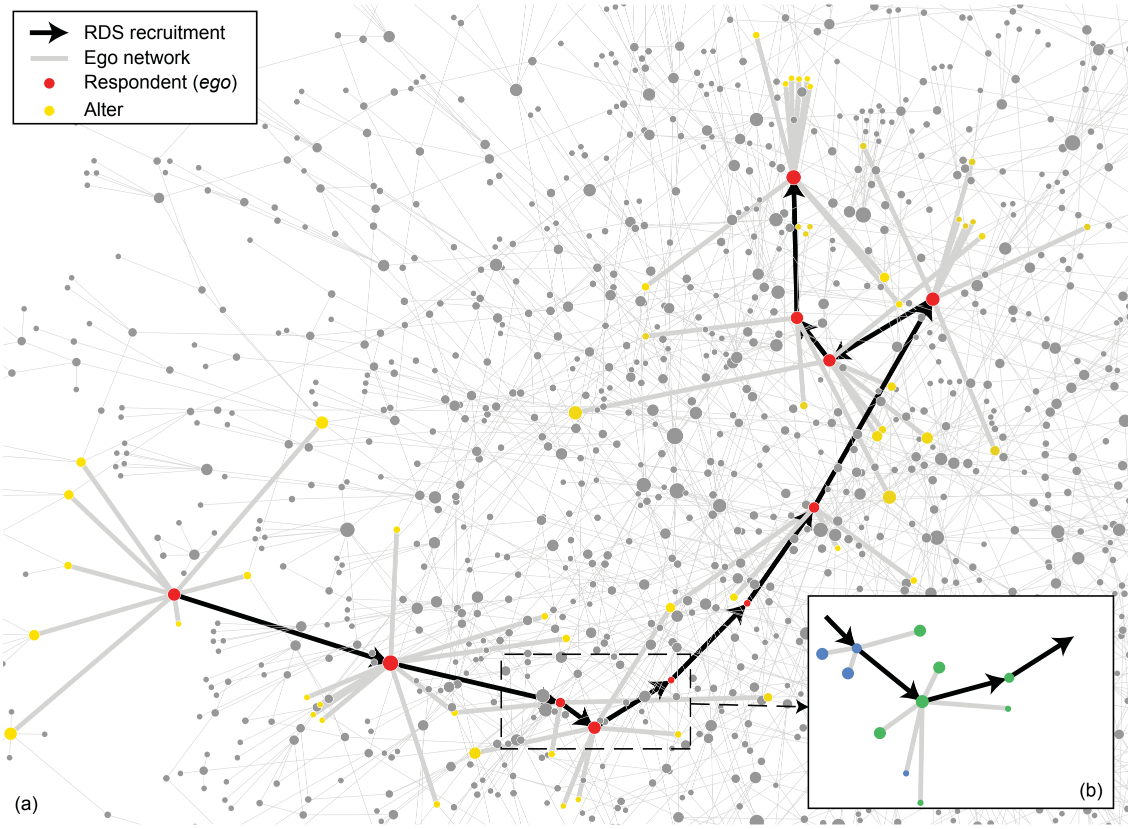


FIGURE 17 AN RDS CHAIN WITH EGO NETWORK DATA

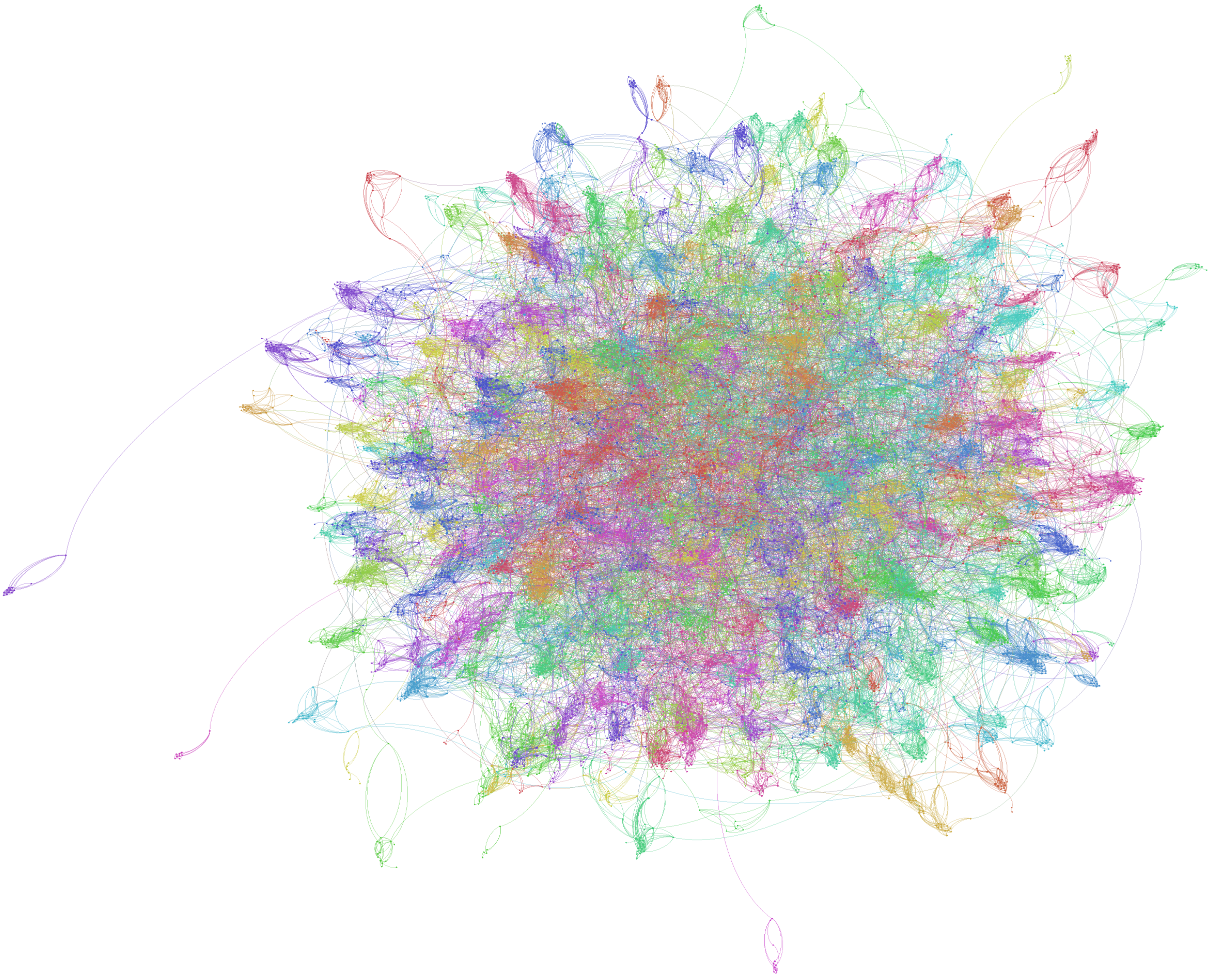


FIGURE 18 VISUALIZATION OF THE KOSKK NETWORK

Network data

In addition to the MSM social network, we have also generated a set of simulated networks with $h_A \in [0, 0.5]$ and $w \in [0.5, 2.5]$ based on the KOSKK model, which is among the best social network models that can produce the most realistic network structure with respect to degree distributions, assortativity, clustering spectra, geodesic path distributions, and community structure, and the like [133]. These networks are configured with population size $N = 10000$, average degree $\overline{D}^* = 10$, and population value $P_A^* = 30\%$. Model parameters are adjusted to produce networks with clear community structures (see Appendix in Paper IV for details).

Simulation procedure

General setting: We simulated RDS samples are collected without replacement with either 6 seeds and 2 coupons or 10 seeds and 3 coupons, sample size is 500. All simulations were repeated 10,000 times, and seeds were excluded from the calculation of estimates.

Random and differential recruitment: We modeled the presence of differential recruitment by the parameter p_A^{diff}, which represents the additional-than-random likelihood of recruiting peers from group A by any respondent.

Reporting error about degree and ego networks: We simulated reporting error at two stages of an RDS process: First, when a respondent reports his or her degree, any alters of type A or B will be missed and not reported with probability p_A^{miss} or p_B^{miss}, respectively; second, when the composition of an ego network is reported, any alters of type A will be misclassified as type B with probability $p_{A \mapsto B}^{error}$, and any alters of type B will be misclassified as type A with probability $p_{B \mapsto A}^{error}$.

RDS estimators: The raw sample composition, RDSI and $\mathrm{RDSI}^{\mathrm{ego}}$.

Measurements: Four measurements are then carried out after the RDS simulations: the *Bias*, the *Standard Deviation* (SD) of estimates, the *Root Mean Square Error* (RMSE) and the *Percentage* an estimator outperforms the rest in all simulations:

$$P^{best} = \frac{times\ the\ estimator\ gives\ closest\ estimate\ to\ s_{AB}^{*}\ or\ P_A^{*}}{m}. \tag{66}$$

7.4.3 Result

Random recruitment vs. differential recruitment

When the peer recruitment is random, RDSI performs as poorly as the raw sample proportion due to the relatively large variance. For age and civil status, the number of times when RDSI provides the closest-to-population estimates is even less than the sample proportion. On the other hand, $\mathrm{RDSI}^{\mathrm{ego}}$ in generally 15% access times gives the best estimates than RDSI for variables of the MSM networks.

When the sampling is done with differential recruitment, specifically under the extreme worst-case scenario that any peer of type A is twice as likely to receive a coupon from the ego compared to any peer of type B, the access time $\mathrm{RDSI}^{\mathrm{ego}}$ gives the best estimates increases to 70%~90%. While RDSI produces biases as large as 0.1~0.2, biases of $\mathrm{RDSI}^{\mathrm{ego}}$ are mostly less than 0.02, regardless of homophily, activity ratio and network community structure.

Sampling with degree reporting error

$\mathrm{RDSI}^{\mathrm{ego}}$ shows strong robustness to degree reporting error, i.e., being unaware of peers within the target population. Even when 20% of alters in respondents'' ego networks are unidentified, $\mathrm{RDSI}^{\mathrm{ego}}$ is still able to produce estimates with bias less than 0.05 most of the time.

Sampling with ego network reporting error

Reporting errors in the composition of ego networks have much larger effect on the precision of the $\mathrm{RDSI}^{\mathrm{ego}}$ estimator. When 20% of alters are misclassified as being group members with the opposite property, estimate bias can easily exceed 0.1. Misclassification errors regarding the group that comprises a large proportion of alters can increase bias and error significantly for $\mathrm{RDSI}^{\mathrm{ego}}$, as a substantial amount of misclassified alters will be reported.

8
CONCLUSIONS AND DISCUSSION

8.1 KEY FINDINGS AND CONTRIBUTIONS

8.1.1 Performance of RDS estimators

Paper I is, to our knowledge, the first comprehensive evaluation of RDS on an empirical social network extracted from a real hidden population. We have found the most harmful conditions as well as some beneficial conditions for RDS estimates.

Ideal scenario: When all the assumptions are fulfilled, the RDSII estimates approached the true population value very quickly but had high design effects, especially for the two variables with high homophily.

Harmful violations: The most harmful violations of assumptions are network directedness and recruitment behaviors (i.e., rejecting coupons, miscounting peers, or preferential recruitment) that depend on participant characteristics. Both the SD and the MAE increase when more coupons are used; it becomes even worse when a large number of coupons are combined with small number of seeds. Presumably this way of implementing RDS would end the sample in a limited number of waves and the respondents all come from a block of the network that is very different from the whole population, leading to large bias and error. On the other hand, when more seeds are used (either selected through simple random sampling or proportional to degree), the diverse starting points result in decreased SD and MAE.

Beneficial or non-relevant violations: There are also violations that do not affect the performance of RDSII significantly, or are even beneficial for the precision of estimates, such as SWOR when sample proportion is small, or participants' behavior of rejecting coupons or miscounting peers is independent of their characteristics.

Other important factors are also evaluated, including homophily, network density, degree distribution, and the like. Generally, the RDSII estimates perform better if there is little homophily, or the network density is high, or the degree distribution is homogenous. Other important factors are also evaluated, such as homophily, network density, degree distribution. These effects on the performance of the RDSII estimator are summarized together with those discussed above in Table 5.

TABLE 5 EFFECTS OF CONDITIONS EVALUATED IN PAPER I

Violation of assumptions	Bias, SD & MAE
If you would not recruit your recruiter (irreciprocal relationships)	**Increase**
If you are not allowed to participate twice (sampling without replacement)	**Increase**: large sample proportion **Decrease**: small sample proportion
If you are more likely to invite your close friends (non-random recruitment)	**Increase**
If you have difficulties in counting your friends, or refuse to participate	**Increase**: behavior depends on group type **Decrease**: behavior is independent of group

(Table 5 *cont'd*)

(degree reporting error, low response rate)	type
If the survey starts with many initial participants and you are allowed to recommend many friends	**Increase**: many coupons and a few seeds **Decrease**: a few coupons and many seeds

Other factors

If network is dense	**Decrease**
If variance in degree is small	**Decrease**
If homophily is low	**Decrease**
If seeds are selected proportional to nodes' degree	No effect

8.1.2 Improved RDS estimate methods

The evaluation study reveals that network directedness and outcome-correlated recruitment behavior are the two most harmful violations of RDS assumptions. We have developed improved estimate methods for each of these conditions.

The primary contribution of Paper III is that it shows that indegree of nodes is a fairly good approximation of inclusion probability for RDS on directed networks, and that this approximation is robust to changes in indegree-outdegree correlation and indegree correlation. We have developed a sensitivity analysis method, based on the attractivity ratio m^*, to incorporate the uncertainties in both network directedness and reported outdegrees. Our results show that, while it is of course best to have correct indegree information on the network, it is possible to get a deeper understanding of how RDS estimation is influenced by network directedness by using sensitivity analysis. An illustration of such a sensitivity analysis has been presented for the estimation of proportions of males and injectors among drug users in New York City (see Figure 19). The sensitivity analysis approach enables us to quantify the level of changes the RDS estimates will be: for each change of 0.1 in the average indegree ratio, the change in the RDS estimates will be about 2 percentage units.

By collecting ego network data with RDS, in Paper IV I developed a new estimator, RDSI^{ego}, to improve the validity and reliability of RDS estimates. RDSI^{ego} is superior to traditional RDS estimators. Most importantly, RDSI^{ego} exhibits strong robustness to differential recruitment, a violation of the RDS assumptions that may cause large bias and estimation error and is not under the control of the researchers. Evaluation studies on the simulated KOSKK networks also show that RDSI^{ego} performs consistently well in networks with varying homophily, activity ratio, and community structures.

Compared to a few other newly developed estimators that require population size as a priori information, such as the SS-estimator and GH-estimator[*], the main advantage of RDSI^{ego} is that it enables researchers to improve the precision of estimates with a feasible implementation: unlike population size, which is usually unknown for hidden population, the ego network data can be collected directly from sample respondents. Such data have been collected in a few RDS studies, for example, in an RDS study of MSM in Campinas City, Brazil, by de Mello et al [198], respondents were asked to describe the percentage of certain characteristics among their friends/acquaintances,

[*] The GH-estimator requires both the population size and characteristics of ego networks from participants.

such as disclosure of sexual orientation to family, HIV status, and the like. An RDS study of opiate users in Yunnan, China, various information about supporting, drug using, and sexual behaviors between respondents and their network members were collected [199]. One of the most thorough RDS studies utilizing ego network information was done by Rudolph et al [200], in which they asked the respondents to provide extensive characteristics for each alter within their personal networks such as demographic characteristics, history of incarceration, and drug injection and crack and heroin use.

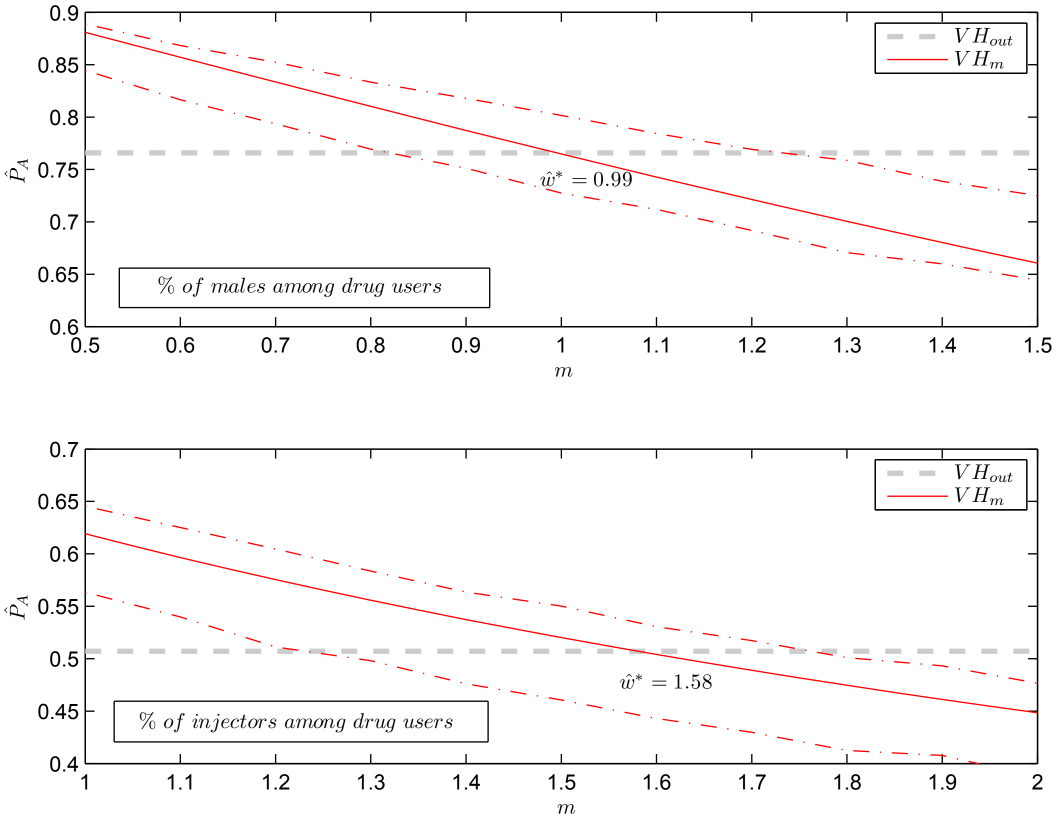


Figure 19 Sensitivity analysis of RDS estimates for proportion of (a) males and (b) injectors among drug users in New York City.

8.1.3 Implementation of WebRDS

We have demonstrated for the first time that it is possible to implement an Web-based RDS system to sample 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 [196]. We successfully used the system to sample and survey 676 MSM on a number of sensitive issues. The plots of sample composition show clear independence of the seeds and stabilization for all variables, with the exception of home province.

By comparing national statistics and other published research data with our estimates, it has been shown that the WebRDS sample is younger and of higher education than the Vietnamese average. The sample is heavily concentrated in the two large metropolitan areas of Ho Chi Minh City and Hanoi, but also spread outside the large metropolitan areas with 32 provinces out of 63 represented, which speaks to the high degree of flexibility of recruiting location-free samples using WebRDS.

WebRDS will in most cases entail a lower costs than a standard RDS study. The cost of monetary incentives in our study was on average 5.9 USD per participant in the cleaned sample (3353 USD in total). Staff hours to interact with seeds, deliver incentives, monitor invalid submissions, and the like, totaled a one-month full-time equivalent (FTE). Adjustment of the site to appeal to the local target group is technically easy but requires research. In comparison, an offline RDS would have shared similar costs for incentives and formative research about the study population (see e.g. [197]) but would also require a survey office and at least five months of staffing (conservative FTE estimate).

8.2 LIMITATIONS

8.2.1 Limitations of VH_m estimator

The information on respondents personal network size (outdegree) is generally collected by asking questions like "How many people do you know?", where "knowing" is usually defined as "You know them and they know you by sight or by name" [202], or by nomination: "Who do you consider to be your friends on this list?" [169,170]. The use of the suggested indegree-based estimator VH_m brings new challenges for RDS practice, because indegrees are difficult to collect. Indegrees reported in the literature are generally inferred from nominations in studies which are conducted within closed networks. There are currently no guidelines for researchers to ask respondents about their indegree in RDS practices; potential questions such as, "In the studied population, how many people do you think will recruit you if they have got a coupon?" might cause respondents to report inaccurate indegrees since it is hard to guess the number of incoming irreciprocal links; some people who should be included might even not be known to the respondent.

There are, however possibilities to gain knowledge about m^* for the studied population. First, it is sometimes reasonable to make assumptions about the ratios of average indegrees between studied groups, thus making it possible to utilize our estimators through sensitivity analysis. In the simplest scenario, for example, one might assume that those with HIV will be less known compared to those without in a population where HIV has a strong social stigma; thus $m^* < 1$, and it is safe to choose an interval of m with a maximum value less than 1.

Second, since many social networks have a positive indegree-outdegree correlation, the activity ratio $\hat{w}^*$, which is observed from the sample, may be an indicator of where to vary m from. Actually, in the MSM network, we find the difference between m^* and w^* is small for the studied variables; the absolute difference is 0.27, 0.17, 0.02, and 0.05 for age, county, civil status and profession, respectively.

Third, prior information about m^* may be obtained by using empirical studies related to the studied population. For example, in the Baltimore Needle Exchange Program [203,204], the authors suggested to use bar-coded syringes to infer the inner needle exchange network among IDUs, where "outdegree" is inferred by the number of people who returned each person's needles, and "indegree" is the number of people for whom each person returned needles. While such estimates will contain many uncertainties, the existence of long-term follow up studies of the networks of friendship, sexual behavior, and needle sharing for HIV-related high-risk populations, such as the HIV Transmission Network Metastudy Project [205,206], enable researchers to gain a deeper understanding of such populations and thus come closer to inferring m^* from such populations.

Lastly, the rapid increases in Internet-based surveys indicate a promising application field for the proposed method. For example, when participants are restricted to recruiting only through established contacts on their membership Website, a Web-based RDS study would easily adopt the new method and utilize indegree information that is already available in the database such as the *qruiser* Website used in this study. Additionally, the indegree-based estimators would have a wide application in sampling Web contents, where the indegree of Webpages is likely to be more accessible than in empirical RDS studies.

8.2.2 Limitations of RDSI^{ego} estimator

The limitation of $\mathrm{RDSI}^{\mathrm{ego}}$ is rooted in the need to collect ego network data. Many RDS studies are designed to sample hidden populations, and members of such populations may be reluctant to share sensitive information with their friends. Consequently, the proposed method is primarily suited to less sensitive variables. Such information may, for example, include sociodemographic variables (e.g., gender, age groups, profession, marital status, etc.) for which survey methods regarding the design and collection of ego network data has been extensively studied [207-210]. Additionally, certain variables, such as drug use may be highly sensitive in the general population but may not be in an IDU population.

By modeling the difficulty in understanding personal network composition as a degree reporting error and ego network reporting error, which quantify the level of mutual knowledge about studied variables shared with friends, we have showed that even with 20% of alters being unidentified, $\mathrm{RDSI}^{\mathrm{ego}}$ was still able to produce estimates with a bias of less than 0.05 most of the time. On the other hand, $\mathrm{RDSI}^{\mathrm{ego}}$ is sensitive to the error of misclassifying alters. If 20% of alters from one group are mistakenly reported as belonging to the other group, estimate bias can exceed 0.1 when the probability of misclassifying members of one group is substantially larger than misclassification of members in the other group (e.g., $p_{A\mapsto B}^{error} \gg p_{B\mapsto A}^{error}$). Fortunately, the result shows that when the studied variables only related to a small proportion of alters, that is, if P_A^* is low and w is relatively small, the increase in error in misclassifying A as B members will have a small influence on the bias. Consequently, for many sensitive variables surveyed in RDS studies, if the reporting error of a low prevalence trait (e.g., HIV status) is mainly “false negatives”, e.g., alters with HIV are reported as healthy friends since they are reluctant to reveal this information to their egos, estimates with small bias are still expected to be able to achieve.

8.2.3 Limitations of WebRDS

Hard to verify the MSM identity of participants: As with standard RDS surveys and other sampling strategies, it is very difficult to identify whether a participant in WebRDS is truly a MSM. We are confident that the seeds belonged to the population, and based on the characteristics of the sample (e.g., 57% had a boyfriend by the time of the survey, 79% supporting same sex marriage and 68% prefer only men as sexual partners), it is very unlikely that MSM participants have invited non-MSM persons on a massive scale. We have also asked around among our test persons whether they had heard about misuse of the survey but we did not get any such reports.

Vulnerable to duplicated submissions: Unlike physically located face-to-face interviews, where the uniqueness of each participant can be highly controlled (e.g., with fingerprint scans), WebRDS allows anonymous persons to participate as long as they have a valid coupon. Therefore, it is very easy for a participant to take the survey repeatedly with the newly generated expected-to-be-distributed coupons. In the sample, for example, 63 participants provided a repeated email address or telephone number among the 634 adult participants. If we consider submissions with any repeated IP numbers to be frauds, the sample size decreases to 490. Fortunately, our analysis has shown that those suspiciously duplicated participants did not make any substantial population estimate difference.

Evidence for violations of RDS assumptions: As with all other empirical RDS studies, our sample suffers from several violations of RDS assumptions. First, each participant was allowed to recruit a maximum four other persons and not all recruitment was successful (violation of assumption vi). 305 participants had recruited at least one respondent (average 2.27); however, given that the sample stopped by itself, the average recruitment per participant was one. Since the recruitment chain sustained as many as 24 waves, according to Paper I, we believe the violation of one coupon assumption did not significantly bias the RDSII estimates.

Second, 8% of participants were recruited by a stranger and only 41% were recruited by a friend, a clear violation of assumption ii and possibly assumption v. Such a level of network directedness was also observed in other studies. We do not think this caused serious bias in this study (see Paper I and Paper III), however estimates should always be interpreted with caution especially when the network directedness may be compounded by non-random recruitment.

Third, all participants were instructed that they were only allowed to participate once in this study, a violation of the SWR assumption. Since our sample size comprises far less than a majority of the MSM population [201], according to Paper I it is very unlikely that the practice of SWOR would have an effect on the estimates.

As it is not possible to evaluate the level of all possible violations, such as whether the recruitment behavior depended on characteristics of participants, or whether the participant had correctly counted all potential MSM friends he could recruit, interpretation of population estimates from the RDS sample should be conditioned on these uncertainties.

8.3 IMPLICATIONS FOR RDS PRACTITIONERS

8.3.1 Accessing hidden populations over the Internet

The Internet provides stigmatized individuals fast and easy access to the study, with minimal exposure of identity. The successful implementation of the WebRDS study for MSM in Vietnam demonstrates that it is possible to recruit hard-to-access groups through the Internet.

We have developed a system that allows researchers to design their own questionnaires, publish the survey on the Internet, and to recruit respondents automatically. Obviously, implementing RDS with such an electronically available system would significantly reduce work load and be cost efficient.

Physically isolated groups may be reached with WebRDS, given that online social links between individuals of these groups exist such as Internet games, online social networks, and the like. With the prevalence of the Internet and smart phones, the overlap between the target hidden population and the hidden population that uses the Internet will be increased.

Because it is able to provide both easy access and population estimates, WebRDS may also be useful for studies of other types of online populations when there is a lack of sampling frame, or a fast first hand sample is needed regardless of its representativeness. Examples of applications include registered Web forum users, university email users, Internet game players, Facebook[*] or Twitter[†] users, etc.

8.3.2 Seeds, coupons and sample size

Determining the number of coupons per participant and the number of and characteristics of the seeds are among the first problems that are encountered by researchers when preparing an RDS study. We have shown that for the empirical MSM network, the SDs and MAEs are almost unchanged if we shift from randomly selected seeds to seeds with higher degree, this property is consistent with the Markov model of RDS, which implies that the dependence of sample composition on the characteristics of original seeds will decrease quickly. However, a few other studies, with simulated RDS on very small networks [175,182], show that the initial selection of seeds biases RDS estimates. As these studies are implemented on small simulated networks ($N = 1000$) and sample sizes are relatively large (≥500), I suspect that the sample composition is not stationary when the sample size is reached.

For RDS users, it is always recommended recruitment be started with seeds as diverse as possible, and that the maximum number of coupons a participant can distribute be limited. Such a design would be most helpful when the target population is loosely connected and has strong community structures, as the diverse seeds and limited number of coupons will force the recruitment chains to move out of local communities and to penetrate into diverse parts of the network, yielding a sample with improved representativeness. According to our experience, a pilot test would be very useful for the sampling design.

Sample size is yet another parameter to be determined at the planning stage. Inadequate samples sizes have been used since the invention of RDS. As discussed in 4.5, most RDS studies have assumed a design effect of one or two. However, our evaluation study of RDS on the empirical MSM network shows that the design effect can be as high as 13; the design effects are 5 even for variables with very low homophily. Consequently, combining with a few other studies, we would suggest a design effect of 5~10 to be used to achieve the proper precision of RDS estimates. Goel and Salganik have illustrated how to use design effect to determine the sample size of RDS [146]. To

[*] https://www.facebook.com/
[†] https://twitter.com/

have sufficient statistical power to detect a decline in unsafe injection practices from 40% to 30%, SRS would need about 350 respondents at each of the two time points; with a design effect of 5, the required sample size for RDS would be 1750!

8.3.3 Treatment for violation of assumptions

Violations of assumptions in RDS practices are inevitable [211-213]. Consequently, it is important for researchers to identify and investigate harmful violations and to make adjustments if possible.

8.3.3.1 Network connectedness

When the network is disconnected, RDS would only recruit respondents from networks that are connected with the seeds, and sample results are only valid from the network where the sample is drawn. If there is a big difference between the connected network and isolated groups, large estimate bias may occur. As the social network is invisible to researchers, it is very hard to infer whether there are isolated groups based on the sample data. However, it is possible to reach isolated groups by increasing the diversity of seeds. Note that when the network is disconnected, inclusion probabilities of individuals would be based on the selection probability of seeds for each component, and the RDS estimators are not valid anymore. Fortunately, network disconnectedness is not a major concern for many populations as human societies are increasingly highly interactive, and the “small-world” phenomenon ensures that most of target populations are connected and can be reached through a limited number of waves.

8.3.3.2 *Irreciprocal relationship*

Despite the widely acknowledged evidence of the existence of directedness among social networks, the effect of directedness on RDS estimates has seldom been evaluated. This could be problematic since all previously reported RDS estimates rely on the assumption that the studied networks are purely reciprocal, the violation of which will result in unknown biases.

Many RDS studies have included a question in the survey to access the relationship between the recruiter and recruit. For example, in an RDS study of IDUs in Sydney, Australia [214], 29% of the respondents considered the relationship to their recruiter to be “not very close”, and in a study of IDUs in Tijuana, Mexico [215], only 62% of the respondents considered their relationship with their recruiter as “friend”. On these occasions, we encourage use of the VH_m to test the sensitivity of population estimates to the changes of attractivity ratio m. Priori information about where to start the test interval is discussed under 8.2.2.

8.3.3.3 Sampling without replacement

Even RDS assumes SWR, empirical RDS studies are conducted by SWOR. We have shown that when the sample size proportion is relatively small compared to the size of population, e.g., ≤10%, implementing RDS by SWOR can actually improve the performance of RDS estimators with decreased SD, MAE and DE. However, when the sample size proportion is relatively large, e.g., ≥50%, it has been shown that SWOR will generate large estimate bias and error and the SS- or GH-estimator can be used.

The trick is that the sizes of hidden populations are mostly unknown, the SS- or GH-estimator, however, needs population size as an input parameter to generate estimates.

We therefore suggest that the SS- or GH- estimator be used as a sensitivity test method to access the magnitude of changes of estimates by varying tested population sizes.

8.3.3.4 Degree reporting error

Relationships are complex, it is extremely difficult for respondents to accurately report their personal network sizes, as relationships maybe directed or weighted, or the respondent can easily misclassify target population members. We have shown that as long as the reporting behavior is not dependent on the outcome variables, the reporting error has little effect on the performance of RDS estimators; however, when the reporting error is correlated with the outcome variables, e.g., if there is a substantial exaggeration of personal network sizes in one group compared to the other, large estimate bias and error may occur.

8.3.3.5 Differential/nonrandom recruitment

According to our study, differential recruitment is one of the most harmful violations of RDS assumptions. Due to the automatic design of RDS, once the sample is started from seeds, the distribution of coupons is largely out of the control of researchers, and non-random recruitment often occurs. For example, respondents may tend to recruit people who they think will benefit most from the RDS incentives [216]. In a study of MSM in Campinas City, Brazil [198], participants were most often reported to recruit close peers or peers they believed practiced risky behaviors. In [137,175,180], it has been shown that all current RDS estimators would generate bias when the outcome variables are related to the tendency of such non-random distribution of coupons among respondents' personal networks.

It is possible to assess the severity of differential recruitment by collecting ego network data and comparing the ego network based estimator for recruitment matrix, $\hat{S}^{ego}$, with the observed raw sample recruitment matrix S. Either when the recruitment is random or when there is substantial differential recruitment, the ego network based estimator for population characteristics, $\mathrm{RDSI}^{\mathrm{ego}}$, can be used to greatly improve the precision of RDS estimates. However, due to the limitations inherent in the collection of sensitive variables from stigmatized group, reliable ego network data may be difficult to collect for sensitive variables.

8.3.3.6 Response rate and number of coupons

To stimulate the recruitment process and to prevent recruitment chains from stopping early due to low response rates, researchers often use more than one coupon in RDS studies [136,149]. We have seen that the estimate bias and error will be large when too many coupons are used per recruiter. This effect occurs because a large number of coupons will make the sample size increase explosively. If all invitations successfully generate new participants, the desired sample size will be reached within a small number of sampling waves. A similar discussion was presented in 8.3.2. Researchers should try to achieve a compromise between recruitment efficiency and response rate: if the response rate is high, fewer coupons should be used to avoid ending recruitment in short waves; otherwise, more coupons should be used to keep the recruitment alive. In short, recruitment chains should be as long as possible.

8.3.4 A note on the raw sample proportion and variance estimation

It is worth noting that current RDS variance estimate methods are far from satisfying. The traditional bootstrapping method has been found to largely underestimate the variance, while other studies have found the MCMC-based method tend to overestimate.

A few newer variance estimate methods, including our VH_m and $\mathrm{RDS}^{\mathrm{ego}}$ estimator and Gile’s SS- and GH- estimator, have shown improved performance regarding the coverage rates for CIs. However, all of these methods, except $\mathrm{RDS}^{\mathrm{ego}}$, require certain population values as input parameters: the VH_m estimator needs attractivity ratio, and the SS- and GH- estimator needs population size.

When population values are not known, which is common for HIV/AIDS-related high-risk populations, $\mathrm{RDS}^{\mathrm{ego}}$ is the only estimator with improved variance estimates. However, $\mathrm{RDS}^{\mathrm{ego}}$ is based on the collection of ego network data and even the improved variance estimates are not able to reach the desired coverage rates. Consequently, future research is needed to develop feasible RDS variance estimate methods with a feasible implementation.

Due to the large variance in RDS estimates, consistent with a few other evaluation studies, we found that the raw sample composition outperforms the RDSI/II estimates in considerable times, implying that it is wise for RDS studies to report both the raw sample composition and the RDS estimates, and that RDS sample results should be always be interpreted with caution.

8.4 CONCLUDING REMARKS AND FUTURE RESEARCH

8.4.1 Concluding remarks

Over the past decade, Respondent-driven sampling (RDS) has been adopted outside the US and has soon become the state-of-the-art sampling method for studying HIV/AIDS-related high-risk populations worldwide. First, it enables researchers to gain access to population members and obtain reliable biological and risk behavior information with high response rate (*accessibility*), and second, it allows them to make population inferences from the sample data, which can then be used to guide to set up efficient prevention and intervention HIV programs (*generalizability*).

I conclude the thesis by revisiting the advantages and disadvantages of RDS:

Advantages

- *Improved recruiting efficiency*: In RDS, a participant is rewarded both for his own participation and for each of his successful recruitments. This kind of dual incentive mechanism stimulates respondents to encourage peers to participate, or to pass coupons to those they think are more likely to participate in the study.
- *Reduced stigma or sensitivity concern*: The peer-driven design allows individuals who received the coupon to decide on their own whether to

participate. Consequently, respondents recruited in the sample are more likely to cooperate and provide reliable answers to sensitive questions.

- *Improved access to target population*: Like all other chain-referral sampling methods, RDS enables researchers to recruit population members who are "remote" in terms of distance or knowledge. RDS strengthens this ability by limiting the number of coupons per participant, forcing the recruitment chains to grow long and allow researchers to explore diverse parts of the target population.
- *Can be combined with Internet or smart phones*: With authorization codes substituting for coupons and electronic surveys substituting for physical interviews, RDS can be implemented online with easy access to participation and maximized anonymity.
- *Cost-effective*: Implementation of RDS is generally considered cost-effective. RDS adopts a rather simple design: once the seed is selected, the recruitment process will continue automatically, and researchers usually conduct interview in fixed locations for respondents seeking participation, thereby minimizing traveling expenses and administration cost. If the sampling is implemented online, cost savings can be more significant with an available automatically RDS recruiting system.
- *Unbiased population estimates*: Under certain assumptions, RDS is able to generate asymptotically unbiased estimates for population characteristics based on sample data. This feature is particularly important for the study of hidden populations, where no sampling frame exists and there is usually a lack a representative random sample.

Disadvantages

- *Sample recruitment relies on social network*. RDS is fundamentally a chain-referral sampling method and thus bears the disadvantage of relying heavily on the social network connections between population members. If the network is disconnected into many isolated groups, or the connection between individuals is too loose such that the recruitment chains fail to proceed, RDS will not be able to recruit sufficient samples.
- *Peer-driven may bring sampling bias*. This is yet another common issue with all chain-referral sampling methods. As newer participants are invited by their friends to participate into the study, respondents may, for various reasons, invite differently, for example to avoid inviting certain friends to protect them from exposure, or to invite relatives or close friends to get rewarded.
- *Estimates rely on rigorous assumptions*: Hardly any assumption of RDS estimators can be met in practice; both our study and a few recent studies have recently found that large bias and estimate error may be generated when certain assumptions are violated. This thesis expends major effort on improving the reliability and validity of RDS estimate methods under real conditions.

Compared to other nonprobability sampling methods, RDS has shown improved accessibility in the study of hidden populations, especially for HIV/AIDS-related high-risk populations such as MSM, IDUs and SWs. The ability to produce population estimates, which are usually not available in nonprobability sampling, has made it an even appealing option.

However, as the variance of estimates is considerably high, and assumptions for RDS estimates can hardly ever be met in practice, we recommend that RDS users thoroughly investigate violation of assumptions and make adjustments if possible. Results from RDS samples should be interpreted with caution, and researchers should bear in mind that the raw sample proportions are very likely to be closer to the true population value than RDS estimates. Researchers are encouraged to collect ego network data through the implementation of RDS to improve the reliability and validity of population estimates.

More precisely, to implement RDS and use RDS estimates properly, researchers are advised to refer to Table 6, where the performance and treatments for RDS under various conditions are summarized based on this thesis.

8.4.2 Future research

These following points are potentially interesting and important directions for future research that relates to the current development of RDS methodology and work of this thesis:

- To improve the performance of RDS estimators by combining different data sources;
- To improve the methods for estimating the variance of RDS estimates;
- To implement WebRDS in other settings of hidden populations;
- To implement RDS for sampling of Web content, Internet users, etc., and to use VH_m or $\mathrm{RDS}^{\mathrm{ego}}$ if applicable;
- To apply $\mathrm{RDS}^{\mathrm{ego}}$ for RDS studies in which ego network data is reported, and to access the data quality of ego network data; and
- To use RDS to collect detailed sex risk behavior data of HIV/AIDS-related high-risk populations and to build epidemiology models.

TABLE 6 SUMMARY AND RECOMMENDATIONS FOR RDS ESTIMATES IN PRACTICE

RDS assumptions	Violation in practice	Consequence for RDS estimates (Bias, SD & MAE)	Recommendation
Connectedness	Population is formed by socially isolated groups	**Increase** *(results only valid for sampled groups)*	Increase diversity of seeds, WebRDS (Paper II)
Reciprocal	You received a coupon from a stranger	**Increase**	Use VH_m to test the sensitivity of estimates to the network directedness (Paper III)
Sampling with replacement	You are not allowed to participate even you get a second coupon from another friend	**Increase**: large sample proportion **Decrease**: small sample proportion	SS- or GH- estimator: large sample proportion No need to adjust: small sample proportion (Paper I)
Accurate degree	You only know approximately the number of your MSM friends	**Increase**: behavior depends on group type **Decrease**: behavior is independent of group type	Define degree clearly, use VH_m to test the sensitivity of estimates to the network directedness (Paper III)
Random recruitment	You prefer to give coupons to your close friends	**Increase**	Collect ego network data, use $RDSI^{ego}$ to improve the precision of estimates (Paper IV)
One coupon	You get four coupons after completing the survey	**Increase**: many coupons and a few seeds **Decrease**: a few coupons and many seeds	Limit the max number of recruits per participant; try to recruit as long as possible. (Paper I)
Other factors			
Homophily	*Increase of homophily*	Increase	$RDSI^{ego}$ is robust to changes of homophily (Paper IV)
Undirected network	*Increase of activity ratio*[*]	Increase	$RDSI^{ego}$ is robust to changes of activity ratio (Paper IV)
	Increase of network density	Decrease	--[†]

[*] Increases from 1.

[†] No need to adjust.

(Table 6 *cont'd*)

	Increase of degree heterogeneity[*]	Increase	----[†]
Directed network	*Increase of attractivity ratio*[‡]	Increase	VH_m (Paper III)
	Increase of directedness	Increase	VH_m (Paper III)
	Increase of indegree correlation[§]	Not relevant	VH_m (Paper III)
	Increase of indegree-outdegree correlation	Decrease	VH_m (Paper III)

[*] From Poisson to skewed degree distribution.
[†] No treatment was discussed.
[‡] Increases from 1.
[§] Assortativity calculated based on indegree.

ACKNOWLEDGEMENTS

Work on this thesis would not have been possible without encouragement and support from many people. The greatest thanks go to my supervisors, I cannot be luckier to have such a great team to guide my study and research. First to you my main supervisor Professor Fredrik Liljeros. Together with Johan Giesecke, you were my very early contacts to Sweden five years ago. Great passion in research, always surprises everyone with brilliant ideas in discussions, broad collaborations among very different disciplines, and most importantly, a kind and selfless heart, life is so much easier to work with you during my PhD. I cannot appreciate more for the inspirations, guidance and support from you. Hope we can continue to work together in the future for a long time.

Anna, you are my main supervisor at Karolinska, and the core person who links me to IHCAR and the wonderful HIV/AIDS and Global Health Research Group. You are such a kind boss who never gets angry or complain. Comfort and strength are always what I get from you. Thank you so much for guiding me through the study and research these years, you and Fredrik have let me be the most free-to-do-what-I-want PhD student. I enjoy all the collaborations with you within and beyond the PhD project, let's do more afterwards.

Monica, thanks for all the help with reports and applications. You live a bit far away from Stockholm, but I know you are here whenever there is a need.

Petter, my mentor as well as a great collaborator, a supervisor and a friend. Thanks for translating Swedish to English and English to Chinese to me, thanks for inviting me to the inspiring IceLab in Umeå, thanks for recommending me to Santa Fe Institute, thanks for all the introductions and discussions on frontier network science, and, most of all, thanks for being such a great collaborator on the Haiti project, your insight on physics and complex network science is a great source of innovations and solutions to research questions. I believe there is more for us to move on in the future.

Linus, the best ever colleague, co-author, collaborator, co-founder and friend, we went through this together. I am always surprised by the thorough thoughts you made into the details. You have also a great ability of networking and collaborating. It has been a great time for the traveling, conferences, meetings and discussions together. Thanks also to all the help on language and Swedish culture tours. Coding, data processing and discussions were what kept us staying late in the office during the last four years, now I am happy to give you back to Nathalie and wish you a wonderful journey in Africa.

Erik, thanks for taking care of Flowminder so well, it has been great to know you and I have much to learn from you, the always positive way of thinking, the finest art of negotiating, and the insight on economic theories. But I do hope your prediction on the economic crisis never come true.

Tom, you are like my supervisor. Thanks for the discussions during workshops and research meetings where we try to improve the methodology of RDS, thanks for taking the call whenever there is a need for mathematical clarification, and thanks for the kind advice on my presentations.

Vinod, you are also the core person who makes me feel part of Karolinska. Thanks for sharing of your broad expert knowledge with an open mind during our meetings and discussions. I enjoy our collaborations and have learned a lot in the ISSC project.

Weirong, thanks for recruiting me into the ISSC project and for sharing your knowledge as a senior researcher. It has been a wonderful journey during the process of this project, and a precious learning experience. Hope we will continue to collaborate for a very long time.

Beom, our meetings and discussions formed the very early research plan of this thesis. Thanks for introducing me many physic approaches, for answering each of my ask-for-help emails in the past years, as well as for the guided walking tour in Stockholm during one of your occasional visit. Next time, I will be the guide.

Diego, thanks for the insightful comments and discussions for the analysis of QX network data. Hope we can continue to work on it in the future.

Jens, thanks for the hard work on Paper III, and all the timely help.

Martin Camitz, thanks for your contribution to our shortest path paper and for all the timely help for the WebRDS system.

Thanks Asli, Gaetano, Gun-Britt, Maud, Kersti, Marie and Bo et al. for your support from IHCAR, and all the timely help whenever I had an ask.

Thanks are also due to people who I have either formal or informal collaborations, your encouragement and inspirations are important to me: Abela Agnarson, Abigail L. Horn, Amir Rostami, Christa Brelsford, Christofer Edling, Ian Wood, Johan Von Schreeb, Marco Dueñas, Oleksandr Ivanov, Susanne Strömdahl, and Richard Garfield et al. Hope we can keep working on what we have achieved and explore more in the future.

My warmest thanks goes to the Sociology Department at Stockholm University, where I spent most of the time worked there and occupied the office on so many weekends and holidays. Thank you Karin, Saemundur, Snorri, Maria Bagger-Sjöbäck, Maria Lind and Thomas et al. for letting me stay smoothly and be least disturbed without going through all the bureaucratic, most have been taken care of without me knowing about it. I was very lucky with my office room-mates: Love, Martin Hällsten, Jani, Lambros, Linda Kridahl, Brita and Alejandro, from whom I often seek for Swedish translation and practical information, thanks for always being available for a chat between all the modeling and simulations.

A general but very special “thank you!” goes to all the present and former colleagues at IHCAR from Karolinska Institutet and the Sociology Department at Stockholm University, thank you for the four years hospitality and support: Atika, Anna Mia, Barbara, Bharati, Linda Sanneving, Lisa, Klara, Elin, Emma, Helena, Hanani, Gergei, Christofer, Gunnar, Martin Gerdin, Mikaela, Yvonne, Li, Andrzej, Veronika, Hernan, Patrick, Theodora, Sandeep, …, the individual name list is simply too long to be included here.

Besides my connections in Sweden, I want to thank my supervisor and colleague at the Department of Information Systems and Management at NUDT in China: YJ Tan, HZ Deng, J Wu, Y Li, B Liu, B Cheng, K Sun, BF Ge, HW Dong, RJ He. Thanks for all the help for keeping my official duties there.

Thanks to my mom, my brother and all my family members for consistent supporting my studies.

Bao, my love, so much had happened since we met, thank you for accompanying me these years.

References

1. GAO [U.S. Government Accountability Office] (2011) 2010 Census: Preliminary Lessons Learned Highlight the Need for Fundamental Reforms.

2. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2009) Survey Methodology: Wiley.

3. Robinson WT, Risser JMH, McGoy S, Becker AB, Rehman H, et al. (2006) Recruiting injection drug users: A three-site comparison of results and experiences with respondent-driven and targeted sampling procedures. Journal of Urban Health-Bulletin of the New York Academy of Medicine 83: I29-I38.

4. Zamani-Alavijeh F, Bazargan M, Shafiei A, Bazargan-Hejazi S (2011) The frequency and predictors of helmet use among Iranian motorcyclists: A quantitative and qualitative study. Accident Analysis and Prevention 43: 1562-1569.

5. Kadilar C, Cingi H (2004) Ratio estimators in simple random sampling. Applied Mathematics and Computation 151: 893-902.

6. Robinson J (1987) Conditioning ratio estimates under simple random sampling. Journal of the American Statistical Association 82: 826-831.

7. Sedransk J, Meyer J (1978) Confidence-intervals for quantiles of a finite population - Simple random and stratified simple random sampling. Journal of the Royal Statistical Society Series B-Methodological 40: 239-252.

8. Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47: 663-685.

9. Vitter JS (1985) Random sampling with a reservoir. ACM Transactions on Mathematical Software 11: 37-57.

10. Salant P, Dillman DA (1994) How to conduct your own survey. New York: John Wiley and Sons.

11. Cruzorive LM (1989) On the precision of systematic-sampling - A review of matherons transitive methods. Journal of Microscopy-Oxford 153: 315-333.

12. Gundersen HJG, Jensen EB (1987) The efficiency of systematic-sampling in stereology and its prediction. Journal of Microscopy-Oxford 147: 229-263.

13. Madow WG (1949) On the theory of systematic sampling .2. Annals of Mathematical Statistics 20: 333-354.

14. Van Spall HGC, Toren A, Kiss A, Fowler RA (2007) Eligibility criteria of randomized controlled trials published in high-impact general medical journals - A systematic sampling review. Jama-Journal of the American Medical Association 297: 1233-1240.

15. Danz NP, Regal RR, Niemi GJ, Brady VJ, Hollenhorst T, et al. (2005) Environmentally stratified sampling design for the development of great lakes environmental indicators. Environmental Monitoring and Assessment 102: 41-65.

16. Imbens GW, Lancaster T (1996) Efficient estimation and stratified sampling. Journal of Econometrics 74: 289-318.

17. Henderson RH, Sundaresan T (1982) Cluster sampling to assess immunization coverage - A review of experience with a simplified sampling method. Bulletin of the World Health Organization 60: 253-260.

18. Hughes G, Madden LV, Munkvold GP (1996) Cluster sampling for disease incidence data. Phytopathology 86: 132-137.

19. Thompson SK (1990) Adaptive cluster sampling. Journal of the American Statistical Association 85: 1050-1059.

20. Rocco E (2003) Constrained inverse adaptive cluster sampling. Journal of Official Statistics 19: 45-58.

21. Scheaffer RL, Mendenhall III W, Ott RL, Gerow K (2011) Elementary survey sampling: Duxbury Press.

22. Non-probability sampling Available: http://dissertation.laerd.com/non-probability-sampling.php. Accessed Jan 11, 2013.

23. Hedt BL, Pagano M (2011) Health indicators: Eliminating bias from convenience sampling estimators. Statistics in Medicine 30: 560-568.

24. Johnston LG, Trummal A, Lohmus L, Ravalepik A (2009) Efficacy of convenience sampling through the Internet versus respondent driven sampling among males who have sex with males in Tallinn and Harju County, Estonia: challenges reaching a hidden population. Aids Care-Psychological and Socio-Medical Aspects of Aids/Hiv 21: 1195-1202.

25. Ozdemir RS, St Louis KO, Topbas S (2011) Public attitudes toward stuttering in Turkey: Probability versus convenience sampling. Journal of Fluency Disorders 36: 262-267.

26. 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.

27. Auerswald CL, Greene K, Minnis A, Doherty I, Ellen J, et al. (2004) Qualitative assessment of venues for purposive sampling of hard-to-reach youth - An illustration in a Latino community. Sexually Transmitted Diseases 31: 133-138.

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

29. Babbie ER (2011) The practice of social research: Wadsworth Publishing Co Inc.

30. Patton MQ (1990) Qualitative evaluation and research methods. Thousand Oaks, CA: Sage Publications, Inc.

31. Henshaw SK, Martire G (1982) Abortion and the public-opinion polls. Family Planning Perspectives 14: 53-60.

32. Kaplowitz SA, Fink EL, Dalessio D, Armstrong GB (1983) Anonymity, strength of attitude, and the influence of public-opinion polls. Human Communication Research 10: 5-25.

33. Welch RL (2002) Polls, polls, and more polls - An evaluation of how public opinion polls are reported in newspapers. Harvard International Journal of Press-Politics 7: 102-114.

34. Ashing KT, Padilla G, Tejero J, Kagawa-Singer M (2003) Understanding the breast cancer experience of Asian American women. Psycho-Oncology 12: 38-58.

35. Grimshaw JM, Thomas RE, MacLennan G, Fraser C, Ramsay CR, et al. (2004) Effectiveness and efficiency of guideline dissemination and implementation strategies. Health Technology Assessment 8: 1-72.

36. Hussey S, Hoddinott P, Wilson P, Dowell J, Barbour R (2004) Sickness certification system in the United Kingdom: qualitative study of views of general practitioners in Scotland. British Medical Journal 328: 88-91.

37. Sullivan M, Kone A, Senturia KD, Chrisman NJ, Ciske SJ, et al. (2001) Researcher and researched-community perspectives: Toward bridging the gap. Health Education & Behavior 28: 130-149.

38. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences 109: 11576-11581.

39. Kenett DY, Portugali J (2012) Population movement under extreme events. Proceedings of the National Academy of Sciences 109: 11472-11473.

40. Cochran WG (1963) Sampling techniques. New York: John Wiley & Sons, Inc.

41. Moser CA, Stuart A (1953) An experimental study of quota sampling. Journal of the Royal Statistical Society Series A-Statistics in Society 116: 349-405.

42. Owen L, McNeill A, Callum C (1998) Trends in smoking during pregnancy in England, 1992-7: quota sampling surveys. British Medical Journal 317: 728-728.

43. Serfling RE, Cornell RG, Sherman IL (1960) The CDC quota sampling technique with results of 1959 poliomyelitis vaccination surveys. American Journal of Public Health and the Nations Health 50: 1847-1857.

44. Steinber.Ha (1963) Generalized quota sampling. Nuclear Science and Engineering 15: 142-145.

45. Goodman LA (1961) Snowball sampling. The Annals of Mathematical Statistics 32: 148-170.

46. Erickson BH (1979) Some problems of inference from chain data. Sociological Methodology 10: 276-302.

47. Heckathorn DD (2011) Snowball Versus Respondent-Driven Sampling. Sociological Methodology 41: 355-366.

48. Heckathorn DD (2002) Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. Social Problems 49: 11-34.

49. Erickson BH (1979) Some Problems of Inference from Chain Data. Sociological Methodology 10: 276-302.

50. Frank O, Snijders T (1994) Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics 10: 53-53.

51. Watters JK, Biernacki P (1989) Targeted sampling: options for the study of hidden populations. Social Problems: 416-430.

52. H. Fisher Raymond, Theresa Ick, Michael Grasso, Jason Vaudrey, Willi McFarland (2007) Resource Guide: Time Location Sampling (TLS). San Francisco Department of Public Health HIV Epidemiology Section, Behavioral Surveillance Unit.

53. Carlson RG, Wang JC, Siegal HA, Falck RS, Guo J (1994) An Ethnographic Approach to Targeted Sampling - Problems and Solutions in Aids-Prevention Research among Injection-Drug and Crack-Cocaine Users. Human Organization 53: 279-286.

54. Karon JM, Wejnert C (2012) Statistical methods for the analysis of time-location sampling data. Journal of Urban Health-Bulletin of the New York Academy of Medicine 89: 565-586.

55. World Health Organization, UNAIDS (2011) Guidelines on surveillance among populations most at risk for HIV. Geneva: WHO.

56. MarpsatA M, RazafindratsimaB N (2010) Survey methods for hard-to-reach populations: introduction to the special issue. Methodological Innovations Online 5: 3-16.

57. Mazzocchi M (2008) Statistics for marketing and consumer research: Sage Publications Limited.

58. Black TR (1999) Doing quantitative research in the social sciences: An integrated approach to research design, measurement, and statistics. Thousand Oaks, CA: SAGE Publications, Inc. .

59. CHGA [Commission in HIV/AIDS and Governance in Africa] (2004) Globalised Inequalities and HIV/AIDS.

60. UNAIDS (2010) UNAIDS report on the global AIDS epidemic 2010. Geneva.

61. UNAIDS (2011) UNAIDS World AIDS Day Report. Geneva.

62. The World Bank. Available: http://data.worldbank.org/indicator/SP.POP.TOTL.

63. Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, et al. (2012) Global epidemiology of HIV infection in men who have sex with men. The Lancet 380: 367-377.

64. Available: http://prostitution.procon.org/view.resource.php?resourceID=000772. Accessed Jan 11, 2013.

65. Countries where sex work ('prostitution') is deemed illegal Available: http://hivand-thelaw.com/countries-where-sex-work-prostitution-dee-med-illegal. Accessed Jan 11, 2013.

66. AIDS Infonet. How Risky is It? Available: http://www.aidsinfonet.org/fact_sheets/-view/152#what_s_my_risk_of_getting_infected_with_hiv. Accessed December 7, 2012.

67. Pisani E (2008) The wisdom of whores: bureaucrats, brothels, and the business of AIDS: Granta London.

68. Malani PN (2010) Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases. JAMA: The Journal of the American Medical Association 304: 2067-2068.

69. Mandell GL, Douglas Jr RG, Bennett JE (1979) Principles and practice of infectious diseases. Volumes 1 and 2: John Wiley & Sons.

70. Basavapathruni A, Anderson KS (2007) Reverse transcription of the HIV-1 pandemic. The FASEB Journal 21: 3795-3808.

71. US Census Bureau. Available: http://www.census.gov/newsroom/releases/-archives/2010_census/cb11-cn181.html. Accessed December 7, 2012.

72. Gates GJ (2012) Same-sex couples in Census 2010: Race and Ethnicity. The Williams Institute.

73. Schwarte N, Cohen R, Ben-Avraham D, Barabási A-L, Havlin S (2002) Percolation in directed scale-free networks. Physical Review E 66: 015104.

74. Newman MEJ (2003) The structure and function of complex networks. SIAM Review 45: 167-256.

75. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Reviews of Modern Physics 74: 47-97.

76. Albert R, Jeong H, Barabási A-L (1999) Diameter of the world-wide web. Nature 401: 130-131

77. Barabási A-L (2009) Scale-free networks: a decade and beyond. Science 325: 412-413.

78. Holme P, Saramaki J (2012) Temporal networks. Physics Reports-Review Section of Physics Letters 519: 97-125.

79. Lu X, Camitz M (2011) Finding the shortest paths by node combination. Applied Mathematics and Computation 217: 6401-6408.

80. Milgram S (1967) The small world problem. Psychology Today 2: 60-67.

81. Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32: 425-433.

82. Strogatz S, Watts DJ, Barabási A-L (2009) Unfolding the science behind the idea of six degrees of separation. Explaining Synchronicity, Network Theory, Adaption of Complex Systems, Six Degrees, Small World Phenomenon in the BBC Documentary: BBC.

83. Schnettler S (2009) A small world on feet of clay? A comparison of empirical small-world studies against best-practice criteria. Social Networks 31: 179-189.

84. Schnettler S (2009) A structured overview of 50 years of small-world research. Social Networks 31: 165-178.

85. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99: 7821-7826.

86. Saramaki J, Kivela M, Onnela J-P, Kaski K, Kertesz J (2007) Generalizations of the clustering coefficient to weighted complex networks. Physical Review E 75.

87. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Physical Review E 69.

88. Rosvallt M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105: 1118-1123.

89. Castellano C, Cecconi F, Loreto V, Parisi D, Radicchi F (2004) Self-contained algorithms to detect communities in networks. European Physical Journal B 38: 311-319.

90. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proceedings of the National Academy of Sciences 101: 2658-2663.

91. Porter MA, Onnela JP, Mucha PJ (2009) Communities in networks. Notices of the AMS 56: 1082-1097.

92. Vincent DB, Jean-Loup G, Renaud L, Etienne L (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008: P10008.

93. Newman MEJ (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103: 8577-8582.

94. Catania JA, Coates TJ, Kegeles S, Fullilove MT, Peterson J, et al. (1992) Condom use in multi-ethnic neighborhoods of San Francisco: the population-based AMEN (AIDS in Multi-Ethnic Neighborhoods) Study. American Journal of Public Health 82: 284-287.

95. Morris M (1995) Data driven network models for the spread of infectious disease. In: Mollison D, editor. Epidemic Models: Their Structure and Relation to Data. Cambridge Cambridge University Press. pp. 302-322.

96. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annual Review of Sociology 27: 415-444.

97. Morris M, Kretzschmar M (1995) Concurrent Partnerships and Transmission Dynamic in Networks. Social Networks 17: 299-318.

98. Rapoport A (1980) A Probabilistic Approach to Networks. Social Networks 2: 1-18.

99. Lu X, Malmros J, Liljeros F, Britton T (2013) Respondent-driven Sampling on Directed Networks. Electronic Journal of Statistics 7: 292-322.

100. Newman MEJ (2002) Assortative Mixing in Networks. Physical Review Letters 89: 208701.

101. Newman MEJ (2003) Mixing patterns in networks. Physical Review E 67: 026126.

102. Foster JG, Foster DV, Grassberger P, Paczuski M (2010) Edge direction and the structure of networks. Proceedings of the National Academy of Sciences 107: 10815-10820.

103. Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5: 17-61.

104. Erdős P, Rényi A (1959) On random graphs I. Publ Math Debrecen 6: 290-297.

105. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440-442.

106. Watts DJ (1999) Small Worlds. Princeton: Princeton University Press.

107. Barrat A, Weigt M (2000) On the properties of small world networks. European Physical Journal B 13: 547-560.

108. Monasson R (1999) Diffusion, localization and dispersion relations on “small-world” lattices. The European Physical Journal B-Condensed Matter and Complex Systems 12: 555-567.

109. Newman MEJ, Watts DJ (1999) Renormalization group analysis of the small-world network model. Physics Letters A 263: 341-346.

110. Newman M, Barabasi AL, Watts DJ (2011) The structure and dynamics of networks: Princeton University Press.

111. Barabási AL, Frangos J (2002) Linked: The New Science Of Networks Science Of Networks: Basic Books.

112. Liljeros F, Edling CR, Amaral LAN, Stanley HE, Aberg Y (2001) The web of human sexual contacts. Nature 411: 907-908.

113. Onnela JP, Saramaki J, Hyvonen J, Szabo G, Lazer D, et al. (2007) Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104: 7332-7336.

114. Price DJdS, Amer. J (1976) A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27: 292-306.

115. Price DJdS (1965) Networks of scientific papers. Science 149: 510-515.

116. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512.

117. Barabasi AL, Albert R, Jeong H (1999) Mean-field theory for scale-free random networks. Physica A 272: 173-187.

118. Gallos LK, Cohen R, Argyrakis P, Bunde A, Havlin S (2005) Stability and topology of scale-free networks under attack and defense srategies. Physical Review Letters 94: 188701.

119. Gallos LK, Argyrakis P (2007) Scale-free networks resistant to intentional attacks. Europhysics Letters 80: 58002.

120. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406: 378-382.

121. Holme P, Kim BJ, Yoon CN, Han SK (2002) Attack vulnerability of complex networks. Physical Review E 65: 056109.

122. Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. Physical Review E 63: 066117.

123. Pastor-Satorras R, Vespignani A (2002) Epidemic dynamics in finite size scale-free networks. Physical Review E 65: 035108.

124. Newman MEJ, Park J (2003) Why social networks are different from other types of networks. Physical Review E 68.

125. Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks. Physical Review Letters 88: 128701.

126. Marsili M, Vega-Redondo F, Slanina F (2004) The rise and fall of a networked society: A formal model. Proceedings of the National Academy of Sciences 101: 1439-1442.

127. Kumpula JM, Onnela JP, Saramaki J, Kaski K, Kertesz J (2007) Emergence of communities in weighted networks. Physical Review Letters 99.

128. Toivonen R, Onnela J-P, Saramäki J, Hyvönen J, Kaski K (2006) A model for social networks. Physica A: Statistical Mechanics and its Applications 371: 851-860.

129. Vázquez A (2003) Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. Physical Review E 67: 056104.

130. Boguñá M, Pastor-Satorras R, Díaz-Guilera A, Arenas A (2004) Models of social networks based on social distance attachment. Physical Review E 70: 056122.

131. Wong LH, Pattison P, Robins G (2006) A spatial model for social networks. Physica A: Statistical Mechanics and its Applications 360: 99-120.

132. Snijders TAB, Pattison PE, Robins GL, Handcock MS (2006) New specifications for exponential random graph models. Sociological Methodology 36: 99-153.

133. Toivonen R, Kovanen L, Kivela M, Onnela JP, Saramaki J, et al. (2009) A comparative study of social network models: Network evolution models and nodal attribute models. Social Networks 31: 240-254.

134. Heckathorn DD (1997) Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems 44: 174-199.

135. Wattana W, van Griensven F, Rhucharoenpornpanich O, Manopaiboon C, Thienkrua W, et al. (2007) Respondent-driven sampling to assess characteristics and estimate the number of injection drug users in Bangkok, Thailand. Drug and Alcohol Dependence 90: 228-233.

136. Malekinejad M, Johnston LG, Kendall C, Kerr L, Rifkin MR, et al. (2008) Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. Aids and Behavior 12: S105-S130.

137. 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.

138. Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology, Vol 34. Weinheim: Wiley-V C H Verlag Gmbh. pp. 193-239.

139. Volz E, Heckathorn DD (2008) Probability Based Estimation Theory for Respondent Driven Sampling. Journal of Official Statistics 24: 79-97.

140. Hansen MH, Hurwitz WN (1943) On the Theory of Sampling from Finite Populations. The Annals of Mathematical Statistics 14: 333-362.

141. Sirken M (2001) The Hansen-Hurwitz estimator revisited: PPS sampling without replacement. ASA Proceedings of the Survey Methods Section.

142. Cochran WG (1977) Sampling techniques. New York: John Wiley & Sons, Inc.

143. Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology, 2004, Vol 34 34: 193-239.

144. Feld SL (1991) Why Your Friends Have More Friends Than You Do. American Journal of Sociology 96: 1464-1477.

145. Heckathorn DD (2007) Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. Sociological Methodology 2007, Vol 37. Oxford: Blackwell Publishing. pp. 151-208.

146. Goel S, Salganik MJ (2010) Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences 107: 6743-6747.

147. 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.

148. Lu X (2012) Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-driven Sampling. Arxiv preprint arXiv:1205.1971. http://arxiv.org/abs/1205.1971v2.

149. 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.

150. Gallagher KM, Sullivan PS, Lansky A, Onorato IM (2007) Behavioral surveillance among people at risk for HIV infection in the U.S.: the National HIV Behavioral Surveillance System. Public Health Reports 122 Suppl 1: 32-38.

151. Kral AH, Malekinejad M, Vaudrey J, Martinez AN, Lorvick J, et al. (2010) Comparing respondent-driven sampling and targeted sampling methods of recruiting injection drug users in San Francisco. Journal of Urban Health-Bulletin of the New York Academy of Medicine 87: 839-850.

152. Available: http://www.respondentdrivensampling.org/.

153. Heckathorn DD, Jeffri J (2001) Finding the beat: Using respondent-driven sampling to study jazz musicians. Poetics 28: 307-329.

154. Heckathorn DD, Jeffri J (2003) Social Networks of Jazz Musicians. Changing the Beat: A Study of the Worklife of Jazz Musicians, Volume III: Respondent-Driven Sampling: Survey Results by the Research Center for Arts and Culture. Washington DC: National Endowment for the Arts Research Division pp. 48-61.

155. Jeffri J, Heckathorn DD, Spiller MW (2011) Painting your life: A study of aging visual artists in New York City. Poetics 39: 19-43.

156. Calafat A, Blay N, Juan M, Adrover D, Bellis MA, et al. (2009) Traffic Risk Behaviors at Nightlife: Drinking, Taking Drugs, Driving, and Use of Public Transport by Young People. Traffic Injury Prevention 10: 162-169.

157. Far AC, Roig DA, Jerez MJ, Franzke NTB (2008) Relationship between alcohol, drug use and traffic accidents related to nightlife among a Spanish youth sample in three Autonomous Communities in 2007. Rev Esp Salud Publica 82: 323-331.

158. Kogan SM, Brody GH (2010) Linking Parenting and Informal Mentor Processes to Depressive Symptoms Among Rural African American Young Adult Men. Cultural Diversity & Ethnic Minority Psychology 16: 299-306.

159. 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. Journal of Studies on Alcohol and Drugs 73: 834-838.

160. Gwadz MV, Cleland CM, Quiles R, Nish D, Welch J, et al. (2010) CDC HIV testing guidelines and the rapid and conventional testing practices of homeless youth. Aids Education and Prevention 22: 312-327.

161. Wejnert C (2010) Social network analysis with respondent-driven sampling data: A study of racial integration on campus. Social Networks 32: 112-124.

162. Qiu PY, Caine E, Yang Y, Chen Q, Li J, et al. (2011) Depression and associated factors in internal migrant workers in China. Journal of Affective Disorders 134: 198-207.

163. Jung M (2012) Immigrant Workers' Knowledge of HIV/AIDS and Their Sexual Risk Behaviors: A Respondent-Driven Sampling Survey in South Korea. Sexuality and Disability 30: 199-208.

164. Parker P, HALPIN B (2011) Eastern European and Baltic Migrant Workers'labour Market Performances in the Republic of Ireland. Research West Review 1.

165. Burnham G, Malik S, Al-Shibli ASD, Mahjoub AR, Baqer AQ, et al. (2012) Understanding the impact of conflict on health services in Iraq: information from 401 Iraqi refugee doctors in Jordan. International Journal of Health Planning and Management 27: e51-e64.

166. Castro MC, Yanez SY (2012) New forms of sampling for minority and hidden populations: respondent samples conducted in a south american immigrant population. Universitas Psychologica 11: 571-578.

167. Song EY, Leichliter JS, Bloom FR, Vissman AT, O'Brien MC, et al. (2012) The Use of Prescription Medications Obtained from Non-medical Sources among Immigrant Latinos in the Rural Southeastern U.S. J Health Care Poor Underserved 23: 678-693.

168. 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.

169. Wallace WL (1966) Student culture: Social structure and continuity in a liberal arts college: Aldine Publishing Company.

170. Feld SL, Carter WC (2002) Detecting measurement bias in respondent reports of personal networks. Social Networks 24: 365-383.

171. South SJ, Haynie DL (2004) Friendship networks of mobile adolescents. Social Forces 83: 315-350.

172. Marsden PV (2005) Recent Developments in Network Measurement. In: Carrington PJ, John Scott, Wasserman S, editors. Models and Methods in Social Network Analysis. New York: Cambridge University Press. pp. 8-30.

173. Frost SDW, Brouwer KC, Cruz MAF, Ramos R, Ramos ME, et al. (2006) Respondent-driven sampling of injection drug users in two US-Mexico border cities: Recruitment dynamics and impact on estimates of HIV and syphilis prevalence. Journal of Urban Health-Bulletin of the New York Academy of Medicine 83: I83-I97.

174. Wang JC, Carlson RG, Falck RS, Siegal HA, Rahman A, et al. (2005) Respondent-driven sampling to recruit MDMA users: a methodological assessment. Drug and Alcohol Dependence 78: 147-157.

175. Gile KJ, Handcock MS (2010) Respondent-driven sampling: An assessment of current methodology. Sociological Methodology: no. doi: 10.1111/j.1467-9531.2010.01223.x.

176. McCreesh N, Johnston LG, Copas A, Sonnenberg P, Seeley J, et al. (2011) Evaluation of the role of location and distance in recruitment in respondent-driven sampling. Int J Health Geogr 10: 56.

177. McCreesh N, Frost SD, Seeley J, Katongole J, Tarsh MN, et al. (2012) Evaluation of respondent-driven sampling. Epidemiology 23: 138-147.

178. 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.

179. Volz E, Wejnert C, Degani I, Heckathorn DD (2007) Respondent-driven sampling analysis tool (RDSAT) Version 5.6. Ithaca, NY: Cornell University.

180. Tomas A, Gile KJ (2011) The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. Electronic Journal of Statistics 5: 899-934.

181. Gile KJ (2011) Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. Journal of the American Statistical Association 106: 135-146.

182. Gile KJ, Handcock MS (2011) Network model-assisted inference from respondent-driven sampling data. arXiv preprint arXiv:11080298 http://arxivorg/abs/11080298.

183. Poon AF, Brouwer KC, Strathdee SA, Firestone-Cruz M, Lozada RM, et al. (2009) Parsing social network survey data from hidden populations using stochastic context-free grammars. PLoS ONE 4: e6777.

184. Handcock MS, Gile KJ, Mar CM (2012) Estimating Hidden Population Size using Respondent-Driven Sampling Data. arXiv preprint arXiv:1209.6241. http://arxiv.org/abs/1209.6241.

185. Rybski D, Buldyrev SV, Havlin S, Liljeros F, Makse HA (2009) Scaling laws of human interaction activity. Proceedings of the National Academy of Sciences 106: 12640-12645.

186. Brem SK (2002) Analyzing Online Discussions: Ethics, Data, and Interpretation. Practical Assessment, Research & Evaluation 8: n3.

187. Ess C (2007) Internet research ethics. In: Joinson. A, McKenna. K, Postmes. T, Reips. U-D, editors. The Oxford handbook of Internet psychology. Oxford: Oxford University Press. pp. 487-502.

188. 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: e49417.

189. 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.

190. U.S. Census Bureau’s International Database (2012). http://www.census.gov/-population/international/data/idb/country.php.

191. Nguyen QC (2010) Sexual risk behaviors of men who have sex with men in Viet Nam. Chapel Hill: North Carolina State University.

192. 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.

193. World Bank (2010) Open data: Urban Population. http://data.worldbank.org/-indicator/SP.URB.TOTL.

194. 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.

195. Fortunato S, Boguñá M, Flammini A, Menczer F (2008) Approximating PageRank from in-degree. Algorithms and Models for the Web-Graph: 59-71.

196. Vu ML, Le Thi MP, Nguyen TV, Doan KT, Tran QT (2009) MSM in Vietnam: social stigma and consequences. Hanoi, Vietnam. http://www.unaids.org.vn/-index.php?option=com_content&view=article&id=561%3A19-july-2011-msm-

in-vietnam-social-stigma-and-consequences&catid=92%3Atwg-lunchtime-seminars&Itemid=109&lang=en.

197. 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 22: 784-792.

198. Mello Md, Pinho AdA, Chinaglia M, Tun W, Júnior AB, et al. (2008) Assessment of risk factors for HIV infection among men who have sex with men in the metropolitan area of Campinas city, Brazil, using respondent-driven sampling. Washington DC: Population Council.

199. Li J, Liu H, Luo J, Koram N, Detels R (2011) Sexual transmissibility of HIV among opiate users with concurrent sexual partnerships: an egocentric network study in Yunnan, China. Addiction 106: 1780-1787; discussion 1788-1789.

200. Rudolph AE, Latkin C, Crawford ND, Jones KC, Fuller CM (2011) Does respondent driven sampling alter the social network composition and health-seeking behaviors of illicit drug users followed prospectively? PLoS ONE 6: e19615.

201. TREATASIA, amfAR (2006) MSM and HIV/AIDS Risk in Asia: What Is Fueling the Epidemic Among MSM and How Can It Be Stopped? www.amfar.org.

202. McCarty C, Killworth PD, Bernard HR, Johnsen EC, Shelley GA (2001) Comparing two methods for estimating network size. Human Organization 60: 28-39.

203. Shrestha S, Smith MW, Broman KW, Farzadegan H, Vlahov D, et al. (2006) Multiperson use of syringes among injection drug users in a needle exchange program: A gene-based molecular epidemiologic analysis. Jaids-Journal of Acquired Immune Deficiency Syndromes 43: 335-343.

204. Valente TW, Foreman RK, Junge B, Vlahov D (1999) Satellite exchange in the Baltimore Needle Exchange Program. Public Health 113: 90-96.

205. Adams J, Moody J (2007) To tell the truth: Measuring concordance in multiply reported network data. Social Networks 29: 44-58.

206. Morris M, Rothenberg R (2011) HIV Transmission Network Metastudy Project: An Archive of Data From Eight Network Studies, 1988-2001. Inter-university Consortium for Political and Social Research (ICPSR).

207. Kogovšek T, Ferligoj A (2005) The quality of measurement of personal support subnetworks. Quality & quantity 38: 517-532.

208. Matzat U, Snijders C (2010) Does the online collection of ego-centered network data reduce data quality? An experimental comparison. Social Networks 32: 105-111.

209. Burt RS (1984) Network items and the general social survey. Social Networks 6: 293-339.

210. Marin A (2004) Are respondents more likely to list alters with certain characteristics?: Implications for name generator data. Social Networks 26: 289-307.

211. Salganik MJ (2012) Commentary: Respondent-driven Sampling in the Real World. Epidemiology 23: 148-150.

212. White RG, Lansky A, Goel S, Wilson D, Hladik W, et al. (2012) Respondent driven sampling—where we are and where should we be going? Sexually Transmitted Infections 88: 397-399.

213. Gile KJ, Johnston LG, Salganik MJ (2012) Diagnostics for Respondent-driven Sampling. arXiv preprint arXiv:12096254.

214. Paquette DM, Bryant J, De Wit J (2011) Use of respondent-driven sampling to enhance understanding of injecting networks: a study of people who inject drugs in Sydney, Australia. International Journal of Drug Policy 22: 267-273.

215. Abramovitz D, Volz EM, Strathdee SA, Patterson TL, Vera A, et al. (2009) Using respondent-driven sampling in a hidden population at risk of HIV infection: who do HIV-positive recruiters recruit? Sexually Transmitted Diseases 36: 750-756.

216. Kogan SM, Wejnert C, Chen Y-f, Brody GH, Slater LM (2011) Respondent-Driven Sampling With Hard-to-Reach Emerging Adults: An Introduction and Case Study With Rural African Americans. Journal of Adolescent Research 26: 30-60.

APPENDIX

Web-RDS Questionnaire for the study of MSM in Vietnam

Question Number	Question English	Response Alternatives English	Question Vietnamese	Response Alternatives Vietnamese
Q1	Do you have a boyfriend now? (“boyfriend” here indicates a stable relationship with emotional attachment)	Yes	Hiện tại bạn có người yêu là nam giới không?	Có
		No		Không
		Don’t want to answer		Không muốn trả lời
Q2	What is the longest time that you have maintained a ("love") relationship with a man?	Never had a ("love") relationship	Mối quan hệ tình cảm dài nhất với một người đàn ông mà bạn từng có từ trước đến nay kéo dài trong bao lâu?	Chưa bao giờ
		Less than 1 month		Dưới 1 tháng
		1 to 6 months		Từ 1 đến 6 tháng
		6 months to 1 year		Từ 6 tháng đến 1 năm
		1 to 3 years		Từ 1 đến 3 năm
		More than 3 years		Hơn 3 năm
		Don’t want to answer		Không muốn trả lời
Q3	If you meet someone for sex, what characteristic of that person is most important to you?	Humorous Rich	Nếu bạn gặp một người đàn ông chỉ để quan hệ tình dục, bạn coi đặc điểm gì của người ấy là quan trọng nhất?	Hài hước
		Intelligent		Thông minh
		Good looking		Đẹp trai
		Understanding		Hiểu tôi
		Rich		Giàu có
		Respect me		Tôn trọng tôi
		Good at sex		Làm tình giỏi
		Don’t care/Don’t want to answer		Không quan tâm/ không muốn trả lời
Q4	If you are looking for a man for a long-term relationship, what characteristic of that person is most important to you?	Humorous	Nếu bạn tìm kiếm một người đàn ông làm người yêu lâu dài, bạn coi đặc điểm gì của người ấy là quan trọng nhất?	Hài hước
		Intelligent		Thông minh
		Good looking		Đẹp trai
		Understanding		Hiểu tôi
		Rich		Giàu có
		Faithful		Chung thủy
		Respect me		Tôn trọng tôi
		Good at sex		Làm tình giỏi
		Don’t care/Don’t want to answer		Không quan tâm/ không muốn trả lời
Q5	Do you think same sex marriage should be permitted in Vietnam?	Yes	Bạn có nghĩ luật pháp Việt Nam nên cho phép kết hôn đồng giới không?	Có
		No		Không
		No opinion/ Don’t want to answer		Không có ý kiến/ không muốn trả lời
Q6	During the last 6 months, how many men have you had sex with (anal, oral or masturbation)?	Text	Trong 06 tháng qua, bạn đã quan hệ tình dục (đường miệng, đường hậu môn hoặc thủ dâm cho nhau) với bao nhiêu nam giới? (Nếu không nhớ chính xác, hãy đưa ra một con số ước lượng gần nhất)	Text
Q7	Which of the following statements best describe your preferences?	Prefer only men as sexual partners	Trong những câu sau, câu nào miêu tả sở thích chọn bạn tình của bạn chính xác nhất?	Chỉ thích bạn tình là nam giới
		Prefer men to women as sexual partners		Thích bạn tình là nam giới hơn nữ giới
		Prefer women to men as sexual partners		Thích bạn tình là nữ giới hơn nam giới
		Prefer only women as sexual partners		Chỉ thích bạn tình là nữ giới
		Don’t want to answer		Không muốn trả lời
Q8	In the last 6 months, have you ever had sex in a public place like a sauna, gym, swimming pool, public toilet or park?	Yes	Trong vòng 06 tháng qua, bạn có quan hệ tình dục ở những chỗ công cộng như phòng tắm hơi, phòng tập thể hình, bể bơi, nhà vệ sinh công cộng hay công viên không?	Có
		No		Không
		Don't want to answer		Không nhớ/ không muốn trả lời
Q9	In what year were you born?	Don't want to answer	Bạn sinh năm nào?	Tôi không muốn trả lời
		1999		1999
		…		…
		1940		1940

Q10	What is your highest level of education? Select only one option.	No schooling	Trình độ học vấn cao nhất của bạn?	Chưa bao giờ đi học
		Primary (Grade 1-5)		Tiểu học (Lớp 1-5)
		Secondary school (Grade 6-9)		Cấp 2 (Lớp 6-9)
		High school (Grade 10 – 12)		Cấp 3 (Lớp 10 – 12)
		University, college or vocational training		Đại học hoặc cao đăng, dạy nghề
		Postgraduate		Sau đại học
		Don't want to answer		Không muốn trả lời
Q11	During the last 12 months, what was the average amount of money you received per each month, from all sources. (include money from salary, parents, interests and all other sources)	<1,000,000 VND	Trong 12 tháng vừa qua, số tiền thu nhập trung bình mỗi tháng từ tất cả các nguồn (ví dụ: lương, tiền bố mẹ cho, từ bạn bè, tiền lãi trong kinh doanh...) của bạn là bao nhiêu?	Dưới 1,000.000 đồng
		1,000,000 – 3,000,000 VND		Từ 1.000.000 Đồng đến dưới 3.000.000 Đồng
		3,000,000 – under 5,000,000 VND		Từ 3.000.000 Đồng đến dưới 5.000.000 Đồng
		5,000,000 – under 10,000,000 VND		Từ 5.000.000 Đồng đến dưới 10.000.000 Đồng
		>10,000,000 VND		Trên 10.000.000 Đồng
		Don’t want to answer		Không nhớ/ không muốn trả lời
Q12	In which province are you living?	Don't want to answer/outside Vietnam	Bạn đang sống ở tỉnh/ thành phố nào?	Không muốn trả lời/ ở bên ngoài lãnh thổ Việt Nam
		Hồ Chí Minh		Hồ Chí Minh
		Hà Nội		Hà Nội
		Hải Phòng		Hải Phòng
		Khánh Hoà		Khánh Hoà
		Cần Thơ		Cần Thơ
		----------------		----------------
		An Giang		An Giang
		…		…
		Yên Bái		Yên Bái
Q13	During the last month, how many days did you use the Internet? If you do not remember exactly, please give your best guess.	Don’t want to answer	Bạn sử dụng Internet bao nhiêu ngày trong vòng 30 ngày vừa qua? (Nếu không nhớ chính xác, hãy đưa ra một con số ước lượng gần nhất)	Không nhớ / không muốn trả lời
		1 or less than 1 day per month		1 ngày hoặc dưới 1 ngày
		2 days		2 ngày
		…		…
		30 days		30 ngày
Q14	During the last 7 days, how many people in your world have you had any type of contact with (in person, on the phone, on chat, facebook, mail or in some other way)?If you do not remember exactly, please give your best guess.	Text	Trong vòng 7 ngày vừa qua, bạn nói chuyện với bao nhiêu người trong giới, dưới bất kỳ hình thức nào như nói chuyện trực tiếp, qua điện thoại, chat, facebook, email, gửi thư hoặc một cách nào khác?	Text
Q15	Out of these in question 14, how many use the internet?If you do not know exactly, please give your best guess.	Text	Trong số những người bạn đề cập ở câu 14 (những người bạn nói chuyện với dưới bất kể hình thức nào trong vòng 7 ngày vừa rồi), có bao nhiêu người từ 18 tuổi trở lên và sử dụng Internet? (Nếu không biết chính xác, hãy đưa ra một con số ước lượng gần nhất)	Text
Q16	What is your relationship to the person who invited you to this study? (You can choose a maximum of two alternatives)	He is a stranger (I have not communicated with him before I got this invitation)	Bạn có quan hệ như thế nào với người mời bạn tham gia nghiên cứu này? (Bạn có thể chọn tối đa 2 phương án trả lời)	Người ấy là người lạ (tôi không có liên lạc gì với người ấy trước khi tôi nhận được lời mời này)
		He is an acquaintance		Người ấy là người quen biết
		He is a friend		Người ấy là bạn
		He is a close friend		Người ấy là bạn thân
		He is a lover/ex lover		Người ấy là người yêu/ người yêu cũ
		Relative		Họ hàng
		Don’t want to answer		Không muốn trả lời

Q17	In what context did you get to know that person? How did you get to know this person? (Choose more than one alternative when appropriate)	Through an MSM web page, chat room, facebook or other Internet site	Bạn biết người ấy trong trường hợp nào? (Bạn có thể chọn nhiều phương án trả lời)	Thông qua các trang web giành cho người đồng tính, phòng chat, facebook hoặc các trang web khác trên internet
		Through people I know (friends, relatives, lovers etc)		Qua người quen (bạn bè, họ hàng, người yêu...)
		Through an MSM club		Qua câu lạc bộ MSM
		Through work		Qua công việc
		Through school, university or other type of education		Ở trường học hoặc các cơ sở đào tạo
		Through a leisure activity		Qua các trò vui chơi, giải trí
		At an MSM venue (bar, disco, sauna, park, street for MSM etc)		Tại các tụ điểm cho MSM (bar, sàn nhảy, phòng tắm hơi (sauna), công viên hay đường phố)
		At a non-MSM venue.		Các tụ điểm khác không dành riêng cho MSM
		Other		Khác
		Don’t remember/ Don't want to answer		Không nhớ/ không muốn trả lời