

Network Distribution and Respondent-Driven Sampling (RDS) Inference About People who inject drugs in Ottawa, Ontario

Kahina Abdesselam

**A thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial
fulfillment of the requirements for the PhD degree in Epidemiology**

School of Epidemiology and Public Health, Faculty of Medicine University of Ottawa

© Kahina Abdesselam, Ottawa, Canada, 2019

Contents

Contents	ii
List Tables.....	v
List of Figures	vi
Acknowledgments.....	vii
Abbreviations	viii
Abstract	ix
1.0 Introduction.....	1
1.1 Social Network Analysis	4
1.2 Network Topology/Structure.....	7
1.3 RDS Inference	9
1.4 RDS PWID data in Ottawa	14
1.5 References	16
2.0 Objectives and Hypothesis.....	22
3.0 Study Population.....	24
This secondary data that is being used for the next two chapters.....	24
3.1 Method of Recruitment	24
3.2 Data Collection.....	24
3.3 Serological Assays	25
3.4 Molecular Analysis	25
3.5 Ethics	26
4.0 Identifying a Heavy Tailed Distribution, Using Frequency of Injections, Among People Who Inject Drugs Network in Ottawa, Canada	27
4.1 Abstract	28
4.2 Introductions.....	29
4.3 Methods	34
4.3.1 Data collection.....	34
4.3.2 Data analysis.....	35
4.3.3 Ethics	37
4.4 Results	37
4.4.1 Descriptive Analysis.....	37

4.4.2 RDS Network	38
4.4.3 Network Topology.....	38
4.5 Discussion	45
4.6 Conclusion.....	47
4.7 References	49
5.0 Social Contacts of People who Inject Drugs in Ottawa Exhibit a Small-World Network	53
5.1 Abstract	54
5.2 Introduction	55
5.3 Data and Methods.....	59
5.4 Results	63
5.5 Discussion	67
5.6 Conclusion.....	72
5.7 References	73
6.0 The Development of Respondent-Driven Sampling (RDS) Inference: A Systematic Review of the Population Mean and Variance estimates.....	78
6.1 Abstract	80
6.2 Introduction	80
6.3 Methods.....	83
6.3.1 Eligibility Criteria.....	84
6.3.2 Information Sources	85
6.3.3 Search Strategy	86
6.3.4 Selection Process	86
6.3.5 Information Items	87
6.4 Results	87
6.4.1 RDS Estimator Developed.....	89
6.4.2 Evaluation of RDS Estimators.....	91
6.4.3 Developed Variance Estimates	96
6.4.4 Evaluated Variance Estimates	98
6.5 Discussion	100
6.6 Conclusion.....	106
6.7 References	108

7.0 Discussion	111
7.1 References	116
8.0 Conclusion	117
9.0 References	119
10.0 Appendix	131
10.1 Search Strategy for Systematic Review	131

List Tables

Table 1: Possible risk factors of injecting drugs associated with HCV status in the PWID network in Ottawa, Canada, 2007.....	39
Table 2: The HCV positivity and use of the needle exchange program (NEP) in the 6-month prior to interview.....	41
Table 3: Parameters of the various distributions used to model the frequency of injection in an Ottawa network consisting of 407 IDU's, Canada, 2007	44
Table 4: Fitted power law distribution to the frequency of injection in the 6-month period prior to interview among 407 Ottawa IDU's compared to other fitted distributions using the Vuong test, cut-off $p > 0.1$	45
Table 5: Demographics of the RDS participants and their alters.	64
Table 6: The relationship PWID and non-PWID alters identified from the interaction grid completed by the RDS participant ($p < 0.0001$)	66
Table 7: Network and Small-world-ness properties of the Ottawa IDU RDS data conducted in 2007.....	67
Table 8: Six critical assumptions associated with the RDS design which when satisfied, estimates from RDS inference are considered reliable ²	82
Table 9: PICOS for systematic review on studies focusing on improving the inferential aspect of RDS.....	84
Table 10: Keywords used for the literature search strategy for studies focusing on RDS inference	86
Table 11: Inclusion and exclusion criteria for eligibility assessment of literature sources.....	87
Table 12: Reasons for exclusion during second screening of articles (full-text review)	88
Table 13: Areas of the inference from 32 RDS articles retained for the systematic literature review.....	88
Table 14: RDS estimators developed since the introduction of RDS in 1997.	90
Table 15: Studies evaluating RDS estimators by year, RDS estimates and population network	91
Table 16: Outcome(s) of evaluations of the most commonly used RDS estimators from studies in Table 15.	94
Table 17: Current core RDS Variance estimators and their main methodological approaches...	97
Table 18: Summary of evaluated variance estimate (s) from 1997 to 2017	99

List of Figures

Figure 1: Summary of the traditional way of conducting RDS inference as well as our potential contribution to RDS inference in people who inject drugs (highlighted in red). Adapted from Saglanik & Heckathorn	22
Figure 2: Visualization of the 7 components (generated by the seeds) consisting of 407 IDU's identified by RDS in Ottawa, Canada, 2007; where square = male, circle = female, triangle = transgender, red = HCV+, blue = HIV+, purple = Co-infection, black = missing and white = not infected.....	42
Figure 3: Visualization of the third component of the network (n=12), where square = male, circle = female, red = HCV+, and white = not infected.....	43
Figure 4: Plotted data of the total frequency of injection in the 6-month period prior to interview among 407 IDUs, Ottawa, Canada	43
Figure 5: Log-log plot of the cumulative distribution for frequency of injection in the 6-month period prior to interview among 407 IDU's, Ottawa, Canada, 2007, fitted to several types of distributions including the power law (black line), lognormal (dashed blue line), Poisson (dotted green line) and exponential (dashed red line)	44
Figure 6: Examples of different types of networks using the Ottawa PWID RDS data with the same number of nodes and links (R Studio, v 1.1456)	58
Figure 7: Grid Interaction used in the RDS questionnaire to identify the degree of each participant, ego-network and the clustering coefficient.....	61
Figure 8: Plotted frequency of social/drug using interactions with alters reported by RDS Participants, Ottawa PWID study, 2007	65
Figure 9: HCV status based on the reported and calculated degrees of the RDS participants (n = 364)	66
Figure 10: PRISMA flow chart of study selection for RDS Inference Studies ⁸	88
Figure 11: Highlights of the thesis findings and its potential contribution and recommendations to RDS inference in people who inject drugs (highlighted in red). Adapted from Saglanik & Heckathorn.	118

Acknowledgments

Foremost I would like to thank my thesis supervisor, Dr. Ann M Jolly, for her continuous support, patience, mentoring, and valuable guidance throughout my PhD journey. Ann is someone you never forget as she is genuine, kind and a caring person. She provided me with constant encouragement and had unwavering belief in me. I could not have imagined having a better advisor and mentor for my PhD study.

I would also like to express my sincere gratitude to my thesis committee, Drs. Ashton Verdery, Franco Momoli, Mark Tyndall, George Wells and Linda Pelude, for their expertise, time, patience, understanding and encouragement. A special mention to Linda Pelude and Dr. Rasha Maal-Bared, who have been amazing mentors both in academic and professional settings, lent an ear to me, and provided me with endless supply of solid advice that kept me going through difficult times. I can truly say that I wouldn't be where I am without their continuous support and friendship.

A special thanks goes to my friends, academic peers and my colleagues from the Public Health Agency of Canada and Health Canada (too many to list by name here but you know who you are) for their encouragement, patience and support. In particular, I would like to thank Denise Gravel, Yogesh Choudhri, Katherine Dinner, Parminder Dhami, Jacob Miller, and Drs. Robert Nesdole, and Aboubakar Mounchili.

A good support system is important for surviving and remaining sane in graduate school. I was lucky to be a part of one we like to call "Team 7/TMNT" and refer to my best friends/siblings Dhara Vyas, Aghiles Abdesselam and Sabrine Abdesselam. Your emotional support and your encouragements have been essential. Thank you for sticking with me through the thick and thin – I couldn't have survived Azkaban (aka thesis writing) without you guys. I love you guys unconditionally. Thank you to N.Uzumaki for constant humour, my thesis is completed, believe it!

Last but certainly not least, I would like to thank my parents, in particular my lovely mother, Dr. Nacéra Mellal. She has been an unwavering source of strength to me, my confidant, my role model and biggest inspiration in pursuing and completing my doctorate; and whose love, guidance, generosity, support and encouragement are beyond description. This thesis is dedicated to my wonderful mother, I love you, always have, always will.

Abbreviations

BA Barabási Albert

CC Clustering Coefficient

CDF Cumulative Distribution Function

CI Confidence Interval

DE Design Effect

DN Direct Network

EIA Enzyme Immunoassay

ELISA Enzyme-Linked Immunoassay

ER Erdős Rényi

HCV Hepatitis C virus

HBV Hepatitis B virus

HIV Human immunodeficiency virus

Hom High order of Markov

FOM First Order Markov

FOMP First Order Markov Process

KS Kolmogorov-Smirnov

MA Model-assisted

MOD Modified

MP Markov Process

MSM Men who have sex with Men

PCR Polymerase chain reaction

PWID People who inject Drugs

RDS Respondent-driven Sampling

RT-PCR Reverse transcription PCR

SB Salganik Bootstrapping

SS Successive Sampling

SRS Simple Random Sampling

STBBI Sexually transmitted and blood
borne infections

STI Sexually transmitted infections

VHE Volz and Heckathorn

Wbc Without branching

WS Watts-Strogatz

Abstract

Respondent-driven sampling (RDS) is very useful in collecting data from individuals in hidden populations, where a sampling frame does not exist. It starts with researchers choosing initial respondents from a group which may be involved in taboo or illegal activities, after which they recruit other peers who belong to the same group. Analysis results in unbiased estimates of population proportions though with strong assumptions about the underlying social network and RDS recruitment process. These assumptions bear little resemblance to reality, and thus compromise the estimation of any means, population proportions or variances inferred from studies. The topology of the contact network, denoted by the number of links each person has, provides insight into the processes of infectious disease spread. The overall objective of the thesis is to identify the topology of an injection drug use network, and critically review the methods developed to produce estimates. The topology of people who inject drugs (PWID) collected by RDS in Ottawa, 2006 was compared with a Poisson distribution, an exponential distribution, a power-law distribution, and a lognormal distribution. The contact distribution was then evaluated against a small-world network characterized by high clustering and low average distances between individuals. Last a systematic review of the methods used to produce RDS mean and variance estimates was conducted. The Poisson distribution, a type of random distribution, was not an appropriate fit for PWID network. However, the PWID network can be classified as a small world network organised with many connections and short distances between people. Prevention of transmission in such networks should be focussed on the most active people (clustered individuals and hubs) as intervention with any others is less effective. The systematic review contained 32 articles which included the development and evaluation of 12 RDS mean and 6 variance estimators. Overall, the majority of estimators perform roughly the

same, with the exception of RDS^{IEGO} which outperformed the 6 other RDS mean estimators.

The Tree bootstrap variance estimate does not rely on modelling RDS as a first order Markov (FOM) process, which seems to be the main limitation of the other existing estimators. The lack of FOM as an assumption and the flexible application of this variance estimator to any RDS point estimate make the Tree bootstrapping estimator a more efficient choice.

1.0 Introduction

Despite the progress in the understanding, detection, and therapy of Hepatitis C (HCV) and HIV, reported cases have increased over the last 5 years worldwide. In North America, a disproportionate number of cases are seen in sex workers, street youth, people who inject drugs (PWIDs) and men who have sex with men (MSM); groups at high risk for infectious diseases. These groups are considered to be “hard-to-reach” or “hidden” populations, for which it is either impossible or extremely expensive to obtain a sampling frame. In addition, due to their stigmatized status, these groups generally prefer not to participate in conventional studies. Consequently, it is difficult to get a comprehensive picture of hard-to-reach populations and to apply inferential statistics to population parameters of interest.

PWIDs are the most at risk population for HCV infection and have the highest prevalence in North America. Most surveillance activities associated with monitoring HCV in PWID involve using non-random sampling methods. Non-random methods of data collection in hidden populations have emerged over the past few decades. However, many of these methods are subject to a high risk of sampling/selection bias. These methods include the popular classical snowball sampling, capture-recapture, targeted sampling, time-location sampling and respondent-driven sampling (RDS) ¹⁻⁵. Snowball sampling has been shown to provide informative research data, but the sample obtained may not be the best representation of the population from which it has been drawn as it relies on specific referrals from select individuals ¹. The capture-recapture method includes sampling from facilities such as corrections facilities, hospitals, sexually transmitted infections (STIs) clinics, and drug treatment and needles exchange centers ³. Facility-based sampling has been shown to be useful in obtaining a large enough sample size. However, as with the previous method, the sample drawn from these

facilities may not represent the whole corresponding population, as individuals seeking certain services in these facilities will differ from the individuals who do not seek these services or do not attend these facilities. Targeted sampling has a similar strategy to the snowball sampling, but mostly emphasizes the qualitative aspect. This method includes an initial ethnographic assessment aimed at identifying the different networks within the hard-to-reach populations ⁴. Target sampling can be quite effective in understanding the behavioral risks in hard-to-reach populations; however, the expense, resources and time required to undertake a thorough ethnographic assessment may not be feasible. Time-location sampling, also known as venue-based sampling, targets individuals from hard-to-reach populations at locations where these groups are more likely to gather ². For instance, sex workers may be more likely to be found in brothels, massage parlors and street corners in ‘red light’ districts, whereas, MSM would gather in bars or at events that would be targeted to MSM, and PWIDs could be found around specific inner-city locations ⁴. The major issue with most venue-based data collection is the fact that many of these venues have a tendency to change over time, which could result in sampling bias. In addition, non-response bias is an issue due to the illicit nature of these activities. This type of sampling can also miss important members who do not frequent these ‘popular’ sites ^{1,3,4}.

As of 2016, RDS has been applied in over 460 studies in over 69 countries ^{6,7}. It is also the main mode of data collection for the Centers for Disease Control and Preventions (CDC) as part of the National HIV Behavioral Surveillance System in 25 US cities involving PWID and sex workers. RDS is a sampling method that was influenced by both snowball sampling and traditional contact tracing. It combines a non-probabilistic chain-referral with a statistical model that allows for estimation and inference of population parameters ^{8,9}. The unique aspect of this method is that the respondents are responsible for recruiting other participants ⁸. The process

starts with initial participants selected by the investigators, also referred to as “seeds”, who are interviewed. These participants are supplied with a set of numbered and uniquely coded referral coupons to distribute to other participants and are provided an incentive for their involvement in the study. The recruitment chain continues until the “desired” sample size is reached^{8,9}. The recruitment process is organized in such a way that it allows for the calculation of probabilities of interest and for the identification of the relationship between recruiters and recruits. By doing so, recruitment biases can be assessed and adjusted for in the analysis. The personal network size of every individual is also collected and allows weighted analysis to be conducted to compensate for the oversampling of respondents with larger social networks. Another advantage of this sampling methodology is that it allows the social interaction among these populations to be visualized and measured. As in all sampling methodology, the RDS consists of two components: data collection and inference. Even though this concept is relatively new, large samples of vulnerable people recruited by RDS have been collected successfully in over 69 countries. However, the concern of RDS lies in the inferential component, which relies on six major assumptions, namely; (1) ‘seeds’ are drawn with probability proportion to their network size, also known as degree, (2) individuals in the target population maintain a reciprocal relationship with their contacts, (3) each individual can be reached by another individual through a series of network ties, (4) sampling is with replacement, (5) individual’s network size, degree, is measured accurately, and (6) referrals are random⁸⁻¹¹. In reality many of these assumptions are violated, thus, all RDS estimates produced are biased.

One of the various differences between RDS and traditional sampling is that RDS sample is first used to construct the social network presumed to represent the social network in the population prior to making inference. As the key to transmission of infection is to know the

potential number of susceptible individuals who are exposed to an infected individual, the social dynamics often referred to as the network structure or topology, within that population would need to be understood. RDS provides the tools to conduct social network analysis, which can facilitate the understanding of the dynamics of disease spread in the target population¹². HCV and HIV require more intimate contact among individuals than air- or waterborne infections. Therefore, the conventional epidemiological models normally used to study the spread of infectious diseases are not accurate representations of human behaviour. The standard assumptions associated with such models, such as the SIR (susceptible, infected and recovered) model are that all individuals in a population are equally susceptible to the pathogen and that each individual has a similar number of contacts¹³⁻¹⁵. Realistically, in order for blood borne or sexually transmitted infections to circulate in a population, individuals need to be having unprotected sex and/or being exposed and subjected to infectious blood (i.e. sharing of contaminated needles). These types of transmission methods typically involve some sort of a social, non-random relationship, between two or more individuals, which cannot be easily implemented in the standard infection models. Network topology can be used to investigate these non-random interactions rather than focusing individuals and their behavior which on their own do not transmit infection^{16,17}.

1.1 Social Network Analysis

Social network analysis is a sub discipline of sociology that originated from network and graph theory that focuses on the patterns of links between items, people or events (represented by nodes) and includes systematic methods to analyse connections, relationships or interactions between them¹⁸⁻²¹. Its immediate and intuitive appeal to public health and infectious disease practitioners may be related to the fact that it's basic concept of humans as socially interactive

beings that allow for the transmission of infection, and its resemblance to a contact tracing or partner notification – one of the primary prevention methods in public health.

Roots of social network analysis date back to the 1800s, and over the years sociology, anthropology, biology, medicine, mathematics and statistics, have contributed theoretical and methodological advances to solidify the foundation of modern social network analysis^{18,19,21}. Use of modern social network analysis in infectious disease dates to 1984; where it was used to demonstrate the links between the first 40 cases of AIDS in a group of young gay men from 10 cities in the United States^{22,23}.

A network can be defined as a set of nodes connected by links or edges of one kind or another, such as friendship or particular activity shared between one node to another. As previously stated, the nodes can represent people; the links may be directed or undirected (i.e. individual A lists B as a friend but individual B does not list individual A as a friend (directed) or individual A has sex with individual B; B automatically would include having sex with individual A (undirected))²³. In general, networks are not necessarily connected. All nodes that make up a network are not directly or indirectly connected with each other; especially in social networks, where it is not feasible to get all nodes and links or edges. Some risky networks have demonstrated that incorporating location (i.e. venues where people inject drugs, bars, communities, etc.) may help demonstrate the nodes are all connected indirectly^{24,25}.

Networks typically consists of several components, especially when the population network consists of assortative mixing (also known as homophily, where individuals link with other individuals with similar characteristics, such as demographics of infectious status)^{21,26}.

Disassortative mixing in networks, also known as heterogeneity, is where individuals link with

other individual opposite of them in terms of demographics, infectious status, etc.^{21,26}. A component in which a node belongs is that set of nodes that can be reached from it by paths running along links/edges of the network^{21,26}. A network is said to have a giant component if a single component contains the majority of nodes in the network. The concept of a giant component is crucial when considering disease dynamics within a network²⁶. Strongly connected component contains nodes that can be reached from each other, where nodes are interconnected, a term known as clustering (i.e. the number of possible individual within a group that would know each other)^{26,27}. Components with high clustering are most at risk from infection imported from a single and random node, since the infection will easily be transmitted to all nodes in the component. Clustering is an important property that defines the social network structure.

Human networks incorporate both an individual's (or 'egocentric') network and a social (or 'sociocentric') network. The term personal network, also known as the degree, refers to an individual's direct contact with other individuals, the social relationships that links these persons to the focal person, and if possible, the social relationships among those linked to the focal person (i.e. among the focal individual's contacts, how many of them also have a social relationship with each other). Egocentric data generally consists of information on a number of individuals (the egos) and their contacts (the alters)^{26,27}. The term social network applies to the whole set of individuals in a particular population and the links or edges connecting them. In a population, there may be as many degrees as there are individuals; in the certain target populations all of these degrees may possibly be connected together to form a single social network^{24,25,28,29}. This leads to another important property that is crucial in understanding the network structure, which influences diseases within social network, known as the degree distribution.

1.2 Network Topology/Structure

The structure or topology of a network has consequences for its individual members and for the network as a whole. Network structures are classified according to the distribution of their connections (degree distribution),³⁰⁻³² which describes the frequency with which nodes or individuals with a certain number of connections or contacts occur in a network. If the degree of a node is defined as k , then the degree distribution is defined as the set of probabilities, $P(k)$, that a node chosen at random will have degree k ³⁰⁻³². The degree distribution is an important property of a network as it naturally captures the heterogeneity in individuals' potential to become infected and potential secondary infections³⁰⁻³². The higher the number of links or edges a node has, the more likely it is to be in contact with an infected node. In addition, the more contacts a node has, the more likely it is to cause a large number of onward cases³⁰⁻³².

Generally in social networks, there are very few individuals who are linked to a significant number of nodes (hubs), whereas the majority of nodes are linked to very few nodes (i.e. individuals generally have few number of sexual partners; whereas, very few have high numbers of sexual partners)³³⁻³⁵. Sexual networks have small-world structures, where people are connected to each other by relatively short paths, and connections are clustered around hubs³³⁻³⁵³⁰⁻³². Sexual networks have small-world structures, where people are connected to each other by relatively short paths, and connections are clustered around hubs³³⁻³⁵. Interventions directed at random individuals in small-world networks are ineffective unless the individuals targeted are the ones with the most contacts. Understanding the connectivity amongst high risk groups will assist in improving public health interventions to prevent or reduce the spread of infection.

For nearly a century, random network structures dominated the study of real world interactions^{14,30,31}. The Erdős and Renyi (ER) random network has been the most popular random model used to study spread of infectious diseases and assumes that each individual within the network has equal probability of being infected and each individual has a similar number of contacts^{36,37}. ER models rely on random interactions which produce a Poisson degree distribution, in which the average of links per node is a good representation of the mean number of contacts in the population. However, this is inconsistent with empirical social network data, especially in sexual networks, above. In PWID populations, a few people inject very frequently in a specified period, while the majority inject relatively seldom within that same time frame^{35,38,39}. Several studies have demonstrated that networks that involve sex workers and PWID exhibit a small-world structure, where connections are created through a combination of randomness and rules^{13,14,33}. Certain small-world networks produce an extremely skewed distribution, best described by some form of power law. Small-world networks that exhibit some form of power-law have been classified into three different types, depending on their degree distribution: scale-free, broad-scale and single-scale networks^{40,41}. Scale-free networks are the distribution that is largely implicated in describing networks of complex human interactions, such as sexual partners and needle sharing^{13,14,33,41-44}.

As previously stated, scale-free networks have a power law degree distribution, where the $P(k) = Ck^{-\sigma}$, where C and α (the decaying coefficient) are constants and $0 < \alpha < 3.5$, and the heavy tail end of the distribution represents the individuals classified as hubs. If bloodborne pathogen networks are best represented by the scale-free networks, then, the usual sample statistics achieved in random, normal distributions, such as the mean, are not very useful in explaining the patterns observed in these types of populations. Power laws are used to

characterize the transition from disorder to order by identifying the highly connected hubs. These connected hubs are observed at a higher frequency than in random networks and play a crucial role in the spread of diseases^{33,45-48}.

There are two well-known mechanisms affiliated with small world properties, known as the Watts-Strogatz (WS) and Barabási-Albert (BA) models^{31,49-53}. The major difference between these two is that WS does not exhibit a power law. The WS model features two important properties: high clustering and the short average path lengths between individuals in the network. The WS better captures real world networks and is a combination of the rules of the regular lattice and random network^{49,54-56}. The BA model, on the other hand, is based on the two important mechanisms of growth and preferential attachment^{52,53,57,58}. Growth refers to the number of individuals within a network that will increase over time and preferentially attach to individuals that are already highly connected and are individuals most likely receive new links over time. This latter concept has been known as the effect of the “rich get richer”. The BA model generates a random scale-free network⁵³.

1.3 RDS Inference

RDS relies on respondents to recruit fellow members of a hidden population. The methodology has proven to be effective in obtaining a large enough sample size in multiple applications worldwide⁵⁹⁻⁶³. However, when it comes to making reliable and valid inference on RDS data to the population of interest, the assumptions associated with statistical component of RDS process make inferences very challenging. For the most part, most applied RDS studies focus on producing population prevalence estimates without fully investigating and/or incorporating the appropriate network structure of the target population, the quality of the data, and the violations of the assumptions⁶⁴⁻⁶⁶.

The RDS assumptions listed earlier may be appropriate for certain hard-to-reach populations but for the most part bears little resemblance to reality. For instance, ‘seeds’ are drawn with probability proportion to their degree and degree is measured accurately (assumption 1 and 5, respectively), is dependent on the respondent recollection and in certain scenarios may be difficult to get accurate numbers which then impacts whether the seed is drawn with probability proportional to their degree. In the PWID population, there are instances where a person will be on a binge of injections in a shooting gallery, and depending on the number of injections, is very likely to inject with needles previously used by other PWID. The individual in such instances may be unaware of the individual who initially used the needle or may not recall the individual as he or she may not be part of the person’s network. Individuals within a network are undirected (assumption 2), that is the individuals in the target population maintain a reciprocal relationship with their contacts. For most bloodborne pathogen networks (i.e. sex workers, PWID), that can be an appropriate assumption. However, if one were interested in seeing a needle sharing network among PWID, there are individuals who will inject used needles from one person; and it may not be reciprocated for that same person (for instance, individual A injects and provides their needle to individual B but individual A does not take used needles from individual B).

The network is one giant component (assumption 3), this may hold true for certain target populations if egocentric and location data are collected and incorporated in the social network (i.e. PWID); however, in the sex worker population, this has been demonstrated to be untrue, as there are tiers of sex workers and the components within the network do not overlap⁶. Sampling is with replacement (assumption 4) and that is violated in every instance of applied RDS since data collection does not allow individuals to participate more than once⁸. Recruitment is

uniform and random (assumption 6). Each individual recruits one person within their personal network at random. It is known that RDS provides each respondent more than one coupon and that individuals preferentially chose recruiters based on certain characteristics and/or economic incentives, as respondents are provided an incentive for participating and, for most RDS studies, for recruiting.

In the last two decades, RDS estimators have emerged with less stringent conditions to facilitate inferences. Many of the developed RDS estimators rely on the assumption that the recruitment process is a Markov Process (MP)^{9-11,67,68}. The MP is equivalent to a random walk on a symmetric (undirected) graph, the most common assumption associated with RDS application^{9,10,69}. In essence, the recruitment process demonstrates a stochastic process with two crucial components. The first being, that the process can assume a limited number of states and the second, the process is state dependent, such that the probability of a certain state being reached depends only on the immediate previous state of the chain. To put into context with RDS design, the recruitment pattern is dependent only on the recruiter and not on the previous recruiters. This type of process is considered to be memoryless and is known as the first order of Markov process (FOMP)^{9,11,67}.

Markov chains are founded on the property of being ergodic. This term means that as a process moves from state to state, it will move to any state in the chain proportionate to the states' probability; thus, there is a zero probability that any state will never recur^{8,9,70}. To apply the ergodic term specifically to RDS, this means that a recruiter with one set of characteristics can recruit another subject with the same or different characteristics^{8,9}. A recurring state means that after one or more recruitment waves a recruit can have the same characteristics as the earlier

recruiter. This important property relies on two chain characteristics: irreducible and aperiodic. The chain is irreducible if for every state there is a positive probability of moving to any other state and the chain is aperiodic if it can explore every state possible^{8,9}. In other words, the recruitment process cannot become trapped within a single group or set of groups. There are two theorems in the modeling the recruitment process as a MP that are relevant in understanding the reliability of statistical indicators drawn from RDS. Theorem 1 is known as the Markov chain law of large numbers and in RDS terminology means that as the recruitment process continues from wave to wave, equilibrium will be attained of the composition of the different characteristics of the recruits that is independent of the characteristics of the seeds (the set of individuals in which recruitment began)^{8,9}. This theorem is the argument that most of the developed estimators use to explain why assumption (1) above, could be relaxed. In essence, at equilibrium, the sample composition is independent of the seeds characteristics. Theorem 2, introduced by Kemeny and Snell (1960), indicates that the set of recruits obtained by a RDS proceeds at a rapid pace, resembling an exponential rate^{8,9}. Thus, as the sample expands wave to wave, the mean composition of the recruits in the leading wave approaches equilibrium approximately at an exponential rate, as long as the recruitment chain is long. Heckathorn (2002) estimated that equilibrium could be reached within six or fewer waves⁷¹. The chain in equilibrium is known to have a stationary distribution from which dependent samples can be drawn.

Modeling RDS procedure as a MP in which estimates about the population can be made, assumes that there is a symmetric relationship (undirected network), the network is connected (irreducible), single recruitment occurs, that sampling is with replacement, and the uniform recruitment assumption is satisfied^{8,9}. These are strong assumptions of which most, if not all, fail

in practice. Multiple referrals were included in the RDS design to reduce bias from the choice of seeds and to avoid the non-response of a potential recruit and termination of the branch^{8,9}. Thus, individuals having trouble recruiting others will not stop the sampling process and given enough time and resources, the recruitment chains of virtually any length can be generated. The problem with including multiple referrals when relying on the Markov chain model is that the actual recruitment process resembles a tree shaped structure; whereas, a linear structure is assumed in the Markov model. Heckathorn argues that the tree-shaped referral structure can be regarded as a set of linear structures in which the respondents from the final wave of the recruitment process can be seen as the product of a linear chain beginning with the seed and whose links correspond to intermediate waves^{8,9}. However, another issue arises with multiple referrals that are often encountered in applied RDS studies and renders RDS estimators substantially biased and that is differential recruitment⁷². Differential recruitment occurs when respondents preferentially recruit their contacts with particular interest⁷².

Several recent studies have emerged assessing the RDS methodology and the extent to which the estimators are biased if assumptions are not satisfied. There are now debatable concerns regarding whether RDS process modeled as MP can actually be considered stationary, as well as whether the memoryless process holds any truth in the recruitment procedure. There is also the concept of “bottlenecking” that has brought much attention in the field of RDS. A “bottleneck” is the term used to describe when sampling occurs only in one part of the entire social network, and may happen due to immeasurable characteristics influencing the direction of the recruitment process⁶⁹.

Researchers have investigated the impact of the assumption violations of RDS on the sampling variance of the RDS estimates. Variances estimates have been demonstrated to be severely underestimated even when assumptions have been met^{64, 69, 73}. Goel and Salganik performed simulations on RDS samples on empirical networks and identified that RDS variances estimates generated 5 to 10 times greater variance than simple random sample (SRS)⁷⁴. McCreesh *et al.* conducted a RDS study on male household heads in rural Uganda, where the true population was known, and they found that 95% confidence intervals (CI) actually represented only 50 to 74%⁷³. So far most of the studies investigating the variance associated with RDS estimators have concluded that the network structure and the RDS process affects the design effect and the downward bias of the RDS variance estimate. As such, it is clear that the quantification of uncertainty associated with the RDS mean estimates remains largely open.

1.4 RDS PWID data in Ottawa

The following body-of-work will mainly focus on investigating properties that influence network structure in order to help identify the most appropriate RDS mean and variance estimate to use in the PWID population in Ottawa. The social network will be constructed by using RDS PWID data, where the nodes are represented by individuals who inject drugs. The links/edges will differ based on the network property assessed. The evaluation of the degree distribution will be using the link/edge of node to represent the activity of injecting (used as proxy of sharing needles). Normally, the link/edge is associated with another node; however, in this particular scenario, the number of individuals sharing needles was not obtainable, and thus, it is assumed that the more the node injects, the more the injector is associated with another injector (i.e. sharing needles). In assessing, whether the RDS PWID data exhibit a small-world network, the clustering and the average path length within the RDS PWID data will be identified. The link

and edges in this scenario represent the social relationship between the injector and their contact (alter). The alters of the injectors can either be PWID or non-PWID.

1.5 References

1. Heckathorn DD. Snowball versus respondent-driven sampling. *Sociological methodology*. 2011;41(1):355-366.
2. Kral AH, Malekinejad M, Vaudrey J, et al. Comparing respondent-driven sampling and targeted sampling methods of recruiting injection drug users in san francisco. *J Urban Health*. 2010;87(5):839-850.
3. Berchenko Y., Frost S.D.W. Capture-recapture methods and respondent-driven sampling: Their potential and limitations. *Sex Transm Infect*. 2011;87(4):267-268.
4. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*. 2005;19 Suppl 2:S67-72.
5. Malekinejad M., Johnston L.G., Kendall C., Kerr L.R.F.S., Rifkin M.R., Rutherford G.W. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS Behav*. 2008;12:S105-S130.
6. Verdery A.M., Merli M.G., Moody J., Smith J.A., Fisher J.C. Brief report: Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in china. *Epidemiology*. 2015;26(5):661-665.
7. Baraff A.J., McCormick T.H., Raftery A.E. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proc Natl Acad Sci U S A*. 2016;113(51):14668-14673.
8. Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Probl*. 1997;44(2):174-199.
9. Heckathorn D.D., Semaan S., Broadhead R.S., Hughes J.J. Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18-25. *AIDS Behav*. 2002;6(1):55-67.
10. Heckathorn DD. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociol Methodol*. 2007;37(1):151-207.
11. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol Methodol*. 2004;34(1):193-240.
12. Young AM, Jonas AB, Mullins UL, Halgin DS, Havens JR. Network structure and the risk for HIV transmission among rural drug users. *AIDS Behav*. 2013;17(7):2341-2351. 2.
13. Liljeros F, Edling CR, Nunes Amaral LA. Sexual networks: Implications for the transmission of sexually transmitted infections. *Microbes Infect*. 2003;5(2):189-196.

14. Keeling MJ, Eames KTD. Networks and epidemic models. *Journal of the Royal Society Interface*. 2005;2(4):295-307.
15. Doherty IA. Sexual networks and sexually transmitted infections: Innovations and findings. *Curr Opin Infect Dis*. 2011;24(1):70-77.
16. Yu X-, Wang X-, Zhang D-, Liang F, Wu X. Mathematical expressions for epidemics and immunization in small-world networks. *Physica A: Statistical Mechanics and its Applications*. 2008;387(5-6):1421-1430.
17. Li X, Chen G, Wang X. On spreading dynamics in discrete small-world networks. . 2005;16:1107-1111.
18. Christley RM, Pinchbeck GL, Bowers RG, et al. Infection in social networks: Using network analysis to identify high-risk individuals. *Am J Epidemiol*. 2005;162(10):1024-1031.
19. Luke DA, Harris JK. Network analysis in public health: History, methods, and applications. *Annu Rev Public Health*. 2007;28:69-93.
20. Martinez-Lopez B, Perez AM, Sanchez-Vizcaino JM. Social network analysis. review of general concepts and use in preventive veterinary medicine. *Transbound Emerg Dis*. 2009;56(4):109-120.
21. Klov Dahl AS. Social networks and the spread of infectious diseases: The AIDS example. *Soc Sci Med*. 1985;21(11):1203-1216.
22. Adler MW, Weller IV. AIDS: Sense not fear. *Br Med J (Clin Res Ed)*. 1984;288(6425):1177-1178.
23. Auerbach DM, Darrow WW, Jaffe HW, Curran JW. Cluster of cases of the acquired immune deficiency syndrome. patients linked by sexual contact. *Am J Med*. 1984;76(3):487-492.
24. D'Angelo-Scott H, Cutler J, Friedman D, Hendriks A, Jolly AM. Social network investigation of a syphilis outbreak in ottawa, Ontario. *The Canadian journal of infectious diseases & medical microbiology = Journal canadien des maladies infectieuses et de la microbiologie medicale*. 2015;26(5):268-272.
25. Al-Azem A. *Social network analysis in tuberculosis control among the aboriginal population of manitoba*. [Ph.D.]. University of Manitoba; 2006.
26. Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. . Structural analysis in the social sciences ed. New York, NY, US: Cambridge University Press.; 1994.
27. Ancel Meyers L, Newman ME, Martin M, Schrag S. Applying network theory to epidemics: Control measures for mycoplasma pneumoniae outbreaks. *Emerg Infect Dis*. 2003;9(2):204-210.

28. Srinivas MN. Social structure. *Sociological Bulletin*. 1964;13(1):12-21.
29. Logan JJ, Jolly AM, Blanford JI. The sociospatial network: Risk and the role of place in the transmission of infectious diseases. *PLoS One*. 2016;11(2)
30. Albert-Laszlo³ Barabási. The physics of the web. *Physics World*. 2001;14(7):33.
31. Hein O, Schwind M, König W. Scale-free networks. *WIRTSCHAFTSINFORMATIK*. 2006;48(4):267-275.
32. Newman ME, Moore C, Watts DJ. Mean-field solution of the small-world network model. *Phys Rev Lett*. 2000;84(14):3201-4.
33. Liljeros F, Edling CR, Stanley HE, Åberg Y, Amaral LAN. Sexual contacts and epidemic thresholds. *Nature*. 2003;423:606.
34. Newman MEJ, Jensen I, Ziff RM. Percolation and epidemics in a two-dimensional small world. *Physical Review E: Statistical, Nonlinear, & Soft Matter Physics*. 2002;65(2 Pt 1):021904.
35. Pelude L. *Networks among injection drug users: Random or scale-free?* [Masters]. University of Ottawa; 2007.
36. Newman ME, Moore C, Watts DJ. Mean-field solution of the small-world network model. *Phys Rev Lett*. 2000;84(14):3201-3204.
37. Newman ME, Watts DJ, Strogatz SH. Random graph models of social networks. *Proc Natl Acad Sci U S A*. 2002;99 Suppl 1:2566-2572.
38. Brady K.A., Nassau T.B., Shinefeld J., Mezzacappa C. Analyzing trends in HIV risks for injection drug users by respondent-driven sampling. *Top Antiviral Med*. 2015;23:473-474.
39. Dombrowski K, Curtis R, Friedman S, Khan B. Topological and historical considerations for infectious disease transmission among injecting drug users in bushwick, brooklyn (USA). *World J AIDS*. 2013;3(1):1-9.
40. Amaral LA, Scala A, Barthelemy M, Stanley HE. Classes of small-world networks. *Proc Natl Acad Sci U S A*. 2000;97(21):11149-52.
41. Amaral LAN, Scala A, Barthélemy M, Stanley HE. Classes of small-world networks. In: Princeton University Press; 2011:207-210.
42. Liljeros F, Edling CR, Amaral LAN, Stanley HE, Åberg Y. The web of human sexual contacts. *Nature*. 2001;411:907.
43. Reed WJ. A stochastic model for the spread of a sexually transmitted disease which results in a scale-free network. *Math Biosci*. 2006;201(1-2):3-14.

44. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data . *SIAM Review*. 2009;51(4):661-703.
45. Dombrowski K, Khan B, Moses J, Channell E, Dombrowski N. Network sampling of social divisions in a rural inuit community. *Identities*. 2014;21(2):134-151.
46. Chen J, Le A, Wang Q, Xi L. A small-world and scale-free network generated by sierpinski pentagon. *Physica A: Statistical Mechanics and its Applications*. 2016;449:126-135.
47. Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs; 16399673. *Journal of Neuroscience*. 2006;26(1):63-72.
48. Andrade Jr. JS, Herrmann HJ, Andrade RFS, Da Silva LR. Apollonian networks: Simultaneously scale-free, small world, euclidean, space filling, and with matching graphs. *Phys Rev Lett*. 2005;94(1).
49. Watts DJ, Strogatz SH. Collective dynamics of small world networks. *Nature*. 1998;393:440.
50. Lou J, Ruggeri T. The dynamics of spreading and immune strategies of sexually transmitted diseases on scale-free network. *Journal of Mathematical Analysis and Applications*. 2010;365(1):210-219.
51. Albert R, Barabási A. Statistical mechanics of complex networks. *Reviews of Modern Physics*. 2002;74.
52. Barabási A, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512.
53. Barabási A. Scale-free networks: A decade and beyond. *Science*. 2009;325(5939):412-413.
54. Zimmermann J-. "Clusters" and "small-worlds":Aproximity issue. *Revue Economique*. 2002;53(3):517-524.
55. Dennie D, Cuccia LA. "Six degrees of separation"-revealing a "small-world phenomenon" through a chemistry literature search activity. *J Chem Educ*. 2014;91(4):546-549.
56. Canale MH. "Small-world" network. *Medico e Bambino*. 2009;28(7):459-460.
57. Zhang Y, Zhang Z, Zhou S, Guan J. Deterministic weighted scale-free small-world networks. *Physica A: Statistical Mechanics and its Applications*. 2010;389(16):3316-3324.
58. Chen G, Hao J, Zhang Z, Tang Y. Getting scale-free network from a small world network without growth. . 2012:7-9.

59. Johnston L.G., Hakim A.J., Dittrich S., Burnett J., Kim E., White R.G. A systematic review of published respondent-driven sampling surveys collecting behavioral and biologic data. *AIDS Behav.* 2016;20(8):1754-1776.
60. Malekinejad M., McFarland W., Vaudrey J., Raymond H.F. Accessing a diverse sample of injection drug users in san francisco through respondent-driven sampling. *Drug Alcohol Depend.* 2011;118(2-3):83-91.
61. Wejnert C. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol Methodol.* 2009;39(1):73-116.
62. Shaghaghi A, Bhopal RS, Sheikh A. Approaches to recruiting 'hard-to-reach' populations into re-search: A review of the literature. *Health promot perspect.* 2011;1(2):86-94.
63. Sperandei S, Bastos LS, Ribeiro-Alves M, Bastos FI. Assessing respondent-driven sampling: A simulation study across different networks. *Soc Netw.* 2017
64. Verdery AM, Mouw T, Bauldry S, Mucha PJ. Correction: Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE.* 2016;11(1):e0148006.
65. Lu X. Linked ego networks: Improving estimate reliability and validity with respondent-driven sampling. *Soc Netw.* 2013;35(4):669-685
66. Gile KJ, Handcock MS. Respondent-driven sampling: An assessment of current methodology. *Sociol Methodol.* 2010;40(1):285-327.
67. Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics.* 2008;24(1):79-97.
68. Salganik M.J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Urban Health.* 2006;83(1).
69. Goel S., Salganik M.J. Respondent-driven sampling as markov chain monte carlo. *Stat Med.* 2009;28(17):2202-2229..
70. Kemeny JG, Snell JL. *Finite markov chains.* The University of Michigan: Van Nostrand; 1960.
71. Heckathorn DD. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Soc Probl.* 2002;49(1):11-34.
72. Tomas A, Gile KJ. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron J Stat.* 2011;5:899-934.
73. McCreesh N., Frost S., Seeley J., et al. An empirical evaluation of respondent-driven sampling. *Sex Transm Infect.* 2011;87:A15-A16.

74. Goel S. Assessing respondent-driven sampling. *Sex Transm Infect.* 2011;87:A15.

2.0 Objectives and Hypothesis

Given that in reality some of the assumptions associated with traditional infectious disease networks and real RDS data may not be satisfied, this work will concentrate on analyzing the data using techniques considered to be robust under departure from the violated assumptions. The overall goal of this doctoral thesis is to provide more insight into the statistical inference for respondent-driven sampling of people at high risk, people who inject drugs, for HCV and HIV. In order to make improvements on RDS inference, it is first important understand the network structure/topology of RDS PWID data prior to assessing which RDS mean and variance estimate to use in an RDS study. RDS estimators rely greatly on the network structure. Thus, the study will use behavioral and egocentric data to better understand the social network structure of the PWID population. Figure 1 summarizes the traditional outline of RDS data collection and inference and highlights the areas how the work contributes to this growing field.

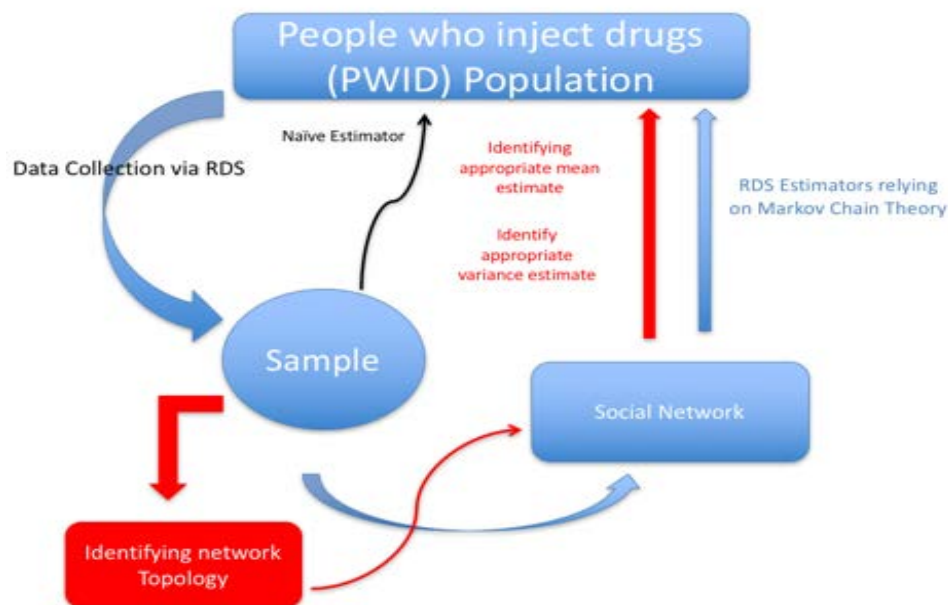


Figure 1: Summary of the traditional way of conducting RDS inference as well as our potential contribution to RDS inference in people who inject drugs (highlighted in red). Adapted from Saglanik & Heckathorn .

The objectives of this thesis are:

1. To determine whether the distribution of the number of injections in the PWID network over a 6 month period exhibits a power law distribution.
2. To determine whether the social contacts within the PWID network constitute a small-world network.
3. To describe and compare the development and progress of RDS mean and variance estimates, through a systematic review, and provide recommended estimators for future RDS studies.

The hypotheses of this thesis are that: (1) the network structure of people who inject drugs, using both frequency of injections and social contacts, do not exhibit properties of an Erdős-Rényi random model but rather have small world attributes; and (2) the performance of the developed RDS mean and variances estimators may perform similarly in both simulation studies and real RDS data to those of the naïve estimator.

3.0 Study Population

This secondary data that is being used for the next two chapters focused on PWIDs in Ottawa, Ontario, which were recruited from September to December 2007, by RDS. The first publication produced from the PWID study was by Pilon *et al.* in 2011 (Pilon *et al.*, 2011).

3.1 Method of Recruitment

The recruitment process started with 7 individuals known to be a part of the PWID population and have participated in similar PWID studies. These individuals are known as seeds and were responsible for initiating the recruitment of other PWIDs. Each respondent was provided with three marked “coupons” (they contain the person’s numeric study code). Seeds and later participants were directed to ask any of their friends who also inject drugs if they would be interested in participating in the study. Study participants presented the coupons, such that sequential waves of enrolment occurred and anonymous linked chains of study participants were created. Wave 1 recruits were those recruited by the seeds; wave 2 participants were those recruited by wave 1, etc. Theoretical calculations and empiric observations, based on Markov chain sampling theory, indicated 6 waves were required for a sample size of 400. Our study included 407 PWIDs in Ottawa (Pilon *et al.*, 2011).

3.2 Data Collection

Each respondent was required to make an interview arrangement with the study coordinators in order to complete a lengthy questionnaire consisting of 2 sections: (1) demographics and behaviors and practices of HIV and HCV infection, and (2) social and risk networks, which incorporated egocentric methods. In addition, each participant was asked to provide separate consent to an anonymous blood test. The SurgiLance One-Step Safety Lance, a

device that collects blood from the fingertip, was used to collect a biological specimen from each respondent. The blood spots were tested at the National Laboratory for HIV and Retrovirology in Ottawa. A compensation of 50\$ CAN, transportation fees (i.e. bus fare) and a prevention pack, containing safer injection and safer sex resources, were provided to each respondent for their time spent in completing the questionnaire/interview and the collection of blood spots.

3.3 Serological Assays

Dried blood spots (DBS) specimens were tested for the presence of both HIV and HCV antibodies. HIV antibodies were first screened by the Genetic Systems rLAV HIV-1 enzyme immunoassay (EIA) (BioRad, Canada). If DBS was tested positive for HIV antibodies, a confirmatory testing using the Genetic Systems HIV-1 Western Blot (BioRad, Canada) was conducted. For assessing whether HCV antibodies were present in the DBS samples, the Genetic Systems rLAV HIV-1 EIA diluent was first conducted and the eluates were then tested using the Ortho HCV Version 3.0 enzyme-linked immunosorbent assay (ELISA). Unlike HIV antibodies testing, no confirmatory HCV assays were available at the time of testing. For further detailed information, please refer to Pilon *et al.* (Pilon *et al.*, 2011).

3.4 Molecular Analysis

Nucleic acids were isolated from all HIV and/ or HCV positive DBS specimens following the manufacturer instructions of the Nuclisens EasyMag System (BioMerieux, Canada). The extracted nucleic acids were used to amplify the HIV and HCV genetic regions using reverse transcriptase polymerase chain reaction (RT-PCR) methods. The amplified templates were sequenced using an ABI Prism genetic analyzer. The sequences were generated by aligning electropherograms to a reference sequence with SeqScape software (Applied Biosystems, Canada). The sequences were then aligned using Clustalx software. The molecular

analysis of the HIV and/or HCV samples obtained from the Ottawa PWID RDS study allowed for molecular epidemiology analysis to be conducted. Further details and results can be referred to Pilon *et al.* (Pilon *et al.*, 2011).

3.5 Ethics

The study design was approved by the Ottawa Hospital Research Ethics Board (certification #: 2007045-1H). As RDS are controlled by the members of the vulnerable population all recruitment is done anonymously, and does not require respondents to divulge sensitive information about identified individuals. Therefore, peoples' privacy is protected and they are more likely to disclose sensitive information relevant to harm reduction. In addition, respondents who wished to undergo HIV and HCV counselling and testing, from which they may receive their results, were referred to the Site needle exchange programme with which an agreement was made to provide services to study participants.

4.0 Identifying a Heavy Tailed Distribution, Using Frequency of Injections, Among People Who Inject Drugs Network in Ottawa, Canada

This chapter contains the first of three manuscripts relating to this dissertation that provides insight into the PWID population in Ottawa, and possibly the PWID in other similar cities, that are normally not captured by traditional non-random sampling methods by using RDS. RDS can provide a more comprehensive picture of how social and risk behaviours influence diseases among this population. This study incorporated methods originating from the disciplines of mathematics and physics to focus on one important aspect of network structure, the degree distribution of individuals engaging in risky bloodborne behaviours in PWID network using RDS.

In the context of PWID, bloodborne pathogens are typically transmitted when an infected individual shares needles with a non-infected individual. The ideal degree distribution in this scenario would be to identify the distribution of the number of individuals each person shares needles with in the network. Unfortunately, due to recall and/or social desirable bias this information could not be obtained. As such, the frequency of injection was used as proxy of needle sharing, which has been further supported by epidemiological and ethnographic studies. Individuals with a greater injection frequency were assumed to be more likely to share needles with many other PWIDs. This chapter provides tools to identify key individuals to target for effective interventions using RDS.

Abdesselam, K.¹, Pelude, L.¹, Verdery, A.², Momoli, F.¹, and Jolly, A.M.¹

¹ School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine,
University of Ottawa

² Department of Sociology and Criminology, Pennsylvania State University

4.1 Abstract

Understanding the topology of contact networks can provide insight on the dynamic processes underlying the spread of infectious diseases, risk behavior interactions and inform preventative strategies. Networks formed around stigmatized activities exhibit structures that may not conform to the assumptions of conventional modelling of infectious diseases. One way to understand how contact networks might influence the spread of infectious diseases is to classify them into different network types according to their topology. By examining this distribution, we can assess whether the contact network is consistent with several different archetypal networks, each of which has different implications for infection transmission at the population level. We tested whether the degree distribution of people who inject drugs (PWID) sampled through respondent-driven sampling fit a Poisson distribution, exponential distribution, power-law distribution, and/or lognormal distribution. We then compared the relative fit of each model to each other. Results indicate that the degree distribution in the contact network was not consistent with a random model, represented by the Poisson distribution. Our results, however, do find that the degree distribution in the contact network is consistent with a power law, which suggests a scale-free network type. There was insufficient evidence to conclude that PWID networks exhibit a scale-free topology, because we cannot distinguish the power-law distribution from other heavy tailed distributions like the lognormal or stretched exponential distributions, such as the Weibull. All told, these results suggest that targeting interventions to PWID in the

‘heavy tail’ of the distribution is more likely than interventions with random members of the network to be successful and efficient. Further work should investigate other topological properties of PWID contact networks.

4.2 Introductions

Even though there has been substantial progress in screening, treatment, and education concerning sexually transmitted and blood borne infections (STBBIs) among high risk groups in Canada and internationally, STBBIs continue to be a major concern. The United Nations has established a “cascade of care” target, referred to as “90-90-90 by 2030”, that aims to have 90% of people living with Human Immunodeficiency Virus (HIV) know their status, 90% of people who know they are HIV positive accessing treatment, and 90% of people on treatment have suppressed viral loads ^{1,2}. The UN has set similar targets for a 30% reduction in new Hepatitis B (HBV) and Hepatitis C (HCV) infections and a 30% increase in treatment of both ¹⁻³. Specifically, among some people who inject drugs (PWID), transmission of HIV, HBV and HCV continues along with increased infection rates despite their increased access to clean needles and other safer injection supplies through harm reduction programs. In addition, many PWID do not know their infection status and even for those that know they are positive it is difficult to access treatment ^{1,4,5}. In order to achieve global targets and implement effective interventions, especially among more hard-to-reach populations, researchers must identify other mechanisms that may optimize reduction of transmission. Understanding the dynamics of infectious disease transmission in hard-to- reach populations such as sex workers, PWID, street youth, those who are homeless, and men who have sex with men (MSM) is a key part of these efforts.

Many of the classic and most influential mathematical and statistical models that are used to understand the dynamics of infectious diseases assume all individuals in a population are equally susceptible to the pathogen and each individual has a similar number of contacts and that each individual is likely to come into contact with any other individual in the population ^{6,7}. Although these assumptions may be valid for airborne or waterborne infections, where individuals within the population are estimated to have approximately equal transmission probabilities upon contact with infected individuals, they do not apply to STBBIs that require more intimate, non-random social contacts between individuals ⁷⁻⁹. The types and diversity of social contacts that circulate the spread of STBBIs are not well represented in standard infection models. Social network analysis and modeling provides an alternative approach to studying the diffusion of HIV and HCV within high-risk populations ⁶.

Network structure, which is defined as the aggregated pathways of local connections that link members of a population together, provides information over and above accepted epidemiologic and mathematical incidence measures. One important means of classifying network structures is by the frequency distribution of the number of direct links - denoting sexual or equipment sharing interactions - between individuals, represented by nodes ; this is often referred to as the “degree distribution” ¹⁰⁻¹². Typical sexual contact networks contain very few individuals who have many sex partners while the majority of individuals have few partners. Networks of PWIDs exhibit similar properties, with a small proportion of people who inject hundreds of times a week which increases the probability of sharing injection equipment with multiple partners and as a result increases the probability of infection and transmission ^{5,6,13-16}.

The social networks literature has developed a series of archetypal networks that, collectively, describe a wide range of contact network types and, individually, suggest very

different transmission dynamics and epidemic potential. For instance, both the Erdős-Rényi (ER) and Barabási-Albert (BA) models have been used to identify the structure or topology of networks in order to better understand the how network contexts can contribute to the spread of infectious disease¹⁷⁻²⁰. The ER model is a random graph approach to understanding network structures that has been the most commonly used model in network theory and mathematical modeling in the absence of empirical data. ER network models assume that (1) each individual within the network has roughly the same number of contacts; (2) the average number of contacts accurately reflects contact frequency; (3) variations in characteristics among the contacts per individual within the network are minimal and (4); when an infectious disease is introduced to this network, each individual within the network has the same probability of getting infected^{10,15,16}. The contact distribution in an ER model resembles a Poisson distribution.

The BA model is known for its scale-free distribution of contacts, wherein hubs are formed through (1) network growth and (2) preferential attachment^{10,18,20,21}. The BA model generates a random, scale-free distribution of connections that takes the characteristics of a power law distribution. In this context, network growth refers to the fact that the number of individuals within a network will increase over time, and preferential attachment is the principal by which individuals who are highly connected (have a large number of contacts) are those most likely to acquire new contacts. The latter concept is known as “rich get richer” or the Matthew effect²¹. Power law distributions have been associated with the preferential attachment mechanism but in certain scenarios can also be generated by multiplicative processes^{122,238-20}. For instance, the well-known Pareto distribution, also a power law distribution, was first developed to explain the income distribution in Italy, where 80% of the income came from 20% of the population^{22,23}. In addition, when income distributions are categorized, the lower end of

the scale resembles a lognormal distribution; whereas the high end fits a power law. For instance, in a classic work describing file sizes on computers, Reed shows that certain real life distributions are difficult to identify as either power law or lognormal; instead, sometimes both distributions are required to properly delineate the distribution of the data ^{22,24}.

Among PWID, frequency of injections may be a more accurate measure of exposure than the number of individuals an injector injects drugs with. An individual or contact with whom an individual shares a needle or contaminated equipment constitutes an exposure which can transmit a virus. One person sharing needles or equipment with only one individual (whether they know them or not) is the most distinctive exposure. However, indirectly sharing needles or equipment includes the following; a person uses the needle first but then gives it to a friend; borrows from a friend while unaware that it was shared with other people or picking up abandoned needles up from a coffee table or on the floor. Also when PWID respond to surveys, many of which are administered by public health staff, they are reluctant to admit to sharing as they know sharing increases the risk of HIV/HCV or HBV transmission. In addition, many PWID feel that admitting to sharing drug injection equipment may result in resources being withdrawn from the community due to poor program evaluation indicators. Secondly, when PWID are on a “binge,” defined as a period of time when individuals are using drugs (injecting) more excessively than their usual pattern, the frequency of needle use is high and the likelihood of having enough needles available is low ^{5,13,16,19}. Lastly, the recall of a person who was on a binge in relation to their risk behaviours at the time is not as accurate as at other times. Therefore, their number of contacts (those they shared drug injection equipment with) may remain largely unknown. Therefore, the distribution of the frequency of injection is more likely to resemble the risk exposure distribution ^{5,13,16,19}. Though this is not as precise a measure as actually knowing the

number of individuals with whom a PWID has shared injection equipment (knowingly or unknowingly) or the number of times they used ‘abandoned’ needles , it appears to be better than relying on PWID responses to risk behaviour questions which have been shown to be underestimated^{5,13}.

Furthermore, several epidemiological and ethnographic studies have demonstrated that individuals are likely to share needles due to injecting in high numbers, injecting stimulants, first time injectors, drug sharing, and a sense of belonging to a community²⁵⁻²⁹. Other epidemiological studies have shown that frequency of injection is significantly associated with needle sharing and that the two risk activities were interrelated³⁰⁻³⁴. Patrick *et al.* argue that though needle sharing is the most likely mechanism by which HCV transmission occurs among PWID, social desirable bias within that population prevents this from being a measure researchers can use. Instead Patrick *et al.* conclude that the measure of high frequency of injection may be a better, albeit surrogate, measure of the core risk behaviour and was further supported by the work conducted by Thorpe *et al.* in 2002^{35,36}.

We sought to characterize the appropriate distribution of drug injection sharing contacts (links) in the PWID network and to test whether this distribution is consistent with different archetypal network models, which have different implications for the transmission of STTBIs. This work fits into a long term goal to elucidate diffusion dynamics in PWID and other hard-to-reach populations for better prevention strategies. We hypothesize that the frequency of injections, which is a key component of network topology that influences the transmission of infectious disease within a PWID network, is not random and does not resemble a Poisson distribution. Our methods allow us to test this hypothesis and to assess whether the PWID contact network is or is not consistent with several different network archetypes.

4.3 Methods

4.3.1 Data collection

Respondent-driven sampling (RDS) was used to collect data from 407 PWIDs in Ottawa in 2007. Initial key recruits, known as “seeds,” who were well known in the PWID community recruited other associates, friends, relatives and acquaintances who also inject drugs through the distribution of cards with unique codes that allowed researchers to track recruitment³⁷. Each consenting participant was asked to recruit up to three other people. These individuals then recruited their associates and so on, until the desired sample size was reached. Each consenting participant answered a detailed questionnaire about demographics, risk behaviours, knowledge, injection practices and their ego network (an individual’s own direct contacts) administered by a trained interviewer at centrally located offices. Blood was obtained for laboratory tests by finger prick, and the resulting dried blood spots were analyzed for HIV and HCV antibodies by the National Microbiology Laboratory of Canada³⁷.

As typical in RDS, the recruitment process was assumed to resemble a Markov chain in which the stationary distribution is reached through sufficient number of waves or generations of referral, which was key in defining the sample size^{38,39}. Generally, RDS practitioners recommend seven waves or more of recruitment²⁷. Past RDS participation rates in Canadian cities have been approximately 50%. Taking into account that each seed and participant would receive three recruitment cards, 7 seeds were enrolled through promoting the study in community agencies, via word of mouth, and actively seeking out PWID, some of whom had participated in previous similar studies⁴⁰.

Risk factors collected consisted mainly of categorical responses, either based on a binary outcome (i.e. positive or negative for HIV and HCV) or predefined response categories and a

few continuous variables (i.e. such as frequency of injection). In the case of key continuous data, data were grouped for use in univariate analysis but the original continuous form was retained for analysis of the distribution. Frequency of injection was calculated based on the following questions asked in the questionnaire: “In the last 6 months, on how many days did you shoot up”; “In the last 6 months, on the days that you injected, how many times a day, on average do you inject?” and “Total injections in last 6 months?” The number of injections for each PWID was then obtained by averaging out the numeric value, provided by the respondent, for the number of injections within a day, a week and 6 months and categorized into total injections in the past 6 months and for univariate analysis categorized: ≤ 100 times, 100 to 400, 401 to 500, 500 to 959 and ≥ 960 times. Age was categorized as < 21 years, 21-29, 30-40, 50-60 and 61+ years while years of injection was categorized as < 5 years, 5-10, 10-20 and 20+ years.

As discussed previously, the distribution of contacts was identified by using the respondent’s reported frequency of injection (a proxy for the frequency of sharing needles and/or equipment).

4.3.2 Data analysis

Descriptive analysis of the socio-demographics and risk factors of respondents was completed in SAS (version 9.3). Three quarters (77%) of the sample did not respond to questions directly related to sharing needles, which is likely due to social desirability bias, but there were several different questions related to sharing and data from 404/407 PWID were included in the final analysis. Differences between HCV⁺ and HCV⁻ individuals were tested by Fisher’s two-sided exact test at the 5% significance level; Wilcoxon’s rank sum test was used for continuous variables.

The PWID network was visualized using Pajek (version 4.1). Participants are represented by nodes and the links between them represent receipt of study coupons from their recruiters. PWID were further distinguished based on their sex and infection status in graphs generated in Pajek.

The frequency of injections was fitted to the power law (I), Poisson (II), log-normal (III) and exponential distributions (IV), using the maximum likelihood estimation method to estimate the goodness of fit parameters, below. Goodness of fit t-tests based on bootstrapping and the Kolmogorov-Smirnov (KS) statistic, were used to assess the plausibility of the fitted models (using PowerLaw package R version 3.2.3)⁴¹. A distribution was rejected with P-value < 0.1; otherwise, the fit was considered statistically plausible. The following models were used:

$$\text{I. } P(x) = Cx^{-\alpha};$$

The power law function has two different constants; α and C are both > 0 and $x \geq x_{\min}$; $x_{\min}$ is the start value of x at the heavy tail end of the distribution. In equation (1) α usually falls in the range of $0 < \alpha < 3.5$.

$$\text{II. } P_{(x)} = e^{-\lambda} \frac{\lambda^x}{x!};$$

The Poisson function has one parameter λ , known as the rate parameter.

$$\text{III. } P_{(x)} = \frac{1}{x} \cdot \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2 \sigma^2}}$$

The lognormal distribution has two parameters: (1) the mean ($e^{\mu + \frac{1}{2}\sigma^2}$); and (2) the standard deviation σ .

$$\text{IV. } P_{(x)} = \lambda e^{-\lambda x}$$

The exponential distribution has only one parameter, λ , known as the rate parameter.

The cumulative distribution function of the power law; lognormal, Poisson and exponential of the frequency of injections over the 6-month period were also plotted. Vuong's likelihood ratio tests were used to compare relative fit of the distribution of interest to alternative

distributions. A large positive test statistic provides evidence of the superiority of distribution 1 over distribution 2; a large negative test statistic suggests the reverse²⁹.

4.3.3 Ethics

All data and specimens were anonymous, a valuable advantage of the RDS process, and the study was approved by the Ottawa Hospital Research Ethics Board (certification #: 2007045-1H).

4.4 Results

4.4.1 Descriptive Analysis

From the seven initial seeds (resulting in seven components) using RDS, a total of 407 PWID were recruited. Table 1 illustrates the participants' characteristics and compares HCV+ and HCV- PWID. The majority of the PWID sample were males (80%), Caucasian (77%), heterosexual (89%), lacked a high school education (52%), depended on government support for a primary source of income (45%), injected stimulants⁴² (i.e. cocaine, crack) (55%), shared needles and/or equipment (68%), did not have stable housing (i.e. living in motels, on the street, shelters, etc.) (54%), injected 100 times or less within the 6-month study period (37%) and based on the RDS process, most likely to be included in the second (out of seven) recruitment chains (62%).

There were significant differences in age ($p < 0.0001$), drug of choice for injection ($p = 0.007$), HIV status ($p < 0.0001$), shared needles and/or equipment, total number of injections within the 6-month period ($p < 0.0001$) and the recruitment chain based on the RDS process ($p = 0.006$) between HCV+ and HCV- PWID in bivariate analysis.

Descriptively, 68% of PWID admitted they share needles and/or equipment, 363 out of 407 (89%) used a needle exchange program (NEP) and 342 (84%) used the NEP in Ottawa (data not shown). However, table 2 shows no difference in use of the NEP in the 6-month HCV infection ($p = 0.16$).

A total of 246 PWID were positive for HCV (60%), 193 (78%) of whom were males and of the 41 PWID who were HIV positive, 19 (46%) were males.

4.4.2 RDS Network

The full network consisting of the 7 different components is illustrated in Figure 2 while Figure 3 shows the RDS process which took place in component 3 ($n=12$).

4.4.3 Network Topology

Figure 4 plots the number of PWID and the frequency with which they injected in the last 6 months, ranging from 1 to 3600. The majority of PWID injected less than 200 times in 6 months while a very small number injected 3,600 times. The frequency of injection revealed a skewed distribution, with a skewness indicator of 3.37. The data were assessed for normality using the Shapiro-Wilk test. Results indicated the data were non-normal and heterogeneous (Shapiro-Wilk test, homogeneity Levene's test, respectively). However, after the data was log transformed, the distribution met the requirements of both homogeneity of variance and normality (p values < 0.05), indicating the lognormal distribution may be an appropriate model for the PWID data.

The log-log plot of the cumulative distribution function (CDF) of the power law and lognormal distribution appeared to fit the data, whereas the Poisson and exponential distribution did not (Figure 5 and Table 3). The decaying coefficient (α) of the power law distribution is

2.24, which indicates a possible scale free distribution (Table 3). Both the power law and lognormal distributions were similar to the empirical data after bootstrapping to generate the $x_{\min}$ and calculating the KS test ($p = 0.360$ and 0.362 , respectively) (Table 4). However, the Poisson and the exponential distribution differed significantly from the observed data, $p \leq 0.05$ (Table 3). Both the power law distribution and the lognormal distributions were closer to the true distribution when compared with the Poisson ($p < 0.01$) and the exponential distributions ($p = 0.09$ and 0.06 , respectively) (Table 4). However, when the power law distribution was compared to the lognormal distribution, both distributions may fit ($p = 0.58$.)

Table 1: Possible risk factors of injecting drugs associated with HCV status in the PWID network in Ottawa, Canada, 2007.

Possible Risk Factors	Sample size (n)	Infectious Status		p value †
		HCV ⁺	HVC ⁻	
*Age				
Median (Range)	406	42.7 (18.1 - 63.9)	36.9 (16 - 62.8)	<0.0001
Gender				
Male	401	193 (79%)	126 (80%)	1.000
Female		51(21%)	31(20%)	
Ethnicity				
Caucasian	402	191(78%)	120 (77%)	0.7235
Black		8 (3%)	5(3%)	
Aboriginal		26 (11%)	21(13%)	
Other		20 (8%)	10 (7%)	
Education				
Did not complete high school	402	130 (53%)	79 (50%)	0.7615
Completed high school		53 (22%)	34(22%)	
Post-secondary		62(25%)	44(28%)	
Source of Income				
Regular Employment	404	47 (19%)	46 (29%)	0.7615
Government Support		112(46%)	70(45%)	
Stigmatized activities		50 (20%)	18 (11%)	
Other		37 (15%)	24 (15%)	
Drug Choice of Injection ^Ω				

Depressants Stimulants	404	125 (51%) 121 (49%)	58 (37%) 100 (63%)	0.0074
HIV Status				
HIV ⁺ HIV ⁻	404	40 (16%) 206 (84%)	1 (1%) 157 (99%)	<0.0001
Sexual Orientation				
Heterosexual Homosexual Bisexual Asexual	397	218 (91%) 7 (3%) 13 (5%) 3 (1%)	137 (87%) 1 (1%) 17 (11%) 1 (1%)	0.1828
Shared Needles/Equipment				
Yes No	404	185 (75%) 61 (25%)	89 (56%) 69 (44%)	<0.0001
Housing				
Stable Unstable (Shelter, Street, motel)	382	101 (44%) 128 (56%)	76 (50%) 77 (50%)	0.2967
Total Number of Injections in the last 6 months				
≤ 100 x 100 to 400 x 401 to 500 x 500 to 959 x ≥ 960 x	407	73 (28%) 93 (38%) 15 (7%) 42 (17%) 26 (10%)	77 (49%) 51 (32%) 6 (4%) 19 (12%) 5 (3%)	<0.0001
End tail of a Power law Distribution (960 + injections in the last 6 months)				
Yes No	407	26 (11%) 223 (89%)	5 (3%) 153 (97%)	0.0064
Recruitment chain				
1 2 3 4 5 6 7	407	156 (63%) 75 (31%) 11 (4%) 3 (1%) 1 (0.5%) 1 (0.5%) 0 (0%)	97 (60%) 51 (33%) 1 (0.5%) 10 (6%) 0 (0%) 0 (0%) 1 (0.5%)	0.0060

* Median age presented and Wilcoxon's rank sum test used to assess differences. † Fisher's two-sided test ^Ω Stimulants include Cocaine, Methamphetamine, Ecstasy, Amphetamines, MDA, Crack, Ritalin; and depressants include the following: Dilaudid, hydromorphone, Benzodiazepines, Percocet, Heroin, Morphine, Oxycontin, Fentanyl, Alcohol, Marijuana, Methadone, Solvents, Barbituates, Tylenol with codeine³⁰

Table 2: The HCV positivity and use of the needle exchange program (NEP) in the 6-month prior to interview

Infectious status	Responses to use of NEP*			Total	<i>p</i> value
	<i>Rarely</i>	<i>Occasionally</i>	<i>Frequently</i>		
HCV ⁻	8	52	61	121	0.1580
HCV ⁺	9	76	131	216	
Total	17	128	192	337	

*Rarely= includes only one time or never, Occasionally= not every week, Frequently= everyday, regularly once or twice or 3 to 4 times week

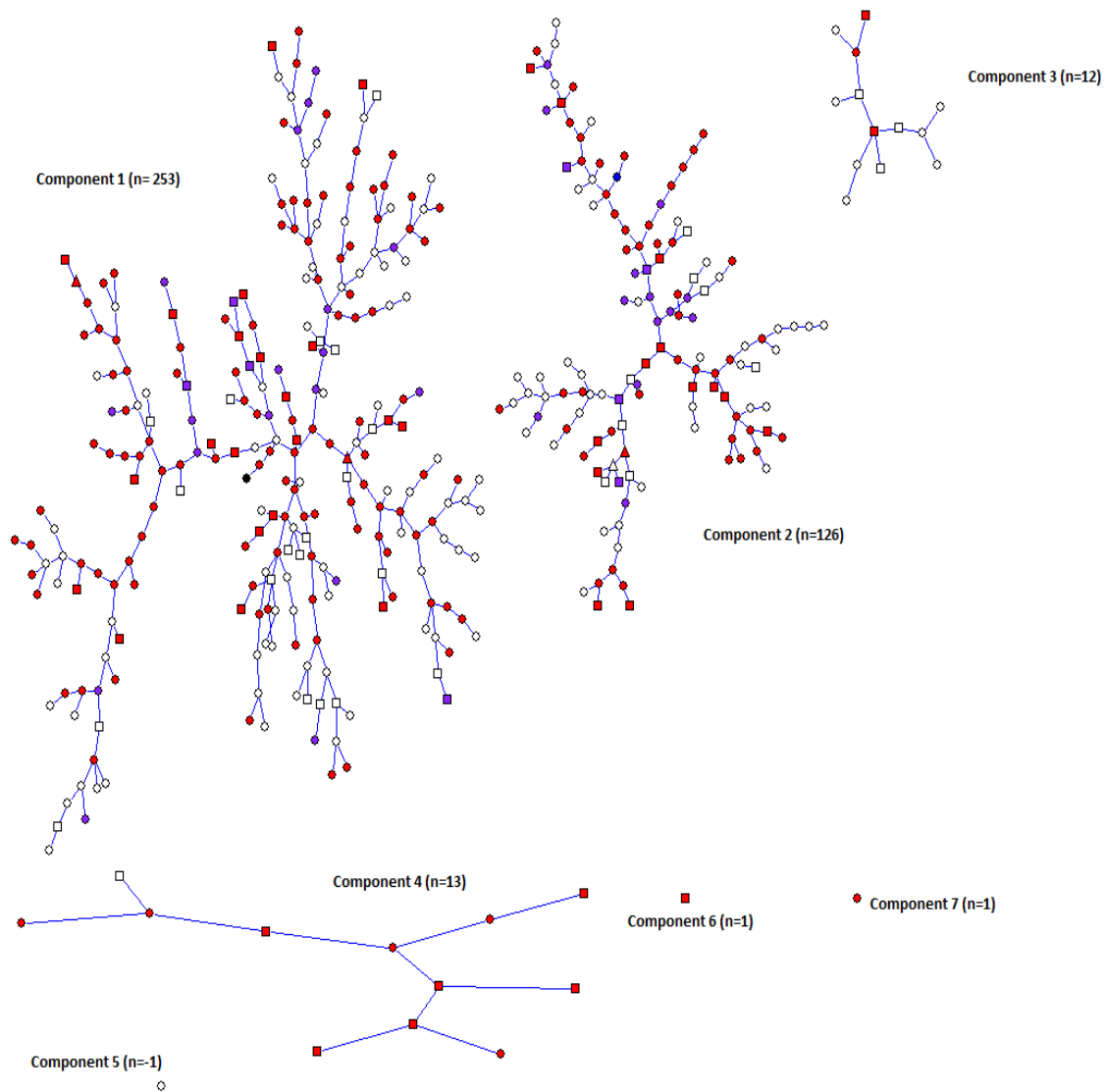


Figure 2: Visualization of the 7 components (generated by the seeds) consisting of 407 IDU's identified by RDS in Ottawa, Canada, 2007; where square = male, circle = female, triangle = transgender, red = HCV+, blue = HIV+, purple = Co-infection, black = missing and white = not infected

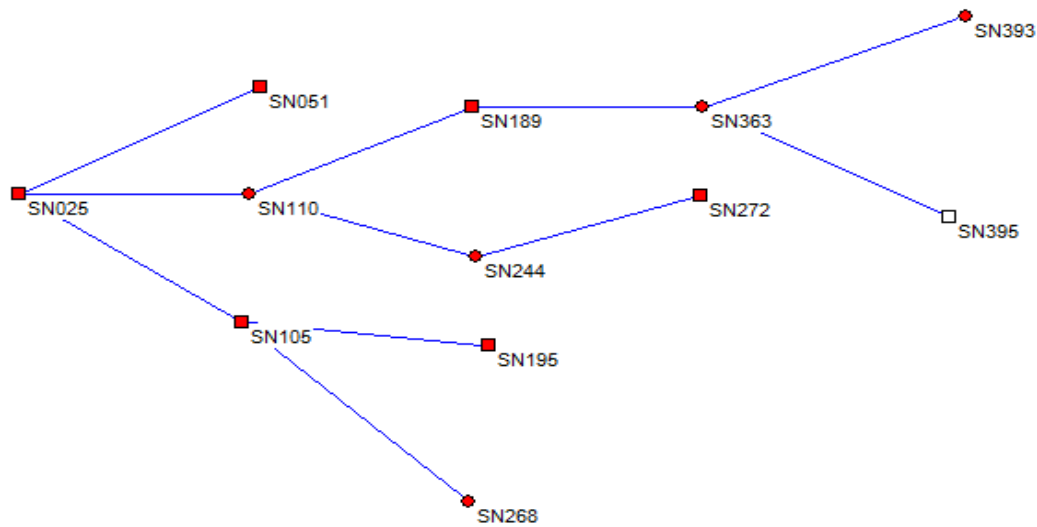


Figure 3: Visualization of the third component of the network (n=12), where square = male, circle = female, red = HCV+, and white = not infected

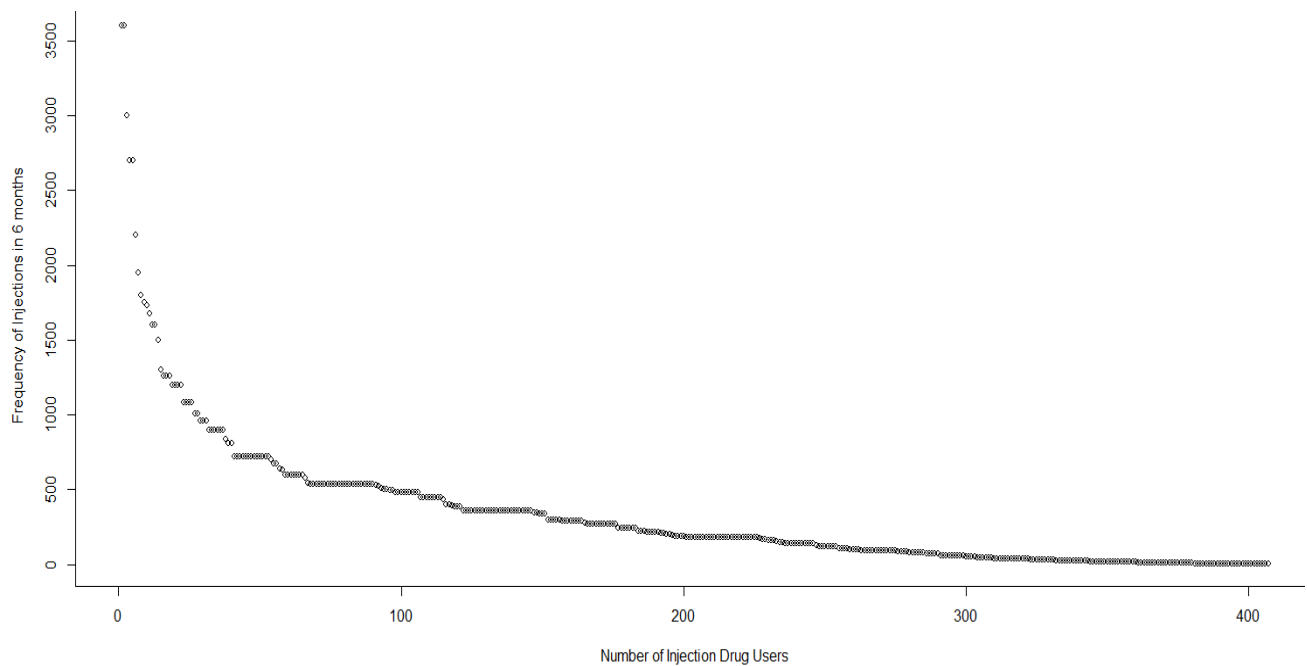


Figure 4: Plotted data of the total frequency of injection in the 6-month period prior to interview among 407 IDUs, Ottawa, Canada

Table 3: Parameters of the various distributions used to model the frequency of injection in an Ottawa network consisting of 407 IDU's, Canada, 2007

Variable of interest	Basic Statistical Measure				Parameters		
	N (sample size)	μ	σ	$x_{\max}$	$x_{\min}$	α	p-value
Discrete Power law	407	346	477	3600	960	2.24	0.36
Lognormal					$x_{\min}$	μ	σ
					110	5.8	0.87
Poisson					$x_{\min}$	λ	p-value
					1800	269	0.05
Exponential					$x_{\min}$	λ	p-value
					33	0.003	0.05

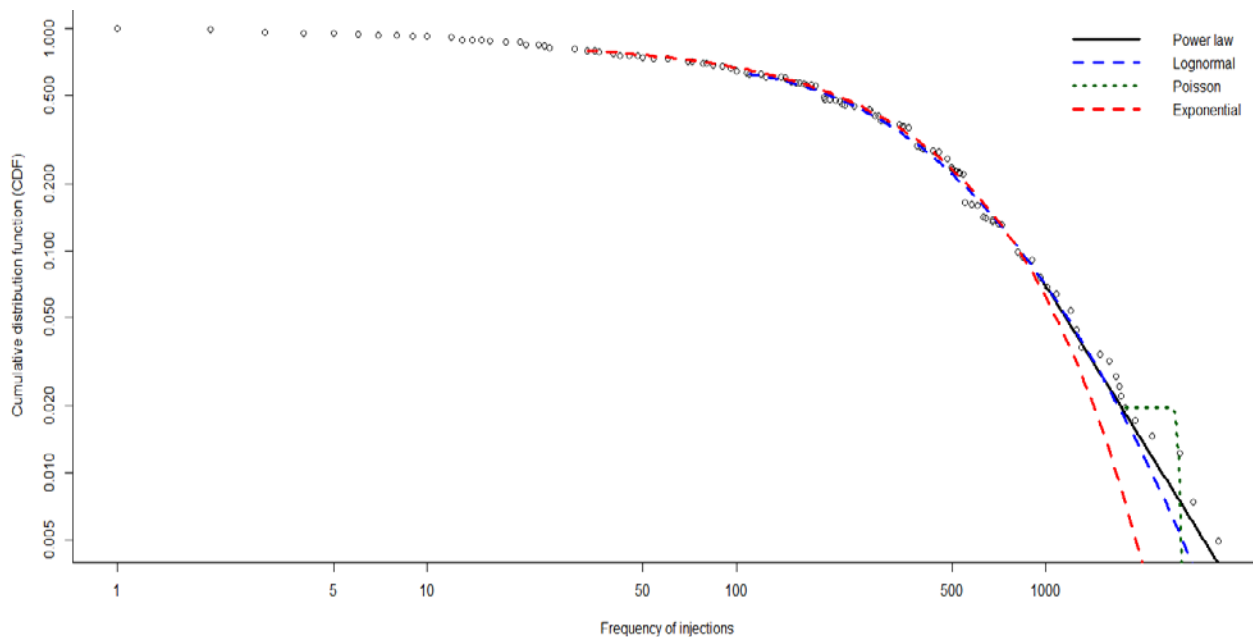


Figure 5: Log-log plot of the cumulative distribution for frequency of injection in the 6-month period prior to interview among 407 IDU's, Ottawa, Canada, 2007, fitted to several types of distributions including the power law (black line), lognormal (dashed blue line), Poisson (dotted green line) and exponential (dashed red line)

Table 4: Fitted power law distribution to the frequency of injection in the 6-month period prior to interview among 407 Ottawa IDU's compared to other fitted distributions using the Vuong test, cut-off $p > 0.1$

Distribution 1		Distribution 2	Vuong test statistic	P value
Power Law	vs.	Poisson	3.64	< 0.01
Power Law		Exponential	1.6	0.09
Power Law		Lognormal	-0.55	0.58
Lognormal		Poisson	3.64	< 0.01
Lognormal		Exponential	1.88	0.06

*Positive Vuong test statistic indicates favor for Distribution 1; Negative Vuong test statistic indicates favor for Distribution 2; p value > 0.10 indicates that either distribution can be closer to the actual distribution

4.5 Discussion

As hypothesized, the Poisson distribution, which represented the ER model, did not fit the distribution of frequency of injections per person among the Ottawa PWID network. Our findings are similar to those of a pilot study conducted among PWID in Winnipeg, Manitoba (Canada) in 2004 (Pelude *et al.*) and in Brooklyn, New York (USA) 1991 to 1993 (Dombroski *et al.*)^{13,19}.

Pelude *et al.* also used the frequency of injection as a proxy to capture the potential multiple routes of disease transmission an injector may be subjected to. They demonstrated that the power law may fit the distribution of the frequency of injecting drugs among PWID in Winnipeg, MB¹⁹. Dombroski *et al.* developed the mean total injection variable, which incorporates the in-degree, out-degree and the number of injecting partners from the respondents and noticed the distribution was similar to the BA model¹³.

Our results indicate the frequency of injections of PWID can conform to a BA distribution. This result was supported by the fact that the decay coefficient of our distribution was 2.24, within the range between 2 and 3 that Barabási and Albert specify²⁰. In addition, several studies have mathematically demonstrated that prophylactic strategies are ineffective in eradicating infectious diseases in networks with decay co-efficient below 3, because there is

always one individual within the population with enough contacts to allow the diseases to keep thriving^{9,18,20,43-46}. The potential hubs of PWID within our network demonstrate that this population may represent the sort of reservoirs lacking an epidemic threshold, requiring more specific knowledge about the hubs and the network structure in order to control and prevent disease transmission.

Our analysis clearly rejects the ER model, but it does not conclusively demonstrate the superiority of the BA model. Although the power law distribution associated with the BA model may fit our data, we cannot rule out whether it is a better fit than the lognormal distribution (see Table 4). The BA model, or the scale-free network, relies on growth and preferential attachment. These two phenomena could be valid assumptions in a PWID network however, they are difficult to assess using data from a cross-sectional study, as it is a “snapshot” in time. We have demonstrated that the frequency of injection appears to follow a group of heavy tailed distributions which include the power law, the lognormal, and stretched exponential distributions, such as the Weibull^{10,20,21,43-48}. The main characteristics shared by all heavy tailed distributions are the distinctive shape of the tail^{10,20,21,43-48}.

The dilemma in not being able to identify whether our data fits a lognormal or a power law distribution is not surprising, as others have had similar difficulties with other real life data, in different fields, such as biology, ecology and economics^{23,49}. Lognormal and power law distributions are quite similar in their skewed shape and in plotting the log-log plot of the CDF, both appear as a straight line. The main difference between the two distributions is that under natural conditions, the lognormal distribution has finite mean and variance; whereas the power law does not^{24,43,50,51}. It becomes even more difficult when both distributions are considered

appropriate fits to the data and the generative process that leads to the distribution could result in either the power law or the lognormal^{12,22,24,50,51}.

In the biological field, lognormal distributions are generally explained by the multiplicative process. The multiplicative process is defined as the law of proportionate effect, which states that the change in the variable of interest at any step of the process is a random proportion of its previous value^{10,50-52}. Variations in animal and plants species, organism growth and population dynamics have been explained by multiplicative processes^{10,50-52}. To further elaborate, a variable of interest that is a result of many small independent factors as a multiplicative product can be modeled as a lognormal distribution. This implies that the data of interest, in this case the frequency of injection, are normal when expressed as a function of the logarithm of the frequency of injection and that large variances in the data are attributed to ‘exceptional’ values that are not considered to be outliers.

4.6 Conclusion

Our goal in attempting to fit as precise a curve as possible, was to; 1) clearly identify possible candidate models to fit the curve of the empirical data; and 2) hypothesize about the processes which result in that shape (generating function). Ideally we wished to obtain a curve or curves, which fit well. However, more important is the mechanism by which the graph (or network) evolves. The nature of injection drug use is that people usually increase their doses, and/or their frequency of injection as they become more inured to the drug, resulting in increased needle sharing. This is similar to preferential attachment in that those who use the drug, need more, which has implications not only for preventing infection, but also for better understanding the context of drug overdoses^{8,13,19,20,53}. By identifying the distribution of individuals engaging

in risky behaviors within the PWID network, strategic interventions targeting those in the ‘heavy tail’ of the distribution are more likely to be successful and efficient, than interventions with random members of the network. The individuals in the heavy-tail distribution are likely the ones sharing needles and/or equipment with a significant number of individuals that compromises the PWID network. It would be far more strategic and cost effective to target interventions to PWID that are well intertwined within PWID community. These individuals have the ability to influence a great proportion of other PWID, rather than PWID who are not very involved, inject with few individuals, and not considered an integral part of the PWID network.

Further investigation into the topological property of PWID networks, including this one, using other models that incorporate the formation of clusters, such as those recently developed for defining network clustering coefficients and random graph models, is needed to test whether the Ottawa PWID network can be classified as a small world network.

4.7 References

1. UNAIDS. Prevention gap report. . 2016.
2. Averting HIV and AIDS. People who inject drugs, HIV and AIDS . *AVERT*. 2017.
3. Sacks-Davis R, Horyniak D, Grebely J, Hellard M. Behavioural interventions for preventing hepatitis C infection in people who inject drugs: A global systematic review. *Int J Drug Policy*. 2012;23(3):176-184.
4. Taylor S. 'A state of emergency': HIV rates in sask. rise again in 2016 ' . *CBC News*. 2017. Available from: <https://www.cbc.ca/news/Canada/saskatchewan/hiv-rates-rise-saskatchewan-1.4101145>.
5. Millson P, Leonard L, Remis RS, Strike C, Challacome L. Injection drug use, HIV and HCV infection in Ontario: The evidence 1992 to 2004. 2005.
6. Bearman PS, Moody J, Stovel K. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*. 2004;110(1):44-91.
7. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. *Clin Microbiol Infect*. 2013;19(11):999-1005.
8. Bansal S, Grenfell BT, Meyers LA. When individual behaviour matters: Homogeneous and network models in epidemiology. *J R Soc Interface*. 2007;4(16):879-891.
9. Miller JC, Volz EM. Incorporating disease and population structure into models of SIR disease in contact networks. *PLoS One*. 2013;8(8):e69162.
10. Hein O, Schwind M, König W. Scale-free networks. *WIRTSCHAFTSINFORMATIK*. 2006;48(4):267-275.
11. Liljeros F, Edling CR, Nunes Amaral LA. Sexual networks: Implications for the transmission of sexually transmitted infections. *Microbes Infect*. 2003;5(2):189-196.
12. Wasserheit JN, Aral SO. The dynamic topology of sexually transmitted disease epidemics: Implications for prevention strategies. *J Infect Dis*. 1996;174 Suppl 2:S201-13.
13. Dombrowski K, Curtis R, Friedman S, Khan B. Topological and historical considerations for infectious disease transmission among injecting drug users in bushwick, brooklyn (USA). *World J AIDS*. 2013;3(1):1-9.
14. Friedman SR, Flom PL, Kottiri BJ, et al. Drug use patterns and infection with sexually transmissible agents among young adults in a high-risk neighbourhood in new york city. *Addiction*. 2003;98(2):159-169.

15. Jolly AM, Muth SQ, Wylie JL, Potterat JJ. Sexual networks and sexually transmitted infections: A tale of two cities. *Journal of Urban Health : Bulletin of the New York Academy of Medicine*. 2001;78(3):433-445.
16. Public Health England, Health Protection Scotland, Public Health Wales, and Public Health Agency Northern Ireland. Shooting up: Infections among people who inject drugs in the UK 2012. 2013.
17. Newman ME, Watts DJ, Strogatz SH. Random graph models of social networks. *Proc Natl Acad Sci U S A*. 2002;99 Suppl 1:2566-2572.
18. Barabási A, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512.
19. Pelude L. *Networks among injection drug users: Random or scale-free?* [Masters]. University of Ottawa; 2007.
20. Barabási A. Scale-free networks: A decade and beyond. *Science*. 2009;325(5939):412-413.
21. Fazekas I, Porvaszsnik B. Scale-free property for degrees and weights in a preferential attachment random graph model. *Journal of Probability and Statistics*. 2013;2013.
22. Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. *Internet Math*. 2003;1(2):226-251.
23. Gabaix X. Power laws in economics and finance. *Annu Rev Econ*. 2009;1(1):255-294.
24. Reed WJ, Jorgensen M. The double pareto-lognormal Distribution—A new parametric model for size distributions. *Communications in Statistics - Theory and Methods*. 2004;33(8):1733-1753.
25. Duncan I, Curtis R, Reyes JC, Abadie R, Khan B, Dombrowski K. Hepatitis C serosorting among people who inject drugs in rural puerto rico. *Preventive medicine reports*. 2017;6:38-43.
26. Wiessing L, Ferri M, Grady B, et al. Hepatitis C virus infection epidemiology among people who inject drugs in Europe: A systematic review of data for scaling up treatment and prevention. *PLoS One*. 2014;9(7):e103345.
27. MacArthur GJ, van Velzen E, Palmateer N, et al. Interventions to prevent HIV and hepatitis C in people who inject drugs: A review of reviews to assess evidence of effectiveness. *Int J Drug Policy*. 2014;25(1):34-52.
28. Pasa MK, Alom KR, Bashri Z, Vermund SH. Sharing of needles and syringes among men who inject drugs: HIV risk in northwest Bangladesh. *PLoS One*. 2016;11(2):e0148276.

29. Grund JP, Kaplan CD, Adriaans NF. Needle sharing in the Netherlands: An ethnographic analysis. *Am J Public Health*. 1991;81(12):1602-1607.
30. Mandell W, Vlahov D, Latkin C, Oziemkowska M, Cohn S. Correlates of needle sharing among injection drug users. *Am J Public Health*. 1994;84(6):920-923.
31. Chen X, Zhu L, Zhou YH, et al. Factors associated with needle sharing among people who inject drugs in yunnan, china: A combined network and regression analysis. *Infect Dis Poverty*. 2016;5(1):73-016-0169-y.
32. Golub ET, Strathdee SA, Bailey SL, et al. Distributive syringe sharing among young adult injection drug users in five U.S. cities. *Drug Alcohol Depend*. 2007;91 Suppl 1:S30-8.
33. Wood E, Tyndall MW, Spittal PM, et al. Unsafe injection practices in a cohort of injection drug users in Vancouver: Could safer injecting rooms help? *CMAJ*. 2001;165(4):405-410.
34. Shaw SY, Shah L, Jolly AM, Wylie JL. Determinants of injection drug user (IDU) syringe sharing: The relationship between availability of syringes and risk network member characteristics in Winnipeg, Canada. *Addiction*. 2007;102(10):1626-1635.
35. Patrick DM, Tyndall MW, Cornelisse PGA, et al. Incidence of hepatitis C virus infection among injection drug users during an outbreak of HIV infection. *Can Med Assoc J*. 2001;165(7):889.
36. Thorpe LE, Ouellet LJ, Hershow R, et al. Risk of hepatitis C virus infection among young adult injection drug users who share injection equipment. *Am J Epidemiol*. 2002;155(7):645-653.
37. Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Probl*. 1997;44(2):174-199.
38. Heckathorn D.D., Semaan S., Broadhead R.S., Hughes J.J. Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18-25. *AIDS Behav*. 2002;6(1):55-67.
39. Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*. 2008;24(1):79-97.
40. Pilon R, Leonard L, Kim J, et al. Transmission patterns of HIV and hepatitis C virus among networks of people who inject drugs. *PLoS One*. 2011;6(7):e22245.
41. Gillespie C. Fitting heavy tailed distributions: The powerLaw package. *Journal of Statistical Software*. 2015;64(2):1-16.
42. Halton Region. Types of drugs.
http://www.halton.ca/living_in_halton/public_health/health_wellness/drugs_alcohol_tobacco/types_of_drugs/. Updated 2015. Accessed 01/15, 2017.

43. Alstott J, Bullmore E, Plenz D. Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS One*. 2014;9(1):e85777.
44. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data . *SIAM Review*. 2009;51(4):661-703.
45. Cooke RM, Nieboer D. Heavy-tailed distributions: Data, diagnostics, and new developments . *Resources for the Future Discussion Paper*. 2011:11-19.
46. Slater N, Louzoun Y, Gragert L, Maiers M, Chatterjee A, Albrecht M. Power laws for heavy-tailed distributions: Modeling allele and haplotype diversity for the national marrow donor program. *PLoS Comput Biol*. 2015;11(4):e1004204.
47. Eguiluz VM, Klemm K. Epidemic threshold in structured scale-free networks. *Phys Rev Lett*. 2002;89(10):108701.
48. Virkar Y, Clauset A. Power-law distributions in binned empirical data. *Annals of Applied Statistics*. 2014;8(1):89-119.
49. Sornette D, Cont R. Convergent multiplicative processes repelled from zero: Power laws and truncated power laws. *Journal de Physique I*. 1997;7(3):431-444.
50. Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: Keys and clues. *BioScience*. 2001;51(5):341-352.
51. Koch AL. The logarithm in biology: II. distributions simulating the log-normal. *Journal of Theoretical Biology*. 1968;23(2):251-268.
52. Champernowne DG. A model of income distribution. *The Economic Journal*. 1953;63(250):318-351.
53. Jones JH, Handcock MS. Sexual contacts and epidemic thresholds. *Nature*. 2003;423:605.

5.0 Social Contacts of People who Inject Drugs in Ottawa Exhibit a Small-World Network

The previous manuscript evaluated whether the PWID RDS Ottawa data exhibited a scale-free distribution by assessing the distribution of frequency of injections among PWID, which was used as a proxy of needle sharing. Although we demonstrated that the distribution of the frequency of injections among PWID fit a heavy tailed distribution rather than a Poisson distribution, it was not possible to conclude the distribution fit a scale-free model.

The structure of both social and contact network is also a key component of the risk environment for members of hard-to-reach populations and also has important implications for diseases transmission and health behaviours. This second manuscript focuses on how PWIDs egocentric data can assess two additional important aspects of network structure, clustering and average path lengths. Networks that have high clustering and short average path lengths are classified as a small-world network. Risk networks that are also small-world networks can lead to more redundant paths increases the likelihood of disease transmission and alters the relationship between concurrency and epidemic potential. However, small-world properties, clustering and short average path length, can also be used to make positive changes within risk networks. Identifying key individuals and/or groups for interventions can increase individual likelihoods of engaging in and spreading health-promoting behaviours.

The traditional clustering coefficient associated with small-world networks was originally intended for complete networks. As networks constructed with RDS data are incomplete, this study demonstrates how social contacts of PWID RDS data can be used to determine the small-worldness of the network, with an appropriate clustering coefficient, and its implication.

Abdesselam, K.¹, Verdery, A.², Pelude, L.¹, Momoli, F.¹, and Jolly, A.M.¹

¹ School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa

² Department of Sociology and Criminology, Pennsylvania State University

5.1 Abstract

The aim of this study was to identify an appropriate network structure, using social contacts and other information that would best reflect the disease transmission topology of people who inject drugs (PWID) in Ottawa. In our previous study, we fitted several distributions, using the frequency of injections obtained from PWID, and ruled out the Poisson and exponential distribution. However, there was not sufficient evidence to conclude that the network exhibited a scale-free distribution. Scale-free networks are a specific type of network that can be considered a sub-category of more general social network structures defined as small world. Small-world networks have short average path lengths and subgroups of people who are more densely connected than others (clustering), which facilitate the transmission of infectious disease by providing several routes for successful disease transmission. Using a respondent-driven sample (RDS) of PWID in Ottawa in 2007, we used two different methods of calculating clustering to assess whether the PWID network can be categorized as a small-world structure. Matlab was used to calculate average path length and the clustering coefficient by using the number of PWID each participant knows (alters), and the connections between them. We constructed a random model with the same number of nodes as the observed RDS network, with links joining participants and their alters randomly and compared it to the observed network. The network of PWID in Ottawa 2007 can be classified as a small world network rather than a random network which has implications for efficient interventions.

5.2 Introduction

In Canada, the estimated population of people who inject drugs (PWID) ranges from 50,000 to 110,000 ^{1,2}. This group has the highest Hepatitis C (HCV) prevalence (~60%) and bears the largest proportion of new HCV infections of any risk group in Canada ¹. In Ottawa alone, it is estimated that between 1,200 to 5,600 individuals injected drugs in 2014 ³⁻⁵. Like other places the increased risk of PWID over other groups is associated with sharing needles and equipment^{2,3}. In this study, we use empirical data on social networks to understand the contact network structure within the PWID population in Ottawa in order to provide insight into transmission which will help improve existing control programs.

The statistical properties of networks - especially the importance of degrees (the number of contacts an individual has) – and hence network analytic methods, have been of particular interest in the area of infectious diseases^{6,7}. The structure of the network has profound effects on the transmission of infection, affecting speed of spread, maintenance of endemicity, probability of an epidemic, and elimination of infection ⁸⁻¹³. Because of the importance of network structure in determining disease dynamics, researchers have sought to define several classes of archetypal network models that characterize a broad range of networks and which have different implications for epidemic potential. Identifying the network type that most closely corresponds to the contact networks of PWID will not only help understand how local interactions contribute to global risk environments that affect the transmission and persistence of communicable diseases but doing so can also provide researchers with a model of which individuals to target within the network for appropriate interventions¹³⁻¹⁶.

In our previous study, we highlighted that hubs, consisting of individuals with a high number of contacts (degree) and/or high numbers of injections, skewed the degree distribution of the network, perhaps indicating that the network was of the small world type. We investigated whether the network was one of those found empirically in the real world. Specifically, we investigated whether the PWID contact network can be characterized as a scale-free network, which is distinguished by a degree distribution that fits a power-law form (regardless of scale)¹⁷⁻²¹. We fitted Poisson, exponential, power-law, and lognormal distribution and ruled out the Poisson distribution as a good fit with our RDS data. Although our results conclusively ruled out the Poisson distribution, which is consistent with the network possibly being scale-free, we were not able to rule out alternative network models including the lognormal model. In this paper we build on that work by determining whether our network exhibited other features of small world networks, namely short average path lengths and high clustering coefficients, which are measures of the extent to which people's acquaintances, friends, or other network contacts know one another.

Networks formed in the real world tend to display three properties: (1) a distinctive degree distribution, (2) short average path lengths, and (3) local clustering, indicated by the clustering coefficient. The degree distribution is the probability distribution of the degrees over a network^{9,16,18,22,23}. Typically, nodes with larger degrees are more influential than the average node is within the network. The average path length is the average number of contacts existing in the shortest chain connecting any two individuals within the network^{9,16,18,22,23}. Local clustering in PWID networks reflects the extent to which individuals inject drugs in pairs or in groups, in which everyone is injecting drugs with each other. Early stylistic network models include regular lattices and random networks, which lack one or more of these three properties^{17,18,24}.

The degrees of individuals in regular or nearly regular lattice networks are more or less the same for each individual within the network, implying low heterogeneity and very little randomness in terms of individual interaction as the probability of two random individuals within the network knowing each other is very slim^{17,18,24}. Key properties in regular lattice networks are long average paths and high clustering, due to individuals within the network being densely connected in sub groups^{17,18,24}. A more widely used model is the random network, which is generated by uniform probabilities of tie formation between each pair in the network; for PWID contact networks, ties in this case represent the types of connections where transmission is possible, including sharing needles or using someone else's needle (which may have been picked up on the ground or otherwise acquired without directly sharing it). This model also has low heterogeneity, since individuals within the network tend to have the same number of connections, short average path, low clustering and a degree distribution resembling the Poisson distribution^{17,18,24,25}. In 1998, Watts and Strogatz (WS) wanted to construct a network model with a short average path, as exists in random networks, but with relatively large clustering coefficients more typical of regular networks¹⁰. They termed their model the small world (SW) network model, which is now considered a fundamental class of networks and is the most frequently used to describe real world networks^{10,26}. A subgroup of small world models - scale-free networks - incorporate both heterogeneity and randomness, the networks can have various forms of clustering within the various subgroups of the network, which means that SW networks can also be scale-free^{18,20,21,26,27}. Scale free networks include hubs and are formed through (1) network growth and (2) preferential attachment^{28,29}.

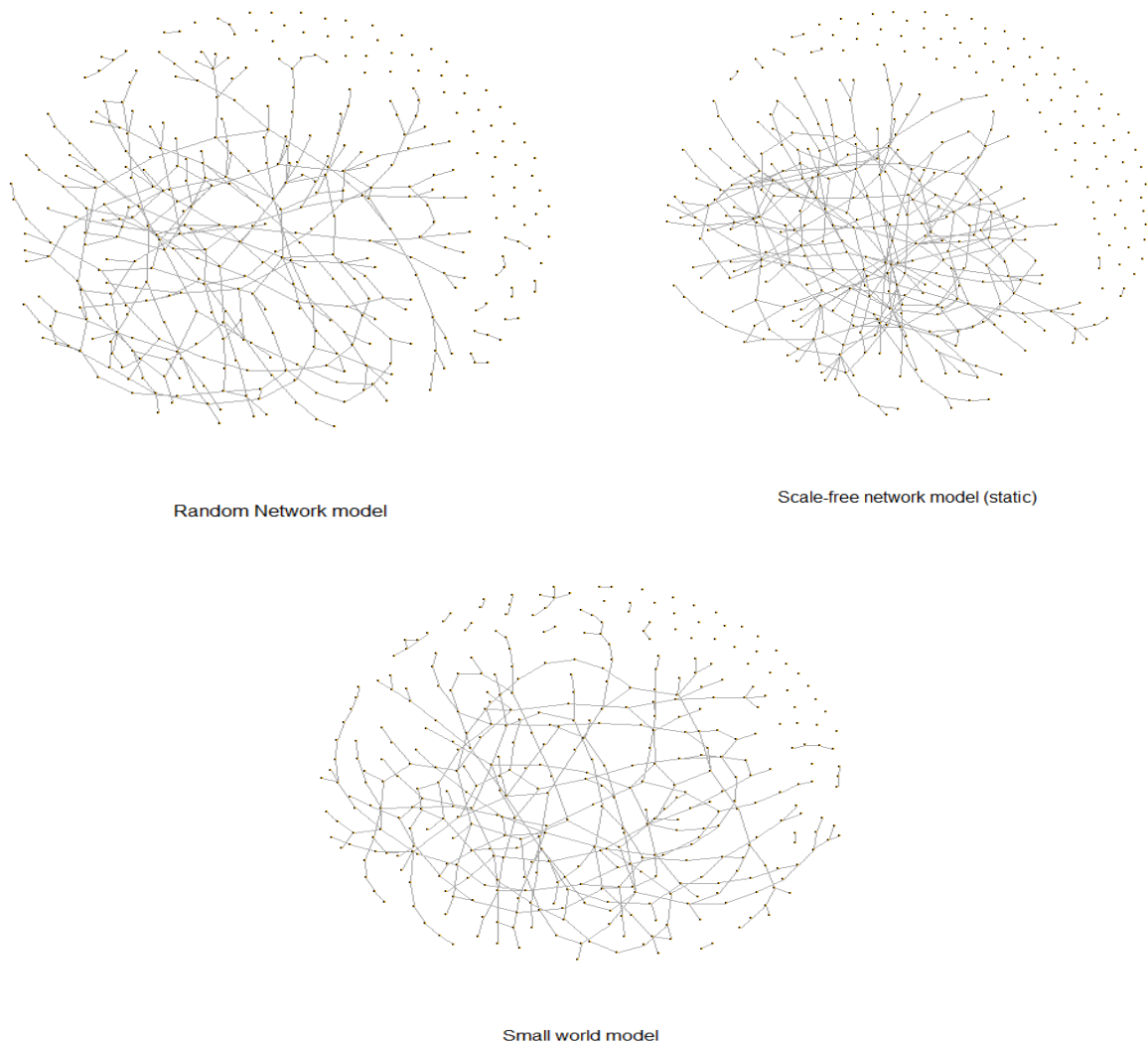


Figure 6: Examples of different types of networks using the Ottawa PWID RDS data with the same number of nodes and links (R Studio, v 1.1456)

The objective of this study was to identify whether the PWID network in Ottawa could be classified as a small-world network. The transmission of infectious disease within a SW network is facilitated by small number of contacts on a path between any two given individuals, and due to clustering, provides multiple routes which facilitate disease transmission^{10,13,30-32}. Links

clustered around hubs render the network robust to random intervention with individuals. As such, interventions strategically designed for highly connected individuals and their local, clustered subgroups are more likely to be efficient and successful in the long term ^{33,34}.

5.3 Data and Methods

We calculated the average path length, the clustering coefficient (CC), and the degree distribution, and used the method developed by Humphries & Gurney (Humphries & Gurney, 2008) to determine whether the Ottawa PWID network can be classified as SW. We use both the WS clustering coefficient (CC^{WS}) ^{10,30} and the Hardiman and Katzir CC estimator tailored to respondent-driven sampling (RDS), CC^{RDS} , ^{13,31,35} to measure the small-world-ness of the Ottawa PWID network.

Respondent-driven sampling was used to collect data from 407 PWIDs in Ottawa in 2007. Initial key recruits, known as “seeds”, were well-known in the PWID community and recruited other associates, friends, relatives and acquaintances who also inject drugs, via the distribution of cards with unique codes, which allowed researches to track recruitment ³⁶. Each consenting participant was asked to recruit up to three other people that are among their network and who inject drugs. These individuals then recruited their associates and so on, until the desired sample size was reached. Each consenting participant answered a detailed questionnaire, administered by a trained interviewer at centrally located offices. The questionnaires collected information on demographics, behaviors, knowledge, injection practices and ego network (an individual’s own direct contacts). Information regarding the sharing of needles and equipment was collected from RDS participants in several different ways. First respondents were asked directly if they shared needles with the individuals they were affiliated with in the RDS process

and their own network: “Have you ever shared a needle with this person?”. Participants were also asked in general if they shared equipment with each of their alters enumerated in the network part of the questionnaire: “Have you ever shared other injecting equipment (such as filters, cookers, water, tourniquets) with this person?”

The size of the ego’s network, including ties to PWID and non-PWID, was calculated based on information gathered on the participant’s contacts. An interaction grid (Figure 7), included in the questionnaire, required that the participant use a code or false name to identify up to 20 members of their personal social network with whom they injected drugs within the last 30 days. Respondents then used the grid to indicate which individuals within their network know each other. We assume that the clustering within the peer network is a proxy for clustering in the contact network.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				

The interaction grid gives us the egocentric networks (each individual participant's social network) and the card system gives us the socio-metric network).

Figure 7: Grid Interaction used in the RDS questionnaire to identify the degree of each participant, ego-network and the clustering coefficient.

Respondents were asked to provide information on their contacts (alters), including; demographics (age, sex and ethnicity), the person's relationship to them, whether they believe this person has injected drugs in the last 6 months, if they have ever shared needles with this person, whether they ever shared other injecting equipment with the person, and if so, how many times. As described in the previous paper we did not use the simple question "Have you ever shared a needle with this person?"; as people tend to underestimate their injection drug contacts, may not know who used equipment prior to them using it, or borrow or lend equipment from or to a friend without knowing any other users. We have also found many people are reluctant to report sharing needles or equipment at all, which underestimates their risk of exposure relative to more specific questions related to interactions with each alter in their personal networks³⁷. As such risk of exposure was evaluated using two sets of questions; the first in the survey itemizing

the behaviours of the respondent and the second in the interaction grid, where interactions with each alter were enumerated. Univariate analysis of the ego network of the RDS participants was conducted using SAS (version 9.3).

The average path length was calculated after networks were constructed in Matlab which computed all path lengths between all paired participants linked in the recruitment chain, their alters, and participant reports on which alters also know each other. Although this computation is based on a sample of the network, we assume that it closely corresponds to the real average path length in the entire network^{10,30}. The clustering coefficient and the average path length of the random network was identified by constructing 1,000 Erdős-Rényi (E-R) random graphs with the same number of nodes as our Ottawa PWID network, and the same number of edges assigned with uniform probability³⁰. The PWID network would be considered a small-world network if the average path length were equal to or greater than the average path length of the constructed random network and if the CC of the PWID network is greater than the CC of the random. We used two different global estimates of the clustering coefficient to evaluate the PWID network. One was from Watts and Strogatz¹⁰; $CC^{WS} = \frac{1}{N} \sum_i^N C_i$; where $C_i = \frac{2E_i}{k_i(k_i-1)}$ (k_i is the degree of node i , and E_i is the number of edges between k_i neighbours of node i). The second clustering coefficient, $CC^{RDS} = \frac{1}{\sum_i^n k_i(k_i-1)} \sum_i^n 2E_i$, was developed by Hardiman and Katzir³⁵ for social networks sampled with a random walk, in which future participants are selected to receive referral coupons from ego's network at random. It was then further expanded for application to the RDS sampling method, where non-random selection, branching recruitment chains, multiple seeds, non-replacement, and other features of the sampling process occur^{13,31,35}.

The overall small world-ness measure (S) is a composite of the average path lengths and clustering coefficients of the random or null models compared with the real network:

$$\frac{CC_{PWID\ network}}{CC_{random}} \bigg/ \frac{L_{PWID\ network}}{L_{random}} \text{ and should be greater than 1 if the real network is a small world}$$

network³⁰. The clustering coefficients and short path lengths were calculated using a package developed by Humphries & Gurney (2008) in MATLAB (R2017B)³⁰. The clustering coefficient tailored to RDS sampling was conducted using Verdery (2016) method in STATA (version 15.0).

5.4 Results

Of the total 407 participating PWID, 398 individuals included details for at least one member of their social network, but only 364 PWID completed the interaction grid of contacts with whom they injected drugs within the last 30 days prior to the interview. No significant difference in demographics, infectious status and number of injections in 6 months were identified from the individuals who completed the interaction grid to those who did not. Approximately 20% of the responses collected from RDS participants showed discrepancies. For example, one contact was identified not to inject drugs with the RDS participant; however, sharing of needles was involved. In another instance, the social contact did inject drugs but sharing was not involved; however, the RDS participant had stated in questions related to them that they do share needles and/or equipment. Contradictions were resolved by including the individual as PWID if the RDS participant said yes to either question. Table 5 highlights the breakdown of the demographic and risk factors that were collected for both the participant and their social contact. It also illustrates the significant differences in demographics and risk factors

amongst the RDS participants and their social contacts (Table 5). Individuals participating in the RDS were older, with the highest proportion being 41 to 50 years old, whereas the individuals identified from the ego network were mainly 31 to 40 years old. Caucasian males were predominant in both groups; however, the proportions of females and Caucasians were much higher in the ego network. We expected drug use habits to differ between respondents and their alters as eligibility for the study was injection in the past six months, but social contacts were not required to be active PWID or even ever to have injected. We also expected and found that the proportion of individuals sharing needles and equipment differed significantly between RDS participants and alters ($p < 0.0001$).

Table 5: Demographics of the RDS participants and their alters.

Demographics/Risk Factors	Sample (n)	RDS participants	Sample (n)	Alters	<i>p</i> value [†]
Age					
≤ 20 years old	406	15 (4%)	1,590	115 (7%)	<0.0001
21 – 30 years old		71 (17%)		409 (26%)	
31 – 40 years old		108 (27%)		541 (34%)	
41 – 50 years old		157 (39%)		392 (25%)	
>50 years old		55 (13%)		133 (8%)	
Gender					
Male	401	319 (80%)	1,587	1,049 (66%)	<0.0001
Female		82 (20%)		538 (34%)	
Ethnicity					
Caucasian	401	311 (77%)	1,592	1,380 (87%)	<0.0001
Aboriginal		47 (12%)		119 (7%)	
Black		13 (3%)		62 (4%)	
Other		30 (8%)		31 (2%)	
Injected drugs in the last 6 months					
Yes	407	407 (100%)	1,586	972 (61%)	<0.0001
No		0 (0%)		614 (39%)	
Shared needles and/or equipment					
Yes	404	274 (68%)	972	478 (49%)	<0.0001
No		130 (32%)		494 (51%)	

[†] Fisher's two-sided test

The degree of the RDS participant roughly resembles a power law curve, (Figure 8) with distinctive small numbers, n=13, of individuals who have more than 10 sharing partners (hubs) where in contrast the majority of individuals have only 1 to 4 contacts (n=315).

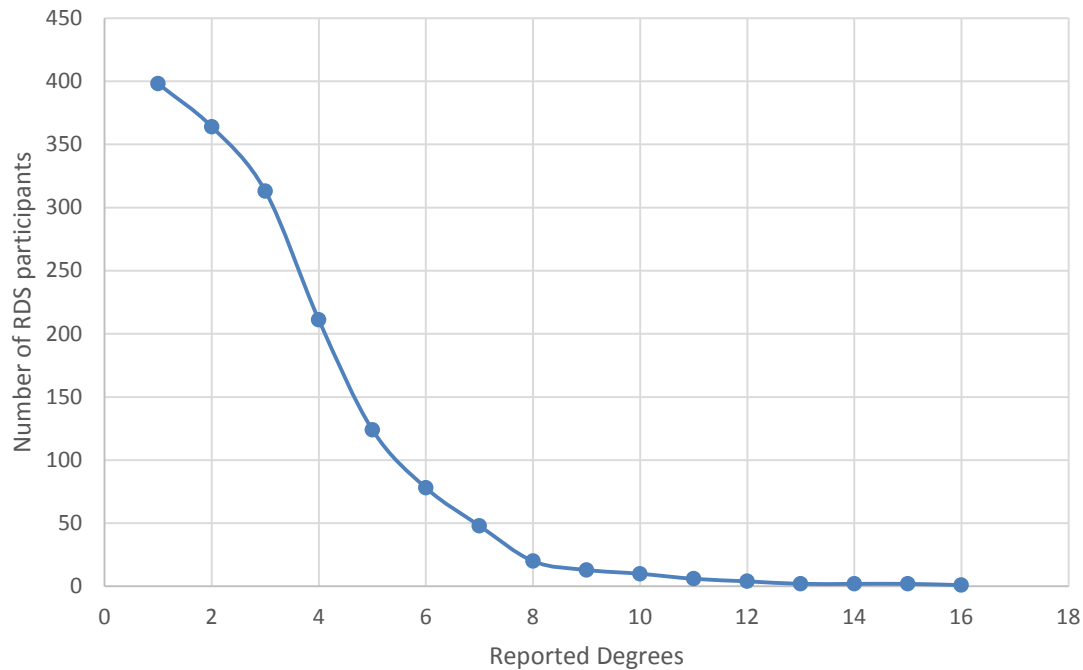


Figure 8: Plotted frequency of social/drug using interactions with alters reported by RDS Participants, Ottawa PWID study, 2007

Most participants associated and/or injected drugs with someone they consider a friend, implying that most individuals surround themselves with people they can trust (Table 6). Only 17% of individuals reported injecting drugs with an acquaintance and 4% with their dealer, both of which may be underestimated if the individual injects very frequently. A Chi-square test of independence was conducted to assess the association between HCV infection status and degree ($p = 0.7532$, Figure 9).

Table 6: The relationship PWID and non-PWID alters identified from the interaction grid completed by the RDS participant ($p < 0.0001$)

Relationship	PWID alters	Non-PWID alters
Acquaintances	154	111
Family member	12	81
Friend	634	327
Dealer	59	0
Partner	95	55
Other	36	29

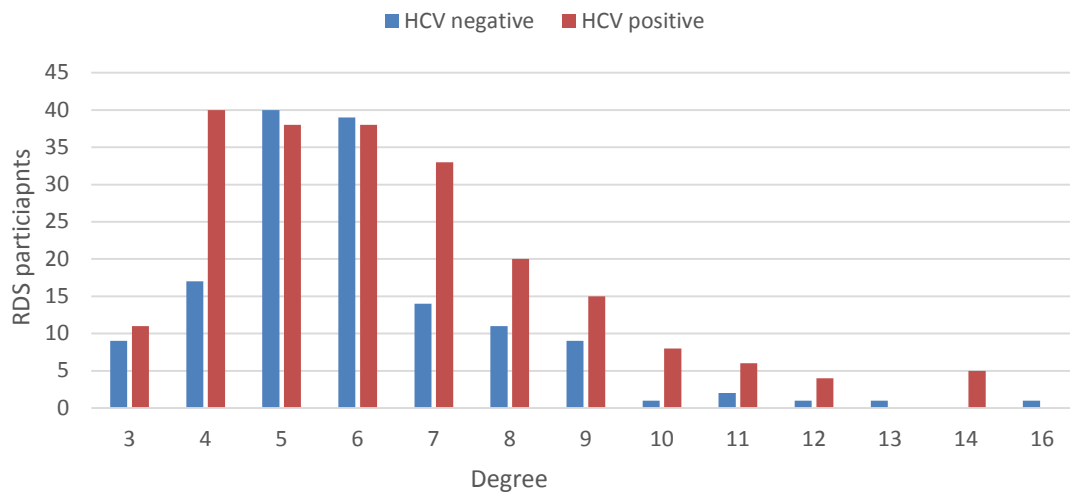


Figure 9: HCV status based on the reported and calculated degrees of the RDS participants

($n = 364$)

The two different ways of calculating clustering coefficient, the conventional WS method³⁰ and the Hardiman and Katzir method, appropriate for RDS process^{13,31,35}, yielded almost identical results in indicating that respondents in this study constitute a small-world network (Table 7).

Table 7: Network and Small-world-ness properties of the Ottawa IDU RDS data conducted in 2007

Properties	Estimates
RDS Sample size	398
Number of total nodes (n)	1,912
Total number of edges (m)	4,154
Average degree $\langle k \rangle$	5.25
Edge density (ξ)	0.0023
Average path Length (L)	4.6
WS Clustering Coefficient (CC^{WS})	0.666
RDS Clustering Coefficient (CC^{RDS})	0.667
WS Small-world-ness (S^{WS})	40.3
RDS Small-world-ness (S^{RDS})	41.8

5.5 Discussion

We used RDS to sample 407 PWID in Ottawa to elucidate the type of network formed by their interactions, which can help to shed light on the epidemic potential of PWID networks for the spread of infections. Although described in more depth in the previous paper, respondents differed from their contacts in that they were older, more likely to be male, belong to a minority and share drug equipment. Variation between respondents and their alters is reassuring as it demonstrates that information on wider groups than just the immediate respondents is revealed. Prior work on RDS has used such information for estimating population prevalence, but it is rarely incorporated for substantive analysis³⁸⁻⁴¹.

The frequency distribution of contacts in the Ottawa PWID network resembles a classic small world network in which 96 % of people have 1 – 4 PWID in their ego networks, while 4 % have 10 – 16. As described previously those with the highest number of contacts form hubs through which infection and information can more easily spread to distant parts of the network (Abdessalam, 2018).

Twenty-one percent of respondents report injecting with an acquaintance or dealer, rather than a friend. Injecting with someone less well-known or trusted is likely to result in an underestimate of exposures, as equipment may have been previously used, unbeknownst to the respondent. In a previous study on the same population, the highest number of injections within 6 months was 960. The average degree of individuals who injected minimum 960 injections in 6 months, was 7 (range 4-12). Interestingly, the highest alter degree reported here was 16 and it was from an individual who reported not injecting drugs frequently and was HCV and HIV negative (Figure 9).

The infection status of respondents was unrelated to the number of injecting or sharing alters. Recent seroconversion to HCV positive is likely to be associated with a high number of exposures to other injecting and sharing partners, but usually it occurs within the first one or two years of injecting⁴². As these data are cross-sectional, they reflect current interactions and not interactions and exposures which likely occurred some years before. It is also likely that after participants test positive for HCV, they are likely to change their behaviours to avoid further transmission of HCV and also protect themselves against contracting HIV. This counter intuitive finding is supported by HCV sequencing results, which demonstrate dissimilar HCV strain types between respondents and their recruits³⁶.

Small networks of 200 to 3000 nodes can be shown to exhibit small-world properties by comparing the clustering coefficient of the generated corresponding random network, with matching numbers of nodes and links^{43,44}. In this study, we showed that the Ottawa PWID network had an average path length and clustering coefficients which differ significantly from the random models, and the composite small world-ness indicator greatly exceeded 1³⁰. Interestingly, the two different approaches used in this study to identify the clustering coefficient

for the Ottawa IDU network provided similar results suggesting that both methods may produce reliable results.

Limitations of this analysis are primarily related to RDS study design. Participants were asked to list a maximum of 20 alters, answer a series of questions about those alters, and then report on alters' interactions with each other. This required a lot of time and attention from the participants and may have discouraged them from listing all social contacts, possibly resulting in underestimation of higher degrees and truncation at the upper end. However, despite these limitations, we found the small world topology despite the potential underestimate, indicating that a small world topology is almost certain. Due to the cross-sectional design we did not find an association between HCV positivity and numbers of contacts, which has been previously noted. Discrepancies between reported sharing practices in general and those responses to injecting interaction with their social contacts occurred in approximately 20% of our participants. Follow-up was not feasible to validate the responses of the RDS participants, but as previously noted, follow-up and dual incentives are difficult to operationalize together while maintaining anonymity. Furthermore, the limitations and assumptions associated with the RDS process should also be noted ^{38,39,45-48}.

Despite these limitations, this RDS sample produced similar findings of prevalence of HCV and HIV in PWID to those of the Ottawa Public Health, and is consistent with estimates at the national level for PWID ¹⁻³. Most important, RDS remains one of the very few methods by which to obtain anonymous, confidential data from a group of interconnected marginalized people who form a subculture in which only they are experts. Our finding that the links between them form a small world structure has valuable implications not only for the spread of infection but for the most recent severe and persistent opioid crisis.

Many studies have focused on the transmission of disease within various types of networks, as the success of diseases spread is heavily influenced by the contact structure in the network⁴⁹⁻⁵³. Infectious diseases need only a small amount of randomness on network interactions in order for them to spread efficiently. High degrees of clustering and short average path lengths can lead to a less than exponential spread of diseases even at the very beginnings of epidemics^{54,55}. Small world networks share similar clustering capabilities as the regular lattice network and also include randomness, which creates short average path lengths, like the random network. Infectious disease spread more easily in scale-free and small-world networks than in regular lattices and random networks^{10,56-58}.

The same small world properties that facilitate the quick spread of disease within a network can also be used to implement efficient and rapid interventions. In our previous study, only 47% of our sample indicated that they go to an NEP at least once a week (Abdesselam, 2018). The majority of PWID in our network sample, 53%, are not using NEP either because they are unaware of the services offered, the NEPs are not closely located, or open hours do not coincide with the individual's injection time frame^{14,15,59-61}. This would suggest that the existing strategies discussed may not impact the majority of PWID. The structure of the small world networks could provide a solution to this problem. Interventions can be targeted towards individuals highly clustered instead of random individuals or people who seek help (i.e. individuals who use NEP, social or/and health services), which is unlikely to be effective to control the PWID population. Strategically interrupting transmission requires interventions targeted to the most active people and centrally located venues for social and health care. It is possible that these people will be valuable members of a prevention team, as by definition they are acquainted with a lot of people, who likely share similar risk practices they do. Peer

education, treatment options, stable housing, rehabilitation, information on contaminated drug supplies, clean injection equipment, and possible future vaccines should reach those most active first. Providing large amounts of needles to the most active users, who then distribute them more widely (secondary distributors) will support active users, disseminate prevention messages and technology, and also reinforce the network connections, by building prevention into a social structure which already exists ^{62,63}.

Several strategies have been put in place to reduce sharing of potential contaminated needles. For instance, opening hours have been increased at the needle exchange programs (NEP) within community health centers across the city of Ottawa, with the hope of obtaining enough resources to have these centers open 24 hours a day ⁶⁴. Vending machines with clean pipes and needles are also now available outside the needle exchange programs, which offer tokens to PWID ⁶⁵. In addition to more health services provided to PWIDs at these NEPs, rapid drug analysis is being introduced that will allow drug users to know the content and quality of their drugs, for the most part, before they use them ⁶⁶. Monitored issue of tokens for vending machines dispensing medical grade hydromorphone have also been suggested, within smaller dedicated highly supervised centers ⁶⁷.

Great strides have been made in anti-virals for HCV treatment. The treatments have been reduced to one pill a day for a duration of 8 to 12 weeks. Treatments are estimated to cost \$45,000 to \$100,000 per patient ⁶⁸. Fortunately, Ontario and British Colombia have recently expanded treatment coverage to all HCV patients, regardless of the severity of the infection. The remaining provinces and territories are working on a similar coverage for HCV infections within their jurisdictions ⁶⁹. These antivirals are very efficient in eliminating the virus and thus, patients treated will not long be able to infect others.

All these strategies will help reduce newly acquired HCV infections and will identify individuals who are unaware of their infectious status by encouraging individuals to access healthcare, counselling and treatment access. This will in turn increase care and treatment among the susceptible population. However, incorporating these strategies within the social dynamics of a small world network will not only provide a successful intervention but a rapid one. In addition, a geographic system analysis can also provide insight on the most effective locations for vending machines, needle exchange programs and health services. Logan *et al.* (2016) demonstrated small (500 m^2) areas of a city can be identified from crossing paths between respondents residences, that of their friends, social venues and all places between, in which vending machines, needle drop boxes and social and health services should be placed⁶¹.

5.6 Conclusion

Small worlds may play an important role in the study of the influence of the network structure upon the dynamics of disease transmission. Small world networks present a much faster epidemic propagation than models based on regular lattices and randomness of a social network^{50,51,53,56,57}. Therefore, the identification of the small-worldness of the PWID in Ottawa can be used to tailor successful interventions for harm reduction, efficient treatment, and prevention and control so that the largest effective reduction can occur with the smallest number of influential individuals. Identifying the clusters within the network and targeting the programs to those particular groups, or to the individuals with highest degrees, will allow for rapid protective approaches to be spread to the rest of the network, due to the short average path length of SW networks.

5.7 References

1. Remis RS. Modelling the incidence and prevalence of hepatitis C infection and its sequelae in Canada, 2007. 2010.
2. Laupland KB, Embil JM. Reducing the adverse impact of injection drug use in Canada. *Can J Infect Dis Med Microbiol*. 2012;23(3):106-107.
3. Public Health Ottawa. Problematic substance use in ottawa: Technical report. . 2016.
4. Navarro C, Leonard L. Prevalence and factors related to public injecting in ottawa, Canada: Implications for the development of a trial safer injecting facility. *International Journal of Drug Policy*. Vol 15.
5. Millson P, Leonard L, Remis RS, Strike C, Challacome L. Injection drug use, HIV and HCV infection in Ontario: The evidence 1992 to 2004. 2005.
6. Bearman PS, Moody J, Stovel K. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*. 2004;110(1):44-91.
7. Kuttler AF, La Greca AM. Linkages among adolescent girls' romantic relationships, best friendships, and peer networks. *J Adolesc*. 2004;27(4):395-414.
8. Barabási A. Scale-free networks: A decade and beyond. *Science*. 2009;325(5939):412-413.
9. Barabási A, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512. doi: 10.1126/science.286.5439.509.
10. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*. 1998;393(6684):440-2.
11. Rocha LE, Liljeros F, Holme P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol*. 2011;7(3):e1001109.
12. Newman L, Dale A. Large footprints in a small world: Toward a macroeconomics of scale. *Sustainability: Science, Practice, and Policy*. 2009;5(1).
13. Verdery AM, Fisher JC, Siripong N, Abdesselam K, Bauldry S. New survey questions and estimators for network clustering with respondent-driven sampling data. *Sociological Methodology*. 2017;47(1):274-306.
14. De P, Connor N, Bouchard F, Sutherland D. HIV and hepatitis C virus testing and seropositivity rates in canadian federal penitentiaries: A critical opportunity for care and prevention. *Can J Infect Dis Med Microbiol*. 2004;15(4):221-225.

15. Pelude L. *Networks among injection drug users: Random or scale-free?* [Masters]. University of Ottawa; 2007.
16. Barzel B, Barabási AL. Universality in network dynamics. *Nat Phys*. 2013;9:10.1038/nphys2741.
17. Sole RV, Valverde S. On evolution and architectural constraints . In: Information theory of complex networks. Springer-Verlag; 2004:190-210.
18. Wang XF, Chen G. Complex networks: Small-world, scale-free and beyond. 2001.
19. Zaidi F, Sallaberry A, Melançon G. Generating artificial social networks with small world and scale free properties. 2012.
20. Chakraborty A, Manoj BS. The reason behind the scale-free world. *IEEE Sensors Journal*. 2014;14(11):4014-4015.
21. Broido A,D., Clauset A. *Scale-free networks are rare*. 2018.
22. Albert-Laszlo, Barabási. The physics of the web. *Physics World*. 2001;14(7):33.
23. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data . *SIAM Review*. 2009;51(4):661-703.
24. Sanabria-Montana CH, Huerta-Quintanilla R. Generalization of clustering coefficient on lattice networks applied to criminal networks. *International Journal of Computer and Information Engineering*. 2017;11(7):868-872.
25. Newman ME, Moore C, Watts DJ. Mean-field solution of the small-world network model. *Phys Rev Lett*. 2000;84(14):3201-3204. 26. Telesford QK, Joyce KE, Hayasaka S, Burdette JH, Laurienti PJ. The ubiquity of small-world networks. *Brain Connect*. 2011;1(5):367-75.
27. Sohn I. Modeling and implementing small world network..2016:414-416.
28. Barabási A, Regan E, Vicsek T. *Deterministic scale-free networks*. Vol 299. ; 2001:559-564. 10.1016/S0378-4371(01)00369-7.
29. Barabási AL, Bonabeau E. Scale-free networks. *Sci Am*. 2003;288(5):60-69.
30. Humphries MD, Gurney K. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS ONE*. 2008;3(4):e0002051.
31. Verdery AM, Mouw T, Bauldry S, Mucha PJ. Correction: Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE*. 2016;11(1):e0148006.

32. Menezes MBC, Kim S, Huang R. Constructing a watts-strogatz network from a small-world network with symmetric degree distribution. *PLoS ONE*. 2017;12(6):e0179120.
33. Centola D, Macy M. Complex contagions and the weakness of long ties. *American Journal of Sociology*. 2007;113(3):702-734.
34. Centola D. The spread of behavior in an online social network experiment. *Science*. 2010;329(5996):1194-1197.
35. Katzir L, Hardiman S. Estimating clustering coefficients and size of social networks via random walk. *ACM Transactions on the Web*. 2015;9(4).
36. Pilon R, Leonard L, Kim J, et al. Transmission patterns of HIV and hepatitis C virus among networks of people who inject drugs. *PLoS One*. 2011;6(7):e22245.
37. Blum K, Han D, Modestino EJ, et al. A systematic, intensive statistical investigation of data from the comprehensive analysis of reported drugs (CARD) for compliance and illicit opioid abstinence in substance addiction treatment with buprenorphine/naloxone. *Subst Use Misuse*. 2018;53(2):220-229.
38. Lu X. Linked ego networks: Improving estimate reliability and validity with respondent-driven sampling. *Soc Netw*. 2013;35(4):669-685.
39. Verdery A.M., Merli M.G., Moody J., Smith J.A., Fisher J.C. Brief report: Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in china. *Epidemiology*. 2015;26(5):661-665.
40. Yamanis TJ, Merli MG, Neely WW, et al. An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in china. *Sociol Methods Res*. 2013;42(3).
41. Merli M.G., Moody J., Smith J., Li J., Weir S., Chen X. Challenges to recruiting population representative samples of female sex workers in china using respondent driven sampling. *Soc Sci Med*. 2015;125:79-93.
42. Robertson J, Bucknall A, Welsby P, et al. Epidemic of AIDS related virus (HTLV-III/LAV) infection among intravenous drug abusers. *Br Med J (Clin Res Ed)*. 1986;292(6519):527-529.
43. Muldoon SF, Bridgeford EW, Bassett DS. Small-world propensity and weighted brain networks. *Sci rep*. 2016;6:22057.
44. Uehara T, Yamasaki T, Okamoto T, et al. Efficiency of a "small-world" brain network depends on consciousness level: A resting-state FMRI study. *Cereb Cortex*. 2014;24(6):1529-39.
45. McCreesh N., Frost S., Seeley J., et al. An empirical evaluation of respondent-driven sampling. *Sex Transm Infect*. 2011;87:A15-A16.

46. Goel S., Salganik M.J. Assessing respondent-driven sampling. *Proc Natl Acad Sci U S A*. 2010;107(15):6743-6747.
47. Gile KJ, Handcock MS. Respondent-driven sampling: An assessment of current methodology. *Sociol Methodol*. 2010;40(1):285-327.
48. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *J R Stat Soc Ser A Stat Soc*. 2015;178(1):241-269.
49. Gallos LK, Makse HA, Sigman M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks; 22308319. *Proc Natl Acad Sci U S A*. 2012;109(8):2825-2830.
50. Enright J, Kao R. *Epidemics on dynamic networks*. *J.Epidem* Vol 24. 2018. 10.1016
51. Brauer F. An introduction to networks in epidemic modeling. *Journal of Mathematical Epidemiology*. 2008;1945.
52. Saramaki J, Kaski K. Modelling development of epidemics with dynamic small-world networks. *Journal of theoretical biology*. 2005:413-421.
53. Demirel G, Barter E, Gross T. Dynamics of epidemic diseases on a growing adaptive network. *Scientific Reports*. 2017;7:42352.
54. Csányi G, Szendroi B. Fractal-small-world dichotomy in real-world networks. *Phys Rev E Stat Nonlinear Soft Matter Phys*. 2004;70(1 2):016122-1-016122-5.
55. Verdasca J, Telo da Gama,M.M., Nunes A, Bernardino NR, Pacheco JM, Gomes MC. Recurrent epidemics in small world networks. *J Theor Biol*. 2005;233(4):553-61.
56. Kuperman M, Abramson G. Complex structures in generalized small worlds. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*. 2001;64(4):471031-471034.
57. Kiss IZ, Green DM, Kao RR. The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. *Journal of the Royal Society Interface*. 2007;5(24):791-799.
58. Matthäus F. Diffusion versus network models as descriptions for the spread of prion diseases in the brain. *Journal of theoretical biology*. 2006;240(1):104-113.
59. Strike C, Miskovic M. Scoping out the literature on mobile needle and syringe programs—review of service delivery and client characteristics, operation, utilization, referrals, and impact. *Harm Reduction Journal*. 2018;15:6.

60. Wylie JL, Jolly AM. Understanding recruitment: Outcomes associated with alternate methods for seed selection in respondent driven sampling. *BMC Med Res Methodol*. 2013;13:93-2288-13-93.
61. Logan JJ, Jolly AM, Blanford JI. The sociospatial network: Risk and the role of place in the transmission of infectious diseases. *PLoS ONE*. 2015;11(2):e0146915.
62. Neaigus A, Reilly KH, Jenness SM, Hagan H, Wendel T, Gelpi-Acosta C. Dual HIV risk: Receptive syringe sharing and unprotected sex among HIV-negative injection drug users in new york city. *AIDS Behav*. 2013;17(7):2501-2509.
63. Neaigus A, Reilly KH, Jenness SM, et al. Trends in HIV and HCV risk behaviors and prevalent infection among people who inject drugs in new york city, 2005-2012. *Jaids*. 2017;75:S325-S332.
64. Kupfer M. City-run injection site opens Tuesday. *CBC News*. 2017. Available from: <https://www.cbc.ca/news/Canada/ottawa/tour-ottawa-public-health-supervised-injection-site-1.4305663>.
65. Lofaro J. Vending machines with clean pipes, needles for drug users open in Ottawa. *CBC News*. 2017. Available from: <https://www.cbc.ca/news/Canada/ottawa/vending-machines-pipes-needles-drug-users-ottawa-1.4292284>.
66. Gillis M. Ottawa's supervised injection site will have high-tech drug analyzing device. *Ottawa Citizen*. 2017. Available from: <https://ottawacitizen.com/news/local-news/pop-up-od-prevention-group-claims-life-saved-ahead-of-awareness-day-thursday>.
67. Woodward J. Safe opioid vending machine closer to reality: BCCDC. *CTV News Vancouver*. 2018. Available from: <https://bc.ctvnews.ca/safe-opioid-vending-machine-closer-to-reality-bccdc-1.3763945>.
68. CATIE. Hepatitis C treatment combinations. <http://www.catie.ca/en/practical-guides/hepc-in-depth/treatment/choosing-drug-combination-chronic-hepatitis-c>. Updated 2018. Accessed 02/18, 2018.
69. Grant K. Deal reduces price of life-saving hepatitis C drugs for Canadians. *The Globe and Mail*. 2017. Available from: <https://www.theglobeandmail.com/news/national/deal-reduces-price-of-life-saving-hepatitis-c-drugs-for-canadians/article34107225/>.

6.0 The Development of Respondent-Driven Sampling (RDS) Inference: A Systematic Review of the Population Mean and Variance estimates

RDS makes strong assumptions associated with both the respondent's behaviour and network structure. In order to better understand the population of interest and to use appropriate RDS estimators to the corresponding RDS data, identifying the appropriate network structure of the population of interest is necessary. Our previous manuscripts have demonstrated that the random and exponential distributions are not appropriate fits to the PWID population network in Ottawa. In fact, the Ottawa PWID data was found to exhibit a small-world structure and the presence of key individuals involved in significant high frequency of injections (i.e. proxy for frequent use of needle sharing). This type of network structure needs to be taken into consideration prior to applying RDS estimators, as it has important implications for the inferential aspect of RDS.

Due to the increasing use of RDS within various hard-to-reach populations across the world, along with the fact that the focus of these studies is on prevalence estimation (i.e. estimation of the population mean or proportion of a focal attribute) there is a dire need to understand the limitations of the RDS mean and variance estimators prior to applying them to RDS data.

RDS mean estimators have shown, via stimulation studies based on empirical RDS data, to produce unbiased estimates so long as the assumptions associated with the RDS process are met. However, there is sufficient evidence to demonstrate that most of the theoretical assumptions are not plausible in real RDS scenarios. Furthermore, studies have shown that RDS variance estimators are greatly underestimated, even when the underlying assumptions of RDS process are met.

The third and last study focuses on a systematic review on the development and evaluation of RDS mean and variance estimators. The purpose of this review is to help inform future new RDS users' insight on the appropriate conditions, network structure and assumptions for which RDS mean and variance estimates to use for their corresponding data. The study emphasizes caution in relying on RDS estimates produced by RDS estimators, particularly when assumptions are violated.

Abdesselam, K.¹, Verdery, A.², Pelude, L.¹, Momoli, F.¹, Dhami, P.¹, and Jolly, A.M.¹

¹ School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa

² Department of Sociology and Criminology, Pennsylvania State University

6.1 Abstract

Respondent-driven sampling (RDS) is a successful data collection method used in hard-to-reach populations. Since its introduction in 1997, there has been an emphasis on identifying appropriate methods for population mean estimates, and more recently, variance estimates. A systematic review using four electronic databases was conducted in order to summarize the progress of RDS inference over the last 20 years and to provide insight to researchers on using the appropriate estimates in analyzing RDS data. Two independent reviewers selected the relevant abstracts and articles. Thirty-two studies were included. The content of the studies was further categorized into developing and evaluating RDS mean and variance estimators. Through rigorous analytical studies, RDS^{IEGO} and Tree boot strapping were identified as the population mean and variance estimate that rely on fewer assumptions, respectively. However, the fact remains that both estimators rely on assumptions that are typically violated in real life, thus, all RDS estimators, both mean and variance, are susceptible to bias.

6.2 Introduction

Certain populations of interest may not have an attainable sampling frame, which makes it very difficult to obtain a representative sample and make any inferences about that particular population. This issue is common in individuals who are unwilling to identify themselves because they engage in illegal and/or stigmatized activities. Examples of hard-to-reach populations include sex workers, people who inject drugs (PWIDs), men who have sex with men (MSM), and street youth (SY)^{1,2}. Many non-random methods of sampling have been developed

over the past few decades, including respondent-driven sampling (RDS) which has become a popular and successful method of sampling these hard-to-reach populations.

In 1997, Heckathorn first introduced RDS, a network-based sampling methodology influenced by snowball sampling and traditional contact-link tracing ². This sampling process was designed to increase the ease of sampling among hidden populations and to approach probability samples as sampling progresses ². What is unique about RDS is that respondents are responsible for recruiting other participants. Participants form a referral chain characterized by waves, where each wave represents one set of recruitments (or step) along the chain. The process starts with initial participants selected by the investigators (referred to as “seeds” and counted as wave zero), who are interviewed. They are supplied with a set of numbered and uniquely coded referral coupons which they then distribute to people they know that are also in their “hidden” population. Typically, seeds are provided with a dual financial incentive: one for their involvement in the study and other(s) once they have successfully recruited another individual. The recruitment chain continues until a large enough sample size or number of waves has been reached ².

Over 460 studies have been conducted illustrating the success of RDS in recruiting hard-to-reach populations and, if assumptions are met, then RDS provides an additional innovation in inference about the target population through the use of weighting techniques ³. However, the assumptions associated with the RDS process are quite stringent and bear little resemblance to reality (Table 8). One assumption that is automatically violated is sampling with replacement, as RDS is usually performed without replacement, disallowing the same individual from being recruited more than once.

Table 8: Six critical assumptions associated with the RDS design which when satisfied, estimates from RDS inference are considered reliable ²

Assumption	Description
(1) Seeds drawn with probability proportional to their degree	Initial individuals are chosen by the researcher based on their network size
(2) Undirected (symmetric) network	Individuals in the target population maintain a reciprocal relationship with their contacts
(3) Irreducible	Each individual can be reached by another individual through a series of network ties
(4) Sampling with replacement	Individuals can be included in the study multiple times
(5) Degree accurate	Individual's network size, which is the number people they are in contact within the target population, is accurate
(6) Random recruitment	Individuals recruit uniformly at random from their personal network.

The widespread use and application of RDS in estimating population characteristics (i.e. disease prevalence, demographics at risk) based on a sample has assisted in informing public health questions, policies and interventions emphasizes the importance of clarifying the statistical foundations of RDS. The recruitment technique is the reason for its success in data collection; however, it is that same mechanism that complicates inferences because it requires the individuals conducting the study to make assumptions about the recruitment process and the structure of the social network connecting individuals within the study population ^{1,4-7}. There are three main approaches to addressing the issues associated with the assumptions underlying inference from RDS data ⁸. The first is the development of estimators that are accompanied by less stringent assumptions. Over the last two decades, the inference estimators have gone through several changes that address the statistical models' unrealistic assumptions about network structure and peer recruitment process. The second approach evaluates these estimators and how violations of assumptions impact them, either analytically or through simulation studies. The third approach assesses the quality of the RDS data and determines whether assumptions have been violated in practice.

In less than a decade, improving inference on RDS data has focussed on addressing RDS reliability and validity. There is more emphasis on providing robust estimates for the measurement of precision and variability. We unify and highlight the main findings on the reliability and validity of estimates conducted so far. This is the first study that systematically assesses all developed RDS mean and variance estimates through content analysis. The findings of this study are meant to guide researchers, interested in conducting RDS, on the strengths and limitations of the developed estimates and to help provide the most promising mean and variance estimates so far.

6.3 Methods

Studies on developed RDS estimators, measures of uncertainty, population size estimates, evaluations of RDS, and RDS methodology were identified through a systematic literature review.

The literature review was conducted using the method proposed by Pai *et al.* and the PRISMA reporting format was followed ^{9,10}. A broad list of RDS studies which shed light on the inferential aspect of RDS was first obtained through the system literature review conducted in October 2017 in search engines Medline, Embase, Scopus, and Web of Science. Abstracts of published articles were screened and a subset was retained for full text review. The PICOS criteria for the systematic review are described in Table 9 below ⁹.

Table 9: PICOS for systematic review on studies focusing on improving the inferential aspect of RDS

P	I	C	O	S
Patients, Population or Problems	Interventions or Exposures	Comparison	Outcomes	Study design
There are several concerns regarding the reliability and validity of RDS studies - what progress has been made since its introduction	- Development of RDS estimators, - Evaluation of RDS estimators	No comparator specified per se but several studies of interest have compared: - RDS Estimators - Variance Estimators	Evaluation/effectiveness of Estimates Development of Estimates	Observational studies Simulations (<i>in silico</i>)

6.3.1 Eligibility Criteria

6.3.1.1 Problems

Addressing the assumptions underlying inference from RDS data, has been the prominent focus on improving RDS inference and was the main consideration in this review. Any studies that provided insight on additional estimates that focused on precision, variability, and uncertainty (i.e. variance estimates and confidence intervals) were also considered.

6.3.1.2 Phenomena of interest (*Intervention or Exposure*)

All studies related to the progress and contributed to RDS inference were considered. This included any studies that assessed the impact of the estimates when assumptions are violated; developed new RDS estimators with less stringent assumptions; and evaluated RDS estimators using simulations, real RDS data and/or a combination of both.

6.3.1.3 Comparison

Prior to the review, there was no comparator specified; however, it was expected that many of the relevant studies that would be included in the two screens would involve comparing and evaluating current RDS estimators; including population prevalence and variance estimates.

6.3.1.4 Outcomes

All studies that provided any insight in improving inference in RDS mean and variance estimates. This would include evaluation of estimates as well as development of estimates studies.

Evaluation of estimates provides strengths and limitations, in various network conditions, in RDS data's reliability and validity by addressing the impact of assumption violations. Development of new estimates and alternative methodological approaches to RDS were retained.

6.3.1.5 Type of Studies (Study Design)

The systematic literature review considered all quantitative components of observational designs (mostly cross-sectional studies) and pseudo-populations created via simulation studies.

6.3.2 Information Sources

A search of Medline, Embase, Scopus and Web of Science electronic bibliographical databases was conducted, using MeSH and Emtree terms for Medline and Embase, respectively. Free-text keywords were used for Scopus and Web of Science. The search covered the period from January 1, 1997 to October 2017. The search was conducted using all identified key words and index terms relevant to the various areas of RDS inference across all included databases (Table 3.) Studies published in French or English were included.

6.3.3 Search Strategy

The search strategy is summarized in Table 10 and includes all relevant keywords used for the systematic literature review.

Table 10: Keywords used for the literature search strategy for studies focusing on RDS inference

Summary of core key words in respondent-driven sampling (RDS) literature search	
Respondent-driven sampling	Evaluation
Bias	Validation
Reproducibility	Data Interpretation
Statistics	Confidence Intervals
Epidemiologic methods	Variability
Reliability	Estimate
Inference	Methodology
Simulation	Limitation
Assumptions	Accuracy
Analysis	Effectiveness

6.3.4 Selection Process

Following automated elimination of duplicate sources found across the four databases, titles and abstracts of all unique records were manually screened using the eligibility criteria described in Table 11. Screening was independently conducted by two assessors, and any conflicts were discussed until consensus was reached. Reasons for exclusion during the first screen are not summarized in the review due to the relatively large number of sources. The full text of sources retained after screening was reviewed in order to confirm eligibility. This step was also conducted independently, with conflicts resolved between the two assessors. Further exclusions were performed based on review of full-text articles and reasons for exclusion were noted.

Table 11: Inclusion and exclusion criteria for eligibility assessment of literature sources

Inclusion Criteria	Exclusion Criteria
Study written in English or French Language Peer reviewed Development of RDS Estimator Evaluation of RDS Estimator Developed of variance estimate Evaluation of variance estimate	Studies focusing on the application of RDS Studies focusing on the data collections of RDS process Studies involving Online-RDS RDS Studies focusing on geospatial analysis Opinion, poster, or editorials Literature review Diagnostic Tools

6.3.5 Information Items

We summarized findings qualitatively, and extracted publication year; type of study, whether a development and/or evaluation of estimates or addressing gaps of RDS inference; study method: design, data source(s), outcome(s), and results: evaluations of estimates, summary of findings.

6.4 Results

Based on the search strategy a total of 1,996 sources were identified. Of these, 1,064 duplicates were removed yielding a total of 932 unique sources that were screened based on their titles and abstract (Figure 10). After the first screen, 132 articles were retained for review of the full text article. A total of 100 articles were excluded during the second screening process and reasons for exclusion are summarized in Table 12. Thirty-two articles remained of which the areas of focus are documented in Table 13.

Table 12: Reasons for exclusion during second screening of articles (full-text review)

Reason for Exclusion	Number of articles excluded
Related to Data Collection	30
Editorial/opinions/comments/poster/abstract	8
Comparisons to other non-random sampling methods	10
Reviews	4
Spatial Analysis	10
Online RDS	5
Application of RDS and population estimates	22
RDS Data quality	9
Design Effect	2
Total	100

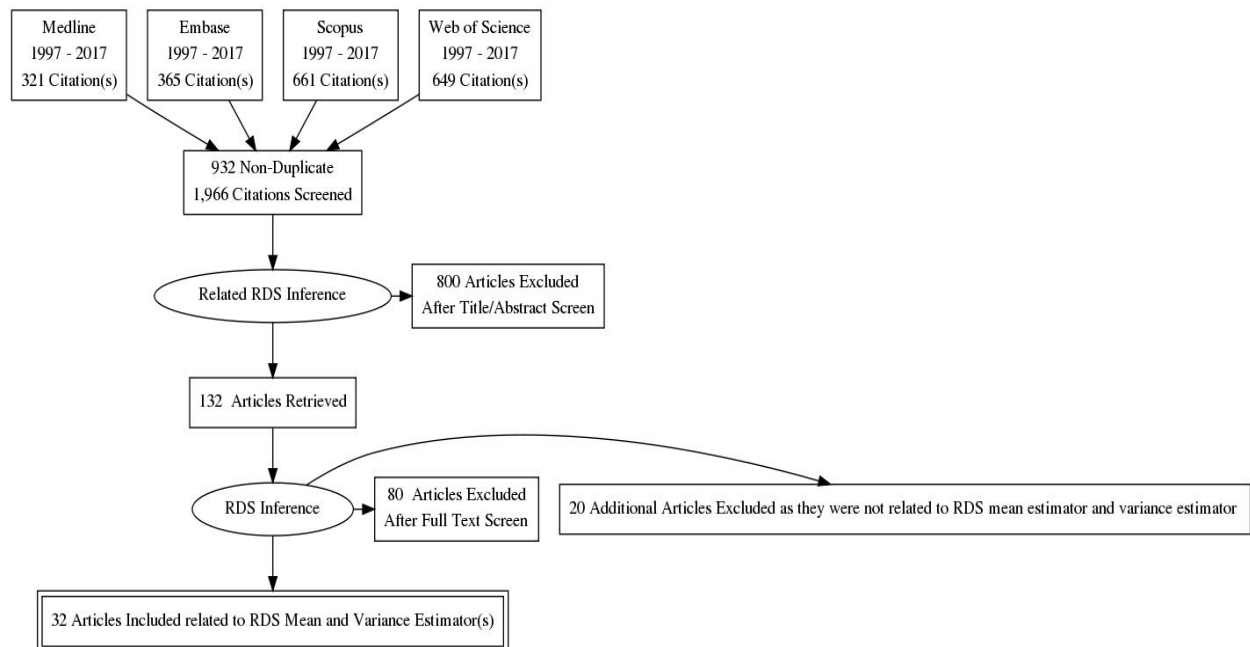


Figure 10: PRISMA flow chart of study selection for RDS Inference Studies⁸

Table 13: Areas of the inference from 32 RDS articles retained for the systematic literature review.

Areas focusing on Inference of RDS	Number of articles *
Developing RDS Estimator	11
Evaluating RDS Estimator	22
Developing Variance Estimate	8
Evaluating Variance Estimates	4

*The sum of the number of articles exceeds the total number of articles retained because this table reflects areas focusing on inference of RDS and some articles include more than one area.

6.4.1 RDS Estimator Developed

Since the introduction of RDS in 1997, Heckathorn demonstrated that RDS produced unbiased means/population estimates based on two analytical models: Markov chain model and theory of biased networks, also known as the homophily model ^{2,11}. He first noticed that RDS creates a stochastic process in which each recruiter's social characteristics affect the characteristics of their recruits and that the Markov model could provide a statistical model of the sampling process. Many of the developed estimators rely on the concept that RDS data can be modeled as a Markov Process (MP), which is equivalent to a random walk on a symmetric (undirected) graph ^{2,11}. In essence, the recruitment process is stochastic with two crucial components. The first being that recruitment can assume a limited number of states and the second, the process is state dependent, such that the probability of a certain state being reached depends only on the previous state of the chain ^{2,11}. To put this into context with the RDS design, the recruitment pattern is dependent only on the recruiter and not on the recruiter's recruiter. This type of process

So far, 10 prevalence estimators and 2 add-on modifications for an existing RDS estimator have been developed during the review period (January 1997 to October 2017, Table 14) and exclude the naïve estimator as this approach is identical to that of simple random sampling (SRS).

Table 14: RDS estimators developed since the introduction of RDS in 1997.

Year	Author	RDS Estimator	Highlight of Estimator	Reference
2002	Heckathorn	**RDS-HK ₁	Estimate includes post-stratification to control for differences in network size and clustering across groups	11
2004	Salganik and Heckathorn	**RDS-I	Estimates the average number of cross-ties from each group to the other group	12
2007	Heckathorn	**RDS-HK ₂	Is similar to RDS-HK ₁ but adjusts for differential recruitment degree and can be used to estimate continuous variables.	13
2008	Volz & Heckathorn	**RDS-II	Estimates the probability of an individual in the population by their network size, degree, and assumes the sampling probabilities are proportional to degree	14
2011	Gile	**RDS-SS	The Successive Sampling (SS) estimator is like RDS-II with the exception that it takes into consideration the feature “without replacement”. It also requires the population size to be known.	15
2013/2016	Lu <i>et al.</i> Malmros <i>et al.</i>	RDS-DN	Adjusts RDS-I and RDS-II for in- and out-degree, which no longer relies on the assumption that the network is undirected.	16,17
2013	Lu	**RDS ^{IEGO}	Incorporates the ego network into RDS-I and RDS-II but demonstrates that RDS-I + ego network (RDS ^{IEGO}) is superior.	18
2015	Gile & Handcock	*RDS-MA	Relies on the exponential-family random graph model to make estimates of the population prevalence and is supposed to adjust for seed selection and differential sampling	19
2015	Aronow & Crawford	*Non-parametric RDS-II	The estimate relies on less stringent assumptions; however, the conditions listed are untestable	20
2016	Selvaraj <i>et al.</i>	RDS-MOD	The estimate has a similar approach to RDS-II. The modification is the additional constant multiplier which incorporates the sample weights and information about correlation between sample units, capturing the clustering effect.	21
2017	Berchenko <i>et al.</i>	*RDS- Count process model	This maximum likelihood estimator requires the counts of individuals and the interview time of the participants. This estimator discards the homogenous random walk model for a stochastic epidemic model.	22

* Does not rely on the First Markov Order theory

** Most commonly used estimators in the literature, with the exception of RDS^{IEGO}, which was included in the study that evaluated the most estimators so far (Ref ID 1481).

6.4.2 Evaluation of RDS Estimators

General understanding of the use and sensitivities of RDS measures comes from statistical simulations and real RDS data. Even though most of the articles reviewed are focused on the evaluation of how accurately RDS estimators truly represent the target population, no gold standard estimator has yet been identified. Twenty-one studies used and compared RDS estimators on a population network (Table 14). Table 15 highlights the effectiveness and bias, a measure of the mean difference between the statistic and parameters across simulated studies that indicates whether the estimate is under or over estimating the population prevalence, of the evaluated RDS estimators. It is important to note that interpretation of Table 14 needs to be taken with caution due to the variation of the estimators tested, targeted population, approach and conditions of simulations, generating distributions, RDS conditions violated and other relevant information in regards to that particular study.

Table 15: Studies evaluating RDS estimators by year, RDS estimates and population network

Year	Evaluated RDS estimates	Population Network	Reference
2008	RDS-I, RDS-HK ₁ and RD- II	Simulation of RDS generated by 5 random network with mild assortative mixing and 10,000 random walks	14
2009	RDS-I and RDS-II	Two WebRDS samples of Undergraduates at the same residential university collected in 2008 and 2004 (Wejnert & Heckathorn, 2008)	23
2009	Naïve, RDS-I and RDS-II	Markov chain Monte Carlo Simulation Three sampling situations: SRS, RDS with single recruitment and RDS with multiple recruitment	24
2010	RDS-II	Simulation based on parameters obtained by the CDC HIV Enhanced Surveillance program (ref Abul-Quader et al., 2006)	25
2011	Naïve, RDS-I, RDS-II, RDS-HK ₂ , RDS-SS	Simulation based on parameters obtained by the CDC HIV Enhanced Surveillance program (ref Abul-Quader et al., 2006)	26
2011	Naïve, RDS-HK ₁ and RDS-II	Simulation based on parameters obtained from an MSM empirical RDS study and considered various different network distribution: - o Random - o Degree-dependent - o Individual covariate-dependent - o Transmission model	27
2011	RDS-SS and RDS-II	Simulation based on parameters obtained by the CDC HIV Enhanced Surveillance program (ref Abul-Quader et al., 2006)	28

2012	Naïve, RDS-I and RDS-II	Ongoing population cohort of 25 villages in rural Uganda (Sampling frame exists)	29
2012	RDS-II	Simulation based on empirical MSM network as well as artificial networks	15
2013/2016	Naïve, RDS-DN and RDS-SS	Ref ID 934 Simulated RDS on three different networks types of network: • Net 1: network with little or no indegree correlation and no indegree outdegree correlation • Net 2: network with varying homophily and positive indegree-outdegree correlation. • MSM Nordic regions largest active web community network Ref ID 641 simulated networks generating a Erdos-Renyi and Power law distributions and on empirical MSM Nordic online data	5,16
2013	RDS ^{IEGO}	Empirical MSM Nordic Online RDS data and simulated network based on the KOSKK model	18
2015	Naïve, RDS-II	Simulation study generating various network distribution: • Homophily • Inverse homophily • Rich-get-richer, and generating different RDS features	30
2015	Naïve, RDS-I, RDS-HK ₁ , RDS-HK ₂ , RDS ^{IEGO} , RDS-II and RDS-SS	Empirical population generated from the PLACE-RDS Comparison Study, which includes two surveys in the same population of female sex workers in Liuzhou, China: RDS and Venue-based. Simulated data generated from PLACE-RDS Comparison study and ideal-world scenario for RDS	31
2015	RDS-I, RDS-II, RDS-SS and RDS-MA	Simulated populations based on parameters obtained by the CDC HIV Enhanced Surveillance program (ref Abul-Quader et al., 2006) and includes three levels of simulation: • the generation of random networks according to specified networks features; • the generation of simulated RDS samples from each network; • the estimation of the prevalence of infection from each set of simulated sampled data	19
2016	Naïve, RDS-I, RDS-II and RDS-SS	Men who have sex with Men (MSM) in Moscow, Russia (RDS data)	32
2016	RDS-I, RDS-II and RDS-SS	Simulation study based on data from an RDS study of gay, bisexual and other MSM (GBMSM) in Vancouver, Canada	33
2016	RDS-HK ₂	Simulated studies, including with and without replacement RDS as well as both branching and non- branching referral process, generating different networks: • Watts-Strogatz (Small world) • Barabasi-Albert (Power law) • Erdos-Renyi (Random)	34
2016	RDS-I	Simulated studies, including with and without replacement RDS, on several different networks: • Erdos-Renyi (100 and 10,000 vertices) • Isolate Join Complete (IJC) • Zachary's Karate Club (Zachary, W., 1977) • Colorado Spring (Potterat, J., 2004)	35
2016	RDS-MOD and RDS-I	Empirical data on RDS: • Churachandpur, India data • Jazz musicians • Bishupur, PHek and Wokha	21

2018		3 other Datasets: faux, fausycamore, faumadrona	
	RDS-Count process model and RDS-SS	Empirical data obtained in R, available from CRAN: chords data Simulation generated data using various generative processes, with different network parameters Empirical Web RDS data of students from a large university in the US in 2008 (Wejnert, 2009)	22
2018	Naïve, RDS-I and RDS-II	A combination of simulation studies generating different networks: • Erdos-Renyi • Watts- Strogatz • Barbasí-Albert • Interconnected-Islands, and the Susceptible-infected model disease transmission process.	36

Many of the studies evaluate, on average, 3 estimators, with the exception of Gile *et al.* and Verdery *et al.* who evaluated 5 and 7 estimators, respectively^{19,31}. Most of the studies, including Gile *et al.* study, demonstrated that one estimator may perform better under certain conditions but overall no one estimator outperforms another. The RDS-II estimator is the most frequently used estimator and has been demonstrated to perform very well when a small sampling fraction is used^{23-25,28,36}. Verdery *et al.* evaluated the most estimators so far, under both ideal and real world RDS conditions, and like other studies concluded that the most common estimators used perform roughly the same, with the exception of RDS^{IEGO}³¹. Similar to Lu (2013), he indicates that the RDS^{IEGO} estimator outperforms all others in simulated samples under both ideal and realistic recruitment scenarios, and that by relying on ego network information collected by respondents, the estimator overcomes the bias introduced by preferential recruitment and high design effect¹⁸. It presents significant implementation challenges, though, in the amount of detail required on participants' networks and still relies on some on network structure assumptions. Although much progress has been made on the development of RDS population prevalence/mean estimates, in this systematic review, no one gold standard can be identified and all RDS estimators are prone to bias introduced by the violation of any of the 6 assumptions.

Table 16: Outcome(s) of evaluations of the most commonly used RDS estimators from studies in Table 15.

Violations on RDS conditions							
Estimators	Seed selection	Differential recruitment	Recruitment effectiveness and non-response	Non-reciprocal relationship	Sampling without replacement	Misreporting degrees	Additional observation
Naive	Exhibits large bias ²⁵	Exhibits large bias ^{25, 27} Unaffected by different recruitment depths ³²	Exhibits large bias ^{25, 23}	Unaffected ³⁰	Unaffected ³⁰	Is not impacted as this estimate does not use degree information ³⁰	• Ignores RDS sample design¹² • Naïve estimator outperforms RDS-I and RDS-II when sampling frame is available²⁸ and when the population network is normally distributed³⁰
	Susceptible to substantial bias from seed selection ²⁷	Performed better than the naïve estimate under various conditions, except under the random model ²⁷	Significant bias introduced when recruitment is not completed and did not reach the number of waves calculated ²⁷	Not evaluated	If sampling fraction is very small, estimate is not impacted ¹⁴	Not evaluated	• When disease transmission and network distribution is taken into consideration, overall this estimator outperforms the naïve estimator²⁷
RDS-I	Susceptible to seed bias ³² However, when seeds selected are highly unrepresentative of the population, it outperforms RDS-II, RDS-SS and RDS-HK ₂ ²⁶	Outperforms RDS-II, RDS-HK ₂ , and RDS-SS with the absence of differential activity and large sampling fraction ²⁶ High bias can occur when non-uniform sampling occurs ³⁵	Managed to control for bias at low sampling fractions ²⁵ and at high recruitment depths ³²	Susceptible to bias on directed network ¹⁵	Induced bias to unequal edge-sampling probabilities among different network distributions ³⁵	Modest bias, larger bias was observed in ER and BA model ³⁶	• Susceptible to bottlenecks anywhere in the network²⁴ • Very sensitive to disproportionate number of recruitments from one group to the other group – heavily biased toward underestimating the population prevalence²⁶ • Overall no significant difference was observed in study³² amongst RDS-II²³ and RDS-SS • Estimate strongly associated with the network distribution³⁶

RDS-HK ₂	Seems to adjust for bias introduced by seed selection ²⁵	Differential activity leads to significant bias ²⁵	Controls for bias at low sampling fractions ²⁵	Not evaluated	Negligible bias for small sampling fractions ($\leq 20\%$) and low bias under 40% for various network distributions ³⁴	Susceptible to moderate to low bias ³⁴	• Performs similarly to RDS-I under various conditions²⁶ • Overall estimator performs well when sampling fraction is under 20% and low to moderate homophily is present (< 0.7)³⁴
RDS-II	Bias induced by seed selection, especially at low number of²⁵ Outperform RDS-I³² Under the rich-get richer/power law distribution, RDS-II adjusts for seed selection³⁰	Differential activity leads to significant bias ^{25, 29}	Bias introduced when sample size exceeds 10% of the target population ²⁹ and at low recruitment depths ³²	Susceptible to bias on directed network ¹⁵ ; but can be adjusted ²⁹	Not an issue with sample size ranging from 500 to 1000 ²⁹	Bias introduced when sample size exceeds 10 % of the target population ³² Larger bias was observed in ER and BA model ³⁶	• Susceptible to Bottlenecks anywhere in the network²⁴ • Susceptible to high homophily²⁵ • High sampling fraction leads substantial bias²⁵ • Accuracy of RDS-II under several network distributions was better than RDS-HK₂²⁶ • Estimates strongly associated with the network degree distribution³⁶
RDS-SS	Susceptible to seed bias at low recruitment depth ³²	Bias when homophily is present. ^{26, 32}	Bias at low recruitment depths ³²	Not evaluated	Not applicable as estimate considers without replacement ¹⁵	Susceptible to bias when degree reporting is incorrect ¹⁵	• Population size is required • Overall no significant difference was observed in study³² amongst RDS-I and RDS-II • Little difference was observed between RDS-II and RDS-SS³²
RDS ^{IEGO}	Does not seem to be impacted by seed selection ¹⁸	Demonstrates strong robustness ¹⁸ Surpasses all estimators in both simulated real and ideal world scenarios ³¹	Surpasses all estimators in both simulated real and ideal world scenarios ³¹	Addresses the undirected aspect of the network ¹⁸	Estimator performs well under this condition ¹⁸	Detailed, lengthy ego questions required, for more accurate degree reporting ^{18, 31} Bias is very low for misreporting degree errors. Susceptible to bias introduced by misclassifying alters ¹⁸	• Perform consistently well with varying homophily and network structures¹⁸ • Requires an extensive questionnaire that may be considered intrusive and time consuming to some participants¹⁸

6.4.3 Developed Variance Estimates

This area of RDS inference hasn't received as much attention as the RDS population prevalence estimates. Though population prevalence estimate(s) in hard-to-reach populations are always of interest to the public health sector, the lack of good variance estimates, renders the population prevalence estimates unreliable. Generally, the RDS variance estimators are generated by bootstrap resampling that approximates the RDS design and incorporates the RDS point estimate methodology. Currently, there are six variance estimators (Table 17). The methodological approach of the RDS-I variance estimate can be used to estimate variance for RDS-II and RDS-DN and RDS^{IEGO} point estimates¹⁶⁻¹⁸. Confidence intervals (CI) for all RDS point estimates can be calculated either using percentile or studentized bootstrap methods³⁷. The percentile method calculates the lower and upper bound percentiles directly from the bootstrap sampling distribution of estimates whereas the studentized bootstrap CI method employs an approximation based on the standard studentized distribution³⁷.

Table 17: Current core RDS Variance estimators and their main methodological approaches

Year	Variance Estimate	Methodological approach	Reference
2006	Salganik bootstrap (SB)	The bootstrapping procedure that splits the sample into groups based upon attribute of interest (eg. HIV infection) and iteratively resamples from these groups accordingly. This method assumes that the sampling is with-replacement.	38
2008	Volz-Heckathorn (VHE)	Incorporates the algebraic approach of the Hansen-Hurwitz estimator to account for the correlation and dependency between individuals within the network. This method assumes that the sampling is with-replacement.	14
2011	Successive Sampling (SS) - bootstrap	The bootstrapping procedure takes into account sampling without replacement. The sampling estimates the probability of population network ties across infection status groups (cross-group probabilities) based on the recruitment patterns in the original sample (Gile 2011). SS-BS samples without replacement, the cross-group probabilities are recalculated after each re-sample member is selected. This bootstrap re-sampling procedure is repeated a large number of times, the estimator of interest is computed for each, and the resulting bootstrap distribution of estimates is used to estimate the uncertainty of the estimator. The SS-BS is applied in conjunction with the Successive Sampling (SS) RDS point estimator.	15
2013	Branching bootstrap	This bootstrap is very similar to the Salganik bootstrap with the exception that it takes into account the branching structure of RDS and treats seed selection as a fixed feature of the sampling design.	39
2015	VHEwbc (with branching) VHEhom (Higher order of Markov)	Improves the VHE by accounting for the branching structure of the RDS chain (VHEwbc) and VHEhom estimates second order Markov transition patterns instead of the first order Markov (FOM)	40
2016	Tree bootstrap	A multilevel bootstrapping procedure that is applied within the hierarchical framework of the RDS recruitment tree and not on the attributes measured on the respondents. This will establish a sampling distribution, in which the statistic can be estimated from the observed RDS trees in a way that respects the dependence within the sample.	3

6.4.4 Evaluated Variance Estimates

As previously mentioned, RDS variance estimate studies have been neglected when compared to RDS means/population prevalence estimates studies. There are a total of twenty-two articles on means or population proportions but only eight articles on variance estimates. Table 17 summarizes six, as the remaining variance estimates introduced by Lu and Lu *et al.* (2013) were minor modifications applied to the existing Salganik Bootstrap to further incorporate ego network and in- and out-degrees, respectively^{16,18}. A recent paper from Spiller *et al.* (2018), published after our review period, evaluated SB and SS-B³⁷. Their results were positive and contradict those of Goel & Salganik, Wejnert, and Verdery *et al.* Spiller *et al.* demonstrate that the CI coverage rates and design effects are acceptable as long as the target populations characteristics and network structure are properly reflected in the simulations^{23,24,37}.

Table 18 highlights the variance estimator(s) evaluated, the design effect (DE), and the simulated network population. The DE is the ratio of the variance of the RDS estimator to that of the naïve estimator from a simple random sample (SRS) of the same population. For example, Salganik (2006)³⁸, recommended a DE of 2 which implies that a sample size for an RDS would require double the sample size if it was an SRS. Only simulated populations of networks are included in the evaluation of variance estimates as a real RDS network cannot be used to evaluate variance estimates. In order to identify the true variability of an estimator, one would need to conduct a large number of independent studies in the same population with the same structure, which is typically not feasible. Therefore, RDS variance estimators can only be assessed/evaluated via simulations.

Table 18: Summary of evaluated variance estimate (s) from 1997 to 2017

Study	Variance Estimator	Design Effect	Simulated population network	Main Findings
Goel & Salganik (2010) ⁴¹	VHE Salganik Bootstrap (SB)	>5	Colorado Springs Project 90 Study (Klov Dahl <i>et al.</i> , 1994) National Longitudinal Survey of Adolescent Health (Add Health)	The variance estimate performs poorly under several simulations of real networks. Even when assumptions are met, RDS mean estimates exhibit very high sampling variance.
Wejnert (2009) ²³	VHE and Salganik bootstrap (SB)	3 - 18	Two WebRDS samples of Undergraduates at the same residential university collected in 2008 and 2004 (Wejnert & Heckathorn, 2008)	- The SB algorithm used to generate RDS-I CIs underestimates the variance of groups making up less than 10% of the population to such an extent that CIs often fail to capture population parameters - CI intervals calculated using VHE easily capture population parameters, but tend to overestimate variance of small groups to such an extent that design effects above 20 are observed.
Verdery <i>et al.</i> (2015) ⁴⁰	VHEwbc (with branching) VHEhom (Higher order of Markov) Salaganik bootstrap (SB)	15 - 30	National Longitudinal Survey of Adolescent Health (Add Health) (115 networks) Facebook 100 datasets	- Conflicting results were discovered when FOM process was tested among variables on the full network vs the RDS sample. The full network doesn't satisfy the FOM assumption and yet the sample indicates otherwise, which greatly impacts the reliability of the RDS variance estimate. - SB outperforms VHE in most situations. - VHEwbc and VHEhom outperform VHE, but VHEhom underperforms compared with the VHEwbc. - The improvements made to VHE did show progress but overall all variance estimates seem to underperform and suffer from limitations in most of the situations tested.
Baraff <i>et al.</i> (2016) ³	Naïve, VHE, SB, SS-B and Tree Bootstrap	Estimate can handle high DE	Colorado Springs Project 90 Study (Klov Dahl <i>et al.</i> , 1994)	Demonstrated that the existing methods for variance estimation: VHE, SB, SS-B,

			National Longitudinal Survey of Adolescent Health (Add Health) RDS data collected from IDUs in 26 Ukrainian cities	fail to capture the variability in RDS estimates. • The Tree bootstrap can perform properly when sampling without replacement. • Tree bootstrap captured the large variability in samples when estimating attributes with high design effects.
--	--	--	---	---

Based on these findings, the tree bootstrap variance estimate appears to be a promising variance estimator. It is the only estimator that does not model RDS as a FOM process (with the exception of VHE hom) and requires independent analysis for each attribute of recruitment interest as the bootstrapping is based on re-sampling through the RDS tree structure of recruitment⁴⁰. Verdery *et al.* demonstrate that the FOM assumptions are violated from the various ADD Health and Facebook simulated population networks, which is known to lead to substantially biased RDS variance estimators⁴⁰. The Tree bootstrap can be used for various RDS point estimates and in networks that do not conform to the FOM process³.

6.5 Discussion

The main focus of this review was to collate all findings related to RDS inference in order to assist researchers in identifying factors they need to consider when conducting and making inferences on RDS data in hard-to-reach populations. As there are various different areas that can help improve RDS inference, this review focused on the two most common components of population parameters: mean and variance. A total of 32 articles published from 1997 to October 2017 were retained, the majority of which focused on the RDS mean estimators. This is the first comprehensive review of the developments in both the RDS means and variances.

RDS relies on respondents to recruit members of a hidden population. The methodology has proven to be quite effective in obtaining a large enough sample size in multiple applications worldwide. However, when it comes to making reliable and valid inferences on RDS data to the population of interest, the assumptions associated with the statistical component of the RDS process make it very challenging. In the past two decades, RDS estimators have emerged with conditions less stringent than the ones first associated with the recruitment process and of RDS methodology. Many of the developed RDS estimators rely on the assumption that the recruitment process is a MP model; however, this type of process assumes that sampling is with replacement and that the sample size attains an equilibrium in which bias from seed selection would be considered negligible, both of which have been shown to be violated. The RDS-SS is very similar to the RDS-II estimator and the naïve estimator depending on the sample size used. When the sample size is a very small fraction of the population, the RDS-SS is similar to the RDS-II estimator. RDS-II assumes that sampling probabilities are proportional to the degree while the RDS-SS estimator substitutes a function of the degree based on approximating the sampling process as successive³¹. The RDS-SS estimator assumes the recruitment process is without replacement. The main drawback of this assumption is that the population size must be known. However, if the whole population is sampled, then the RDS-SS behaves like the naïve estimator. Two new estimators were developed which do not involve Markov chain theory but require an estimate of population size. The RDS-MA estimator goes one step further by discarding the assumption of the seed dependence and network structure. The RDS-Count process model completely discards the random walk and models the recruitment process like an epidemic model.

The main issues with real RDS data are that it is very difficult to assess whether some of the respondents' assumptions are true and which estimators perform best when the true parameters in hard-to-reach populations are unknown. Simulations play a pivotal role in RDS evaluation by assessing complex statistical methods in populations with known parameters, especially when it is not feasible to fully standardize operations and procedures in real-life. The usefulness and accuracy of simulation studies usually relies on the soundness of methods, procedures and realistic choice of network parameters. Past simulations studies on RDS estimators have demonstrated reductions in bias with each new estimator that has been developed, suggesting that each estimator produced different results. However, the majority of the studies have pointed to the same conclusion, that even though these estimators may reduce bias when certain RDS conditions are violated, overall these estimators perform similarly to each other. Using both real and ideal RDS conditions, Lu (2013) and Verdery *et al* (2016), demonstrates that RDS^{IEGO} , which is similar to RDS-II estimator but relies on the ego network information collected by respondents, overcomes the bias introduced from preferential recruitment and high design effect^{18,31}.

In most RDS studies, the reason for the estimator selection is rarely discussed in the methods. In many cases, selection may be based more on preference for statistical software than consideration for the estimator itself. The majority of the studies also lack an assessment of the quality of the RDS data; thus, neglecting the numerous assumptions that may be violated. As such, we suggest all researcher to use the diagnostic tools, developed by Gile *et al* (2015), to better detect and understand the assumptions violated with their RDS data, and how the bias induced would impact the RDS estimates⁸. The diagnostic tools can help assess the impact of: sampling without-replacement, failure to attain the desired sample size, seed dependence,

bottleneck, reciprocation of ties, degree accuracy, and recruitment effectiveness. In addition, Gile *et al.* provide qualitative improvements and recommendations that can be incorporated in the questionnaire provided to respondents during the RDS data collection process. In fact, providing naïve proportions or prevalence of the sample itself, in addition to a detailed section of the findings produced from the diagnostic tools, would be more informative and provide a better understanding of the hard-to-reach population, than the RDS estimates produced.

Most researchers, however, are still interested in providing estimates using RDS estimators, even when the assumptions have not been met. In this case, we recommend that the data be assessed to identify which RDS estimators are more appropriate for your data (Table 16). Using a few RDS estimators rather than one, can also help to identify whether the estimate produced is consistent and may be reflective of the population estimate. We have summarised the options and theories behind each estimator and recommend the following strategies in selecting from the measures. For instance, if you know that your network likely follows a mechanism of preferential recruitment such as film actors who star in many movies are likely to be recruited for more movies based on their popularity and “draw”, you may consider choosing RDS II. If you know that an outbreak of syphilis started in people who inject drugs (PWID) and then moved into men who have sex with men, then you may avoid measures susceptible to bias from bottlenecks. RDS^{IEGO} in this past example would appear to be a better option. However, practicality is vitally important in selecting measures. If the network detail required results in lengthy and complex questionnaires, it is unlikely that PWID may answer them. Likewise if the population size of PWID is unknown, but is required to estimate the population proportions as in RD-SS, that method is unlikely to be helpful. This last is particularly challenging as the size of many hidden populations is almost always unknown, by definition. Finally, you may use a

second measure which is commonly used such as RDS II, or at least commonly used in studies of similar populations, to facilitate comparisons across studies. By selecting measures which emulate the study conditions, avoiding those where obvious bias in recruitment is not adjusted, and using a second, more commonly used measure in addition the first, the reader may be better able to justify which methods they used, and why.

Despite the on-going improvements in the development of RDS estimators, the evaluation of how accurately these estimators truly represent the target population is still debatable and even though RDS^{IEGO} shows promising results, no clear gold standard estimator has been identified. The value of a point estimate lies in a good variance estimate, which gives the researcher a range of error around the point estimate, useful in calculating appropriate sample sizes for further studies. Several of the early RDS studies recognized that the variability of RDS estimates was not fully understood. It was first believed that a design effect of 2 was needed for RDS studies and that the Salganik bootstrap produces asymptotically biased estimates³⁸. However, Goel and Salganik demonstrated that there was very high variability associated with the most frequent RDS estimators, even when RDS conditions were met²³. Goel and Salganik also demonstrated that the design effect is much bigger than 2, and the variance estimators produced low coverage CI⁴¹. Many studies after Goel and Salganik agreed with their findings^{16-18,39,40}. However, a study by Spiller *et al.*, not included in the review, concluded that the variance estimators (SB and SS-B) provided good CI coverage and a $DE < 3$, when taking into account sampling without replacement and using the studentized bootstrap for calculating CI instead of the percentile bootstrap³⁷.

The majority of the variance estimators rely on the FOM process for estimating the variance of any attribute of interest within the network. Though the FOM process has been an appropriate model for the recruitment of individuals within a social network, it has been demonstrated that the moment you stratify the network by an attribute, the FOM assumption is violated, which leads to substantial bias in the estimation of variance³⁹. It is also important to note that even when FOM assumptions have been met, the variance estimates produced are greatly underestimated⁴⁰. Baraff *et al.* developed a variance estimate, known as the Tree bootstrap, that is based on the resampling of the structure of the RDS recruitment process and does not rely on any attributes collected; thus, bypassing the FOM condition³.

The findings of this review can provide guidance to beginner RDS users. Though not all studies' findings coincide with each other or address all RDS concerns, the general consensus is that all estimates perform more or less in the same way. Verdery *et al.* had the most comprehensive list of estimators in their evaluation study and they demonstrated that RDS^{IEGO} did out-perform the other estimators in various conditions and was robust to differential recruitment as well as varying homophily and network distribution, which corresponded to the findings of Lu (2013)³¹. RDS^{IEGO} requires detailed ego network information for all variables of interest and still relies, to a certain extent, on assumptions related to the respondent's behaviour and network structure (i.e. degree is reported accurately and that each individual is connected directly or indirectly to every individual in the network). This extensive questionnaire isn't necessarily included in all RDS conducted; thus, authors of the studies which have included the ego-network component are strongly encouraged to use this estimator. This would consist of asking participants about their alters and information regarding behavioural and risk associated factors with each of their alters, including infectious status. Those authors who have not included

ego network information in their questionnaire may consider using any other estimators but will need to be cautious with the application and be transparent with the limitation of this estimator. In RDS studies, RDS-II appears to be the most frequent estimator used^{23-25,28,36}. Out of the 22 studies evaluating RDS estimates, 15 of them assessed RDS-II. The general conclusion of RDS-II is that it performs well in low sampling fraction and low to moderate homophily networks, and is susceptible to large bias due to differential recruitment. The Tree bootstrap variance estimate appears to be the most promising and can be applied to any of the mean estimates.

6.6 Conclusion

This systematic review can assist researchers in identifying the appropriate RDS mean and variance estimators for use in their studies by providing an overview of the current work on RDS inference. To our knowledge, this is the first the paper that collates the findings associated with both RDS mean and variance to provide a better understanding of the fundamentals of RDS inference. Ideally, including a meta-analysis which would compare all existing estimators in both real and ideal RDS conditions using simulation studies, based on different hard-to-reach populations with various parameters and network distributions, may provide more information. This type of study would require substantial time and collaborations from various RDS experts but may be essential in defining best methods of estimation. However, RDS^{IEGO} and Tree bootstrapping estimators appear to be quite promising in providing more reliable estimates of mean and variance, respectively, in hard-to-reach populations due to the reduced number of assumptions associated with these estimators. However, since in most hard-to-reach populations, these assumptions are violated, our recommendation is to: (1) focus on the quality of the RDS questionnaire and data, (2) provide a detailed evaluation of the RDS data (using recommended diagnostic tools by Gile *et al.* (2015)⁸), (3) identify the network structure of the target

population, and (4) use various RDS estimators (including the naïve estimator) that are most appropriate to your RDS data to determine if consistency is present. These suggested steps will provide you with a comprehensive picture of the hard-to-reach population of interest.

6.7 References

1. Berchenko Y, Frost SD. Capture-recapture methods and respondent-driven sampling: Their potential and limitations. *Sex Transm Infect.* 2011;87(4):267-268.
2. Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Probl.* 1997;44(2):174-199.
3. Baraff A.J., McCormick T.H., Raftery A.E. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proc Natl Acad Sci U S A.* 2016;113(51):14668-14673.
4. Chua A.C., Chen M.I.C., Cavailler P., et al. Challenges of respondent driven sampling to assess sexual behaviour and estimate the prevalence of human immunodeficiency virus (HIV) and syphilis in men who have sex with men (MSM) in Singapore. *Ann Acad Med Singapore.* 2013;42(7):350-353..
5. Clark JL, Konda KA, Silva-Santisteban A, et al. Sampling methodologies for epidemiologic surveillance of men who have sex with men and transgender women in Latin America: An empirical comparison of convenience sampling, time space sampling, and respondent driven sampling. *AIDS Behav.* 2014;18(12):2338-2348.
6. Gyarmathy V.A., Johnston L.G., Caplinskiene I., Caplinskas S., Latkin C.A. A simulative comparison of respondent driven sampling with incentivized snowball sampling - the "strudel effect". *Drug Alcohol Depend.* 2014;135(1):71-77.
7. Johnston LG, Chen YH, Silva-Santisteban A, Raymond HF. An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS Behav.* 2013;17(6):2202-2210.
8. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *J R Stat Soc Ser A Stat Soc.* 2015;178(1):241-269.
9. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ.* 2009;339.
10. Pai M, McCulloch M, Gorman JD, et al. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. *Natl Med J India.* 2004;17(2):86-95.
11. Heckathorn DD. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Soc Probl.* 2002;49(1):11-34.
12. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol Methodol.* 2004;34(1):193-240. Accessed 6 October 2017. doi: 10.1111/j.0081-1750.2004.00152.x.

13. Heckathorn DD. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociol Methodol.* 2007;37(1):151-207.
14. Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics.* 2008;24(1):79-97.
15. Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J Am Stat Assoc.* 2011;106(493):135-146.
16. Lu X, Malmros J, Liljeros F, Britton T. Respondent-driven sampling on directed networks. *Electron J Stat.* 2013;7(1):292-322.
17. Malmros J, Masuda N, Britton T. Random walks on directed networks: Inference and respondent-driven sampling. *J Off Stat.* 2016;32(2):433-459.
18. Lu X. Linked ego networks: Improving estimate reliability and validity with respondent-driven sampling. *Soc Netw.* 2013;35(4):669-685.
19. Gile KJ, Handcock MS. Network model-assisted inference from respondent-driven sampling data. *J R Stat Soc Ser A Stat Soc.* 2015;178(3):619-639.
20. Aronow PM, Crawford FW. Nonparametric identification for respondent-driven sampling. *Stat Probab Lett.* 2015;106:100-102.
21. Selvaraj V., Boopathi K., Paranjape R., Mehendale S. A single weighting approach to analyze respondent-driven sampling data. *Indian J Med Res.* 2016;144:447-459.
22. Berchenko Y, Rosenblatt JD, Frost SDW. Modeling and analyzing respondent-driven sampling as a counting process. *Biometrics.* 2017.
23. Wejnert C. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol Methodol.* 2009;39(1):73-116. A
24. Goel S., Salganik M.J. Respondent-driven sampling as markov chain monte carlo. *Stat Med.* 2009;28(17):2202-2229.
25. Gile KJ, Handcock MS. Respondent-driven sampling: An assessment of current methodology. *Sociol Methodol.* 2010;40(1):285-327.
26. Tomas A, Gile KJ. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron J Stat.* 2011;5:899-934.
27. Albuquerque EM, Codeço CT, Bastos FI, eds. *Assessing respondent-driven sampling in the estimation of the prevalence of sexually transmitted infections (STIs) in populations structured in complex networks.* Rio de Janeiro ed. ; 2011 Communications in Computer and Information Science; 2nd International Workshop on Complex Networks, CompleNet 2010; No. 116 CCIS.

28. McCreesh N., Frost S.D.W., Seeley J., et al. Evaluation of respondent-driven sampling. *Epidemiology*. 2012;23(1):138-147.
29. Lu X, Bengtsson L, Britton T, et al. The sensitivity of respondent-driven sampling. *J R Stat Soc Ser A Stat Soc*. 2012;175(1):191-216.
30. Nesterko S, Blitzstein J. Bias–variance and breadth–depth tradeoffs in respondent-driven sampling. *J Stat Comput Simul*. 2015;85(1):89-102.
31. Verdery A.M., Merli M.G., Moody J., Smith J.A., Fisher J.C. Brief report: Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in china. *Epidemiology*. 2015;26(5):661-665.
32. Wirtz A.L., Mehta S.H., Latkin C., et al. Comparison of respondent driven sampling estimators to determine HIV prevalence and population characteristics among men who have sex with men in moscow, russia. *PLoS ONE*. 2016;11(6):e0155519.
33. Lachowsky NJ, Sorge JT, Raymond HF, et al. Does size really matter? A sensitivity analysis of number of seeds in a respondent-driven sampling study of gay, bisexual and other men who have sex with men in Vancouver, Canada. *BMC Med Res Methodol*. 2016;16(1)
34. Barash VD, Cameron CJ, Spiller MW, Heckathorn DD. Respondent-driven sampling-testing assumptions: Sampling with replacement. *J Off Stat*. 2016;32(1):29-73.
35. Ott MQ, Gile KJ. Unequal edge inclusion probabilities in link-tracing network sampling with implications for respondent-driven sampling. *Electron J Stat*. 2016;10(1):1109-1132.
36. Sperandei S, Bastos LS, Ribeiro-Alves M, Bastos FI. Assessing respondent-driven sampling: A simulation study across different networks. *Soc Netw*. 2017.
37. Spiller MW, Gile KJ, Handcock MS, Mar CM, Wejnert C. Evaluating variance estimators for respondent-driven sampling. *Journal of Survey Statistics and Methodology*. 2018;6(1):23-45.
38. Salganik M.J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Urban Health*. 2006;83(1).
39. Yamanis TJ, Merli MG, Neely WW, et al. An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in china. *Sociological methods & research*. 2013;42(3).
40. Verdery AM, Mouw T, Bauldry S, Mucha PJ. Correction: Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE*. 2016;11(1):e0148006.
41. Goel S., Salganik M.J. Assessing respondent-driven sampling. *Proc Natl Acad Sci U S A*. 2010;107(15):6743-6747.

7.0 Discussion

Data collected from a population where no sampling frame exists may not be representative of the population, leading to questions of reliability and validity of estimates, which in turn may lead to inaccurate inferences made about that population. Hard-to-reach or hidden populations typically have no sampling frame. Respondent-driven sampling (RDS), a methodology that is influenced by both snowball sampling and traditional contact tracing, relies on respondents to recruit fellow members of the hidden population network. The overall objective of the thesis was to address some of the gaps currently present in respondent-driven sampling (RDS) inference. The three manuscripts provided insight on the challenges of network structure within the populations and statistical approaches to obtain reliable estimates from hard-to-reach populations, in particular on people who inject drugs (PWIDs) in Ottawa, Ontario.

RDS methodology has proven to be effective in obtaining large enough sample sizes in multiple applications worldwide ¹⁻⁷. However, when it comes to making reliable and valid inferences from RDS data to the population of interest, the assumptions associated with statistical component of RDS processes present challenges. Throughout the past few decades, RDS mean and variance estimators have emerged with conditions less stringent than the ones first associated with the recruitment process and of RDS methodology. The main issues concerning the reliability of RDS inference are the lack of knowledge of the underlying network structure, characteristics of the population and size, and assumptions associated with recruitment process.

The first two studies focused on different aspects of the hard-to-reach population in order to better understand the network topology or structure of the spread of disease within the

network of PWIDs in Ottawa. In the first paper, we used the frequency of injection as a proxy for disease transmission and assessed whether the degree distribution or number of exposures to different people exhibited a power law form, better known as a scale-free network. Random and exponential models have been, and are also currently being used to study disease spread among social contact networks. Our findings have illustrated that the network distribution in PWID in Ottawa does not conform to random nor exponential models; both log normal and scale-free networks fit the PWID RDS data.

The second paper focused on the social contacts and contact interaction, in the network structure of PWID in Ottawa to understand the spread of the disease. We were unable to fully determine whether the frequency of injection within the Ottawa PWID network is scale-free, but we demonstrated that the social contacts within our target population exhibit a small world network. Scale-free networks are small world networks; however, not all small world networks are scale-free. Several studies have demonstrated that small-world networks are more susceptible to epidemics than a random and regular network; scale-free networks, however, lack a critical threshold of connectedness below which pathogens cannot spread. This lack of critical threshold can lead to propagated spread of diseases and epidemics even with known weaker infectious pathogens. Clusters, an essential component of small world networks, and hubs (nodes with many more contacts than others) associated with scale-free networks, are crucial factors in both the success of disease transmission and efficient interventions. Our work has demonstrated that high clustering, with short average path length, and hub formation were present in the Ottawa PWID network.

Public health interventions aimed at preventing or reducing the spread of infection requires us to understand the network structure. Epidemiologically, assessing whether a critical threshold exists for disease transmission within a network would have major implications in preventing, reducing and controlling the spread of infectious diseases. Interventions directed to random individuals in a small world and/or scale-free network are inefficient and expensive. Selection of random people or places for intervention as being representative of a skewed distribution network is ineffective in altering the threshold-free diffusion process. However, interventions targeting the clusters and the small number of individuals with a large number of injections and/or contacts are a better way of preventing or limiting the spread of disease rather than targeting everyone. Our network structure/topology findings also have important implications in the statistical methodology that we use to make inference in the PWID population.

The last paper included a systematic review on the development of RDS inference since the introduction of RDS in 1997. The purpose was to provide guidance for RDS users on the appropriate estimators for their particular RDS study and determine whether there was a gold standard RDS mean and variance estimator. The reliability and validity of RDS mean and variance estimators depend not only the assumptions associated with the data collection process but also on the network structure. Several RDS studies have demonstrated that RDS estimators are not robust under random network structures and when certain RDS assumptions are violated. The majority of researchers publishing RDS studies do not describe the network structure of their target population before RDS estimators are applied to their data. Our systematic review on the RDS inferences studies, which focused on RDS mean and variance estimators, illustrated that certain RDS mean estimators out-performed others with specific network structure. For instance,

RDS II, the most common mean estimator used in RDS studies, performs better when applied to small-world network structures than in random network structures. Our findings highlighted that RDS^{IEGO} and Tree bootstrapping mean and variance estimators, respectively, outperform other estimators and have demonstrated to be associated with less assumptions and robust under various network distribution, including small-world networks, when conditions are met.

Methods to assess RDS data were not included in the systematic review as it was out of the scope of the initial research question. One study in particular stood out that provided a step by step process of conducting RDS (including recommended questions to be included in the data collection) and evaluating the quality of RDS data. Gile *et al.* (2015) developed diagnostics based on three specific features of RDS studies: time sequences of responses, contact with respondents who visited the study site twice due to dual incentive (respondents receive incentive for participating and for recruiting), and the multiple seeds used to begin the recruitment process³. With these features taken into consideration, the diagnostic methods focused on addressing and detecting: the impacts of sampling without-replacement, seed bias, reciprocation of ties, failure to attain the desired sample size, bottleneck, reciprocation of ties, degree accuracy, and recruitment effectiveness. Ideally, if these diagnostic procedures are conducted in parallel with the RDS process, they will provide insight of the actual recruitment process and identify violations of assumptions while it can still be corrected. If this is not possible, there are still recommended diagnostics that should be applied after the completion of the RDS recruitment process. Giles *et al.* (2015) used the developed diagnostic tools in 12 RDS studies conducted in 4 large cities in the Dominican Republic. They concluded by strongly recommending that the reciprocation rate and non-response rate be analyzed; and provide adjustments for the analysis of RDS data. Furthermore, they recommend the Motivation-Outcome analysis, to assess the

violations of sampling with replacement. The Motivation-Outcome analysis is used to determine whether the recruitment process was affected by sampling high rate of individuals in a well-connected sub-group and whether the study population, as whole, is sampled at high enough rate.

Overall, the systematic review demonstrated that all RDS estimators are susceptible to significant bias. Nonetheless, RDS data provides a lot of relevant information that contributes to the comprehensive understanding of hard-to-reach populations. When RDS is performed properly, it produces significantly more details about the demographics and behaviours, as well as the social network existing within that population, that are not captured by traditional non-random data methods (i.e. convenience, targeted and venue-based sampling). If new RDS users are interested in obtaining population estimates, it is recommended to first assess the RDS data, using developed diagnostic tools by Gile *et al.* (2015)³, identify the network structure, and then use multiple RDS estimators most appropriate to the RDS data in order to identify whether RDS estimates produced are consistent with each other.

7.1 References

1. Verdery A.M., Merli M.G., Moody J., Smith J.A., Fisher J.C. Brief report: Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in china. *Epidemiology*. 2015;26(5):661-665.
2. Gile KJ, Handcock MS. Respondent-driven sampling: An assessment of current methodology. *Sociol Methodol*. 2010;40(1):285-327.
3. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *J R Stat Soc Ser A Stat Soc*. 2015;178(1):241-269.
4. Abdul-Quader A.S., Heckathorn D.D., Sabin K., Saidel T. Implementation and analysis of respondent driven sampling: Lessons learned from the field. *Urban Health*. 2006;83(1).
5. Barros AB, Dias SF, Martins MRO. Hard-to-reach populations of men who have sex with men and sex workers: A systematic review on sampling methods. *Syst Rev*. 2015;4(1). Accessed 6 October 2017. doi: 10.1186/s13643-015-0129-9.
6. Berchenko Y., Frost S.D.W. Capture-recapture methods and respondent-driven sampling: Their potential and limitations. *Sex Transm Infect*. 2011;87(4):267-268.
7. Chen L, Crawford FW, Karbasi A. Seeing the unseen network: Inferring hidden social ties from respondent-driven sampling. *AAAI Conf Artif Intell* , *AAAI*. 2016:1174-1180.

8.0 Conclusion

We recommend modified statistical methods and strategies to select those that incorporate the network degree distribution of the target population. They better correspond with the RDS recruitment process and will improve the overall reliability of RDS inference. This in turn will allow us to better understand the disease spread within networks and identify key components of the specific network to implement strategic interventions. In figure 11, we highlight that the egocentric data should be included in the questionnaires provided to the respondents. In addition, diagnostic tools should be used in parallel to RDS data collection in order to detect and correct for bias. The RDS sample collected from the PWID population can then be used to identify the network topology which should be considered when conducting RDS inference. Diagnostic tools should be assessed after the RDS recruitment. If researchers find it necessary to calculate population estimates, despite the limitations associated with them, we recommend RDS^{IEGO} and Tree bootstrapping mean and variance estimators. Based on our content analyses in the systematic review, these estimators rely on fewer assumptions and have demonstrated to out-perform the other estimators. In addition, including non-random sampling methodologies, such as time-location sampling can provide further information on the population characteristics; contribute to the topology of the networks; and may provide missing pieces of the puzzle to understand realistic interactions between these individuals. The topology and appropriate RDS estimators will also help provide a better understanding of disease spread within a network that strong relies on social contacts and will provide the necessary information needed to design public health policies in various hard-to-reach populations.

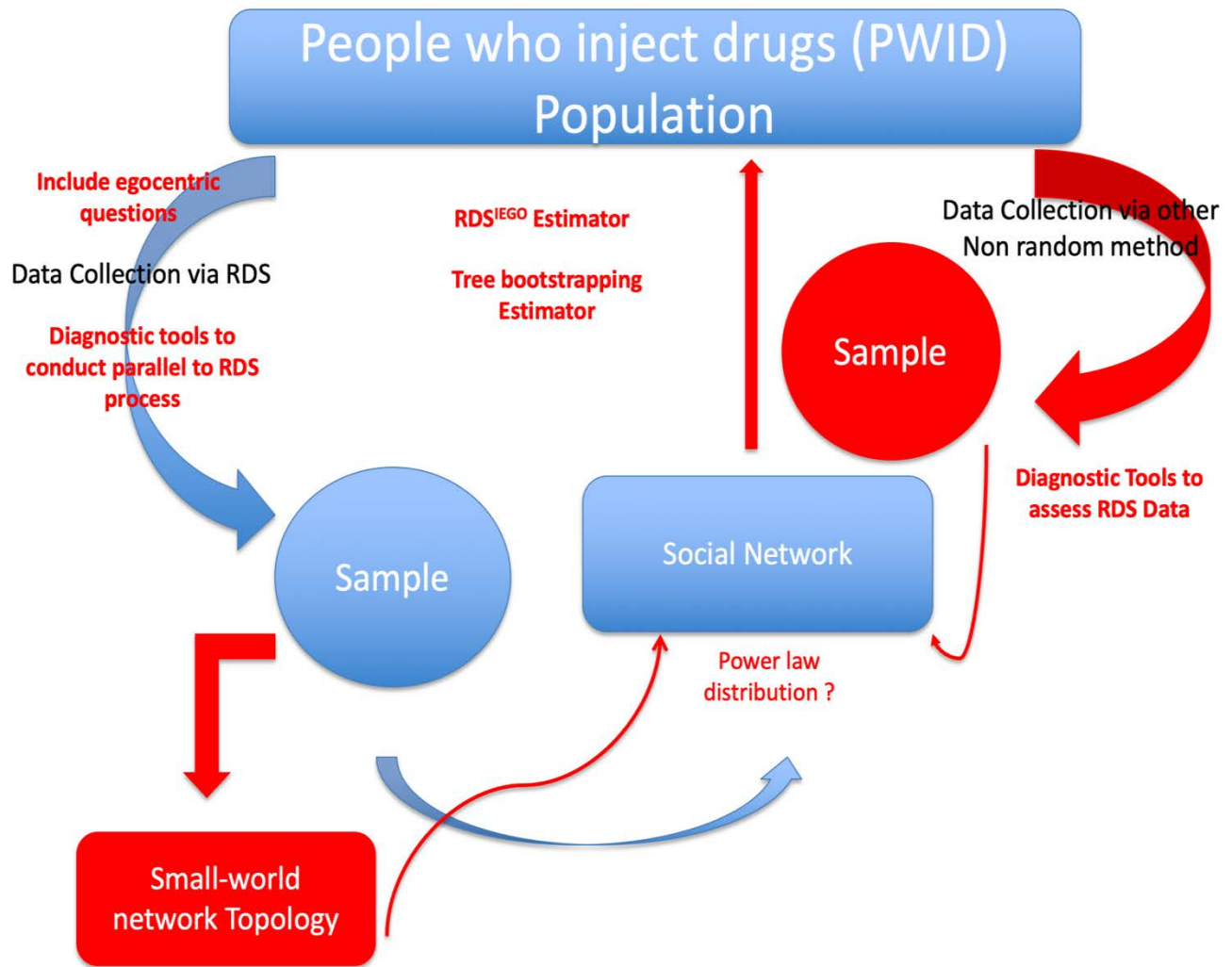


Figure 11: Highlights of the thesis findings and its potential contribution and recommendations to RDS inference in people who inject drugs (highlighted in red). Adapted from Saglanik & Heckathorn.

9.0 References

- Abdul-Quader A.S., Heckathorn D.D., Sabin K., Saidel T. Implementation and analysis of respondent driven sampling: Lessons learned from the field. *Urban Health*. 2006.83(1).
- Adler MW, Weller IV. AIDS: Sense not fear. *Br Med J (Clin Res Ed)*. 1984;288(6425):1177-1178.
- Albert R, Barabási A. Statistical mechanics of complex networks. *Reviews of Modern Physics*. 2002.74.
- Albert R, Barabási. The physics of the web. *Physics World*. 2001.14(7):33.
- Albuquerque EM, Codeço CT, Bastos FI, eds. Assessing respondent-driven sampling in the estimation of the prevalence of sexually transmitted infections (STIs) in populations structured in complex networks. Rio de Janeiro ed. ; 2011Communications in Computer and Information Science; 2nd International Workshop on Complex Networks, CompleNet 2010. No. 116 CCIS.
- Alstott J, Bullmore E, Plenz D. Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS One*. 2014.9(1):e85777.
- Al-Azem A. *Social network analysis in tuberculosis control among the aboriginal population of manitoba*. [Ph.D.]. University of Manitoba; 2006.
- Amaral LAN, Scala A, Barthélémy M, Stanley HE. Classes of small-world networks. In: Princeton University Press; 2011.207-210.
- Amaral LA, Scala A, Barthelemy M, Stanley HE. Classes of small-world networks. *Proc Natl Acad Sci U S A*. 2000.97(21):11149-52.
- Ancel Meyers L, Newman ME, Martin M, Schrag S. Applying network theory to epidemics: Control measures for mycoplasma pneumoniae outbreaks. *Emerg Infect Dis*. 2003.9(2):204-210.
- Aronow PM, Crawford FW. Nonparametric identification for respondent-driven sampling. *Stat Probab Lett*. 2015.106:100-102.
- Auerbach DM, Darrow WW, Jaffe HW, Curran JW. Cluster of cases of the acquired immune deficiency syndrome. patients linked by sexual contact. *Am J Med*. 1984;76(3):487-492.
- Averting HIV and AIDS. People who inject drugs, HIV and AIDS. AVERT. 2017.
- Bansal S, Grenfell BT, Meyers LA. When individual behaviour matters: Homogeneous and network models in epidemiology. *J R Soc Interface*. 2007.4(16):879-891.

- Barabási A, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509- 512.
- Barabási A, Regan E, Vicsek T. Deterministic scale-free networks. *Physics A: Statistical Mechanics and its Applications*. 2001. Vol 299:559-564.
- Barabási A. Scale-free networks: A decade and beyond. *Science*. 2009.325(5939):412-413.
- Barabási A, Bonabeau E. Scale-free networks. *Sci Am*. 2003.288(5):60-69.
- Baraff A.J., McCormick T.H., Raftery A.E. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proc Natl Acad Sci U S A*. 2016.113(51):14668-14673.
- Barash VD, Cameron CJ, Spiller MW, Heckathorn DD. Respondent-driven sampling-testing assumptions: Sampling with replacement. *J Off Stat*. 2016.32(1):29-73
- Barros AB, Dias SF, Martins MRO. Hard-to-reach populations of men who have sex with men and sex workers: A systematic review on sampling methods. *Syst Rev*. 2015.4(1).
- Barzel B, Barabási A. Universality in network dynamics. *Nat Phys*. 2013.
- Bearman PS, Moody J, Stovel K. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*. 2004.110(1):44-91.
- Berchenko Y, Frost SD. Capture-recapture methods and respondent-driven sampling: Their potential and limitations. *Sex Transm Infect*. 2011.87(4):267-268.
- Berchenko Y, Rosenblatt JD, Frost SDW. Modeling and analyzing respondent-driven sampling as a counting process. *Biometrics*. 2017.
- Berchenko Y., Frost S.D.W. Capture-recapture methods and respondent-driven sampling: Their potential and limitations. *Sex Transm Infect*. 2011.87(4):267-268.
- Blum K, Han D, Modestino EJ, et al. A systematic, intensive statistical investigation of data from the comprehensive analysis of reported drugs (CARD) for compliance and illicit opioid abstinence in substance addiction treatment with buprenorphine/naloxone. *Subst Use Misuse*. 2018.53(2):220-229.
- Brady K.A., Nassau T.B., Shinefeld J., Mezzacappa C. Analyzing trends in HIV risks for injection drug users by respondent-driven sampling. *Top Antiviral Med*. 2015.23:473-474.
- Brauer F. An introduction to networks in epidemic modeling. *Journal of Mathematical Epidemiology*. 2008. Vol 1945.
- Broido A,D., Clauset A. Scale-free networks are rare. *Physics and Society*.2018. Canale MH. "Small-world" network. *Medico e Bambino*. 2009.28(7):459-460.

CATIE. Hepatitis C treatment combinations. <http://www.catie.ca/en/practical-guides/hepc-in-depth/treatment/choosing-drug-combination-chronic-hepatitis-c>. Updated 2018. Accessed 02/18, 2018.

Cauchemez S, Bhattarai A, Marchbanks TL, et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc Natl Acad Sci USA*. 2011;108(7):2825-2830.

Centola D, Macy M. Complex contagions and the weakness of long ties. *American Journal of Sociology*. 2007;113(3):702-734.

Centola D. The spread of behavior in an online social network experiment. *Science*. 2010;329(5996):1194-1197.

Chakraborty A, Manoj BS. The reason behind the scale-free world. *IEEE Sensors Journal*. 2014;14(11):4014-4015.

Champernowne DG. A model of income distribution. *The Economic Journal*. 1953;63(250):318- 351.

Chen L, Crawford FW, Karbasi A. Seeing the unseen network: Inferring hidden social ties from respondent-driven sampling. *AAAI Conf Artif Intell*, AAAI. 2016. pp.1174-1180.

Chen X, Zhu L, Zhou YH, et al. Factors associated with needle sharing among people who inject drugs in yunnan, china: A combined network and regression analysis. *Infect Dis Poverty*. 2016;5(1):73-016-0169

Christley RM, Pinchbeck GL, Bowers RG, et al. Infection in social networks: Using network analysis to identify high-risk individuals. *Am J Epidemiol*. 2005;162(10):1024-1031.

Chua A.C., Chen M.I.C., Cavailler P., et al. Challenges of respondent driven sampling to assess sexual behaviour and estimate the prevalence of human immunodeficiency virus (HIV) and syphilis in men who have sex with men (MSM) in singapore. *Ann Acad Med Singapore*. 2013;42(7):350-353.

Clark JL, Konda KA, Silva-Santisteban A, et al. Sampling methodologies for epidemiologic surveillance of men who have sex with men and transgender women in latin america: An empiric comparison of convenience sampling, time space sampling, and respondent driven sampling. *AIDS Behav*. 2014;18(12):2338-2348.

Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data . *SIAM Review*. 2009;51(4):661-703.

Cooke RM, Nieboer D. Heavy-tailed distributions: Data, diagnostics, and new developments. *Resources for the Future Discussion Paper*. 2011.

Csányi G, Szendroi B. Fractal-small-world dichotomy in real-world networks. *Phys Rev E Stat Nonlinear Soft Matter Phys*. 2004;70(1 2):016122-1-016122-5.

D'Angelo-Scott H, Cutler J, Friedman D, Hendriks A, Jolly AM. Social network investigation of a syphilis outbreak in ottawa, Ontario. *The Canadian journal of infectious diseases & medical microbiology = Journal canadien des maladies infectieuses et de la microbiologie medicale*. 2015;26(5):268-272.

De P, Connor N, Bouchard F, Sutherland D. HIV and hepatitis C virus testing and seropositivity rates in Canadian federal penitentiaries: A critical opportunity for care and prevention. *Can J Infect Dis Med Microbiol*. 2004.15(4):221-225.

Demirel G, Barter E, Gross T. Dynamics of epidemic diseases on a growing adaptive network. *Scientific Reports*. 2017.7:42352.

Dennie D, Cuccia LA. "Six degrees of separation"-revealing a "small-world phenomenon" through a chemistry literature search activity. *J Chem Educ*. 2014;91(4):546-549.

Doherty IA. Sexual networks and sexually transmitted infections: Innovations and findings. *Curr Opin Infect Dis*. 2011.24(1):70-77.

Dombrowski K, Curtis R, Friedman S, Khan B. Topological and historical considerations for infectious disease transmission among injecting drug users in Bushwick, Brooklyn (USA). *World J AIDS*. 2013.3(1):1-9.

Duncan I, Curtis R, Reyes JC, Abadie R, Khan B, Dombrowski K. Hepatitis C serosorting among people who inject drugs in rural puerto rico. *Preventive medicine reports*. 2017;6:38-43.

Eguiluz VM, Klemm K. Epidemic threshold in structured scale-free networks. *Phys Rev Lett*. 2002.89(10):108701.

Enright J, Kao R. Epidemics on dynamic networks. *Epidemics*. 2018. Vol 24

Fazekas I, Porvaszsyik B. Scale-free property for degrees and weights in a preferential attachment random graph model. *Journal of Probability and Statistics*. 2013.

Friedman SR, Flom PL, Kottiri BJ, et al. Drug use patterns and infection with sexually transmissible agents among young adults in a high-risk neighbourhood in new york city. *Addiction*. 2003.98(2):159-169.

Gabaix X. Power laws in economics and finance. *Annu Rev Econ*. 2009;1(1):255-294.

Gallos LK, Makse HA, Sigman M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks; 22308319. *Proc Natl Acad Sci U S A*. 2011.109(8):2825-2830.

Gile KJ, Handcock MS. Network model-assisted inference from respondent-driven sampling data. *J R Stat Soc Ser A Stat Soc*. 2015.178(3):619-639.

Gile KJ, Handcock MS. Respondent-driven sampling: An assessment of current methodology. *Sociol Methodol*. 2010.40(1):285-327

Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *J R Stat Soc Ser A Stat Soc*. 2015.178(1):241-269.

Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J Am Stat Assoc*. 2011.106(493):135-146

Gillespie C. Fitting heavy tailed distributions: The powerLaw package. *Journal of Statistical Software*. 2015.64(2):1-16.

Gillis M. Ottawa's supervised injection site will have high-tech drug analyzing device. Ottawa Citizen. 2017. Available from: <https://ottawacitizen.com/news/local-news/pop-up-od-prevention-group-claims-life-saved-ahead-of-awareness-day-thursday>.

Goel S., Salganik M.J. Assessing respondent-driven sampling. *Proc Natl Acad Sci U S A*. 2010.107(15):6743-6747.

Goel S., Salganik M.J. Respondent-driven sampling as markov chain monte carlo. *Stat Med*. 2009.28(17):2202-2229.

Golub ET, Strathdee SA, Bailey SL, et al. Distributive syringe sharing among young adult injection drug users in five U.S. cities. *Drug Alcohol Depend*. 2007;91 Suppl 1:S30-8.

Grant K. Deal reduces price of life-saving hepatitis C drugs for Canadians. The Globe and Mail. 2017. Available from: <https://www.theglobeandmail.com/news/national/deal-reduces-price-of-life-saving-hepatitis-c-drugs-for-canadians/article34107225/>.

Grund JP, Kaplan CD, Adriaans NF. Needle sharing in the Netherlands: An ethnographic analysis. *Am J Public Health*. 1991;81(12):1602-1607.

Gyarmathy V.A., Johnston L.G., Caplinskiene I., Caplinskas S., Latkin C.A. A simulative comparison of respondent driven sampling with incentivized snowball sampling - the "strudel effect". *Drug Alcohol Depend*. 2014.135(1):71-77.

Halton Region. Types of drugs.
http://www.halton.ca/living_in_halton/public_health/health_wellness/drugs_alcohol_tobacco/ty

[pes of drugs/](#). Updated 2015. Accessed 01/15, 2017.

Heckathorn D.D., Semaan S., Broadhead R.S., Hughes J.J. Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18-25. *AIDS Behav.* 2002.6(1):55-67.

Heckathorn DD. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociol Methodol.* 2007.37(1):151-207.

Heckathorn DD. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Soc Probl.* 2002.49(1):11-34

Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Probl.* 1997.44(2):174-199.

Heckathorn DD. Snowball versus respondent-driven sampling. *Sociological methodology.* 2011.41(1):355-366.

Hein O, Schwind M, König,W. Scale-free networks: The Impact of fat tailed degree distribution on diffusion and communication processes. *WIRTSCHAFTSINFORMATIK.* 2006.48(4):267-275.

Humphries MD, Gurney K. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS ONE.* 2008.3(4):e0002051.

Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. *Clin Microbiol Infect.* 2013.19(11):999-1005.

Johnston L.G., Hakim A.J., Dittrich S., Burnett J., Kim E., White R.G. A systematic review of published respondent-driven sampling surveys collecting behavioral and biologic data. *AIDS Behav.* 2016.20(8):1754-1776.

Johnston LG, Chen YH, Silva-Santisteban A, Raymond HF. An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS Behav.* 2013.17(6):2202-2210.

Jolly AM, Muth SQ, Wylie JL, Potterat JJ. Sexual networks and sexually transmitted infections: A tale of two cities. *Journal of Urban Health.*2001.78(3):433-445.

Jones JH, Handcock MS. Sexual contacts and epidemic thresholds. *Nature.* 2003.423:605.

Katzir L, Hardiman S. Estimating clustering coefficients and size of social networks via random walk. *ACM Transactions on the Web.* 2015.9(4).

- Keeling MJ, Eames KTD. Networks and epidemic models. *Journal of the Royal Society Interface*. 2005;2(4):295-307.
- Kemeny JG, Snell JL. Finite markov chains. The University of Michigan: Van Nostrand; 1960.
- Klov Dahl AS. Social networks and the spread of infectious diseases: The AIDS example. *Soc Sci Med*. 1985;21(11):1203-1216.
- Kiss IZ, Green DM, Kao RR. The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. *Journal of the Royal Society Interface*. 2007;5(24):791-799.
- Koch AL. The logarithm in biology: II. distributions simulating the log-normal. *Journal of Theoretical Biology*. 1968;23(2):251-268.
- Kral AH, Malekinejad M, Vaudrey J, et al. Comparing respondent-driven sampling and targeted sampling methods of recruiting injection drug users in San Francisco. *J Urban Health*. 2010;87(5):839-850.
- Kuperman M, Abramson G. Complex structures in generalized small worlds. *Physical Review. Statistical, Nonlinear, and Soft Matter Physics*. 2001;64(4):471031-471034.
- Kupfer M. City-run injection site opens Tuesday. CBC News. 2017. Available from: <https://www.cbc.ca/news/Canada/ottawa/tour-ottawa-public-health-supervised-injection-site-1.4305663>.
- Kuttler AF, La Greca AM. Linkages among adolescent girls' romantic relationships, best friendships, and peer networks. *J Adolesc*. 2004;27(4):395-414.
- Lachowsky NJ, Sorge JT, Raymond HF, et al. Does size really matter? A sensitivity analysis of number of seeds in a respondent-driven sampling study of gay, bisexual and other men who have sex with men in Vancouver, Canada. *BMC Med Res Methodol*. 2016;16(1).
- Laupland KB, Embil JM. Reducing the adverse impact of injection drug use in Canada. *Can J Infect Dis Med Microbiol*. 2012;23(3):106-107.
- Li X, Chen G, Wang X. On spreading dynamics in discrete small-world networks. *IFAC*. 2005. (16):1107-1111.
- Luke DA, Harris JK. Network analysis in public health: History, methods, and applications. *Annu Rev Public Health*. 2007;28:69-93.
- MacArthur GJ, van Velzen E, Palmateer N, et al. Interventions to prevent HIV and hepatitis C in people who inject drugs: A review of reviews to assess evidence of effectiveness. *Int J Drug Policy*. 2014;25(1):34-52.

Mandell W, Vlahov D, Latkin C, Oziemkowska M, Cohn S. Correlates of needle sharing among injection drug users. *Am J Public Health*. 1994;84(6):920-923.

Martinez-Lopez B, Perez AM, Sanchez-Vizcaino JM. Social network analysis. review of general concepts and use in preventive veterinary medicine. *Transbound Emerg Dis*. 2009;56(4):109-120.

Neaigus A, Reilly KH, Jenness SM, Hagan H, Wendel T, Gelpi-Acosta C. Dual HIV risk: Receptive syringe sharing and unprotected sex among HIV-negative injection drug users in new york city. *AIDS Behav*. 2013.17(7):2501-2509.

Nesterko S, Blitzstein J. Bias–variance and breadth–depth tradeoffs in respondent-driven sampling. *J Stat Comput Simul*. 2015.85(1):89-102.

Newman L, Dale A. Large footprints in a small world: Toward a macroeconomics of scale. Sustainability: *Science, Practice, and Policy*. 2009.5(1).

Newman M,E.J. Power laws, Pareto distributions and Zipf's law.*Statistical Mechanics*.2004. Vol 46.

Newman ME, Moore C, Watts DJ. Mean-field solution of the small-world network model. *Phys Rev Lett*. 2000.84(14):3201-3204.

Newman ME, Watts DJ, Strogatz SH. Random graph models of social networks. *Proc Natl Acad Sci U S A*. 2002.99 Suppl 1:2566-2572.

Ott MQ, Gile KJ. Unequal edge inclusion probabilities in link-tracing network sampling with implications for respondent-driven sampling. *Electron J Stat*. 2016;10(1):1109-1132.

Pai M, McCulloch M, Gorman JD, et al. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. *Natl Med J India*. 2004.17(2):86-95.

Pasa MK, Alom KR, Bashri Z, Vermund SH. Sharing of needles and syringes among men who inject drugs: HIV risk in northwest Bangladesh. *PLoS One*. 2016;11(2):e0148276.

Patrick DM, Tyndall MW, Cornelisse PGA, et al. Incidence of hepatitis C virus infection among injection drug users during an outbreak of HIV infection. *Can Med Assoc J*. 2001;165(7):889.

Pelude L. Networks among injection drug users: Random or scale-free? [Masters]. University of Ottawa; 2007.

Pilon R, Leonard L, Kim J, et al. Transmission patterns of HIV and hepatitis C virus among networks of people who inject drugs. *PLoS One*. 2011.6(7):e22245.

Public Health England, Health Protection Scotland, Public Health Wales, and Public Health

Agency Northern Ireland. Shooting up: Infections among people who inject drugs in the UK 2012. 2013.

Public Health Ottawa. Problematic substance use in Ottawa: Technical report. 2016.

Reed WJ, Jorgensen M. The double pareto-lognormal Distribution—A new parametric model for size distributions. *Communications in Statistics - Theory and Methods*. 2004.33(8):1733-1753.

Reed WJ. A stochastic model for the spread of a sexually transmitted disease which results in a scale-free network. *Math Biosci*. 2006.201(1-2):3-14.

Remis RS. Modelling the incidence and prevalence of hepatitis C infection and its sequelae in Canada, 2007. Public Health Agency of Canada. 2010

Robertson J, Bucknall A, Welsby P, et al. Epidemic of AIDS related virus (HTLV-III/LAV) infection among intravenous drug abusers. *Br Med J (Clin Res Ed)*. 1986.292(6519):527-529.

Rocha LE, Liljeros F, Holme P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol*. 2011.7(3):e1001109.

Sacks-Davis R, Horyniak D, Grebely J, Hellard M. Behavioural interventions for preventing hepatitis C infection in people who inject drugs: A global systematic review. *Int J Drug Policy*. 2012.23(3):176-184.

Salganik M.J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Urban Health*. 2006.83(1).

Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent- driven sampling. *Sociol Methodol*. 2004.34(1):193-240.

Sanabria-Montana CH, Huerta-Quintanilla R. Generalization of clustering coefficient on lattice networks applied to criminal networks. *International Journal of Computer and Information Engineering*. 2017.11(7):868-872.

Saramaki J, Kaski K. Modelling development of epidemics with dynamic small-world networks. *Journal of theoretical biology*. 2005. pp.413-421.

Selvaraj V., Boopathi K., Paranjape R., Mehendale S. A single weighting approach to analyze respondent-driven sampling data. *Indian J Med Res*. 2016.144:447-459.

Shaghghi A, Bhopal RS, Sheikh A. Approaches to recruiting 'hard-to-reach' populations into re- search: A review of the literature. *Health promot perspect*. 2011.1(2):86-94.

Shaw SY, Shah L, Jolly AM, Wylie JL. Determinants of injection drug user (IDU) syringe sharing: The relationship between availability of syringes and risk network member

characteristics in Winnipeg, Canada. *Addiction*. 2007;102(10):1626-1635.

Slater N, Louzoun Y, Gragert L, Maiers M, Chatterjee A, Albrecht M. Power laws for heavy-tailed distributions: Modeling allele and haplotype diversity for the national marrow donor program. *PLoS Comput Biol*. 2015.11(4):e1004204.

Sohn I. Modeling and implementing small world network. 2016 International Conference on Information and Communication Technology Convergence (ICTC) . 2016. pp.414-416.

Sole RV, Valverde S. On evolution and architectural constraints. In: Information theory of complex networks. Springer-Verlag; 2004. pp.190-210.

Sornette D, Cont R. Convergent multiplicative processes repelled from zero: Power laws and truncated power laws. *Journal de Physique I*. 1997.7(3):431-444.

Sperandei S, Bastos LS, Ribeiro-Alves M, Bastos FI. Assessing respondent-driven sampling: A simulation study across different networks. *Soc Netw*. 2017.

Spiller MW, Gile KJ, Handcock MS, Mar CM, Wejnert C. Evaluating variance estimators for respondent-driven sampling. *Journal of Survey Statistics and Methodology*. 2018.6(1):23-45.

Srinivas MN. Social structure. *Sociological Bulletin*. 1964;13(1):12-21.

Strike C, Miskovic M. Scoping out the literature on mobile needle and syringe program review of service delivery and client characteristics, operation, utilization, referrals, and impact. *Harm Reduction Journal*. 2018.15:6.

Taylor S. 'A state of emergency': HIV rates in Sask. rise again in 2016. CBC News. 2017. Available from: <https://www.cbc.ca/news/Canada/saskatchewan/hiv-rates-rise-saskatchewan-1.4101145>.

Telesford QK, Joyce KE, Hayasaka S, Burdette JH, Laurienti PJ. The ubiquity of small-world networks. *Brain Connect*. 2011.1(5):367-75.

Thorpe LE, Ouellet LJ, Hershow R, et al. Risk of hepatitis C virus infection among young adult injection drug users who share injection equipment. *Am J Epidemiol*. 2002;155(7):645-653.

Tomas A, Gile KJ. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron J Stat*. 2011;5:899-934.

Tsai Y, Hsiao P-, Lin C-, Huang W-. Scale-free, small-world model for network communication. *Hindawi*. 2004.(2):903-908.

Uehara T, Yamasaki T, Okamoto T, et al. Efficiency of a "small-world" brain network depends

on consciousness level: A resting-state FMRI study. *Cereb Cortex*. 2014.24(6):1529-39.

UNAIDS. Prevention gap report. 2016.

Verdasca J, Telo da Gama,M.M., Nunes A, Bernardino NR, Pacheco JM, Gomes MC. Recurrent epidemics in small world networks. *J Theor Biol*. 2005.233(4):553-61.

Verdery A.M., Merli M.G., Moody J., Smith J.A., Fisher J.C. Brief report: Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in china. *Epidemiology*. 2015.26(5):661-665.

Verdery AM, Fisher JC, Siripong N, Abdesselam K, Bauldry S. New survey questions and estimators for network clustering with respondent-driven sampling data. *Sociological Methodology*. 2017.47(1):274-306.

Verdery AM, Mouw T, Bauldry S, Mucha PJ. Correction: Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE*. 2016.11(1):e0148006.

Virkar Y, Clauset A. Power-law distributions in binned empirical data. *Annals of Applied Statistics*. 2014.8(1):89-119.

Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*. 2008.24(1):79-97.

Wang XF, Chen G. Complex networks: Small-world, scale-free and beyond. *IEE Circuits and Systems Magazine*. 2001.

Wasserheit JN, Aral SO. The dynamic topology of sexually transmitted disease epidemics: Implications for prevention strategies. *J Infect Dis*. 1996.174 Suppl 2:S201-13.

Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*. 1998.393(6684):440-2.

Wejnert C. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol Methodol*. 2009.39(1):73-116.

Wiessing L, Ferri M, Grady B, et al. Hepatitis C virus infection epidemiology among people who inject drugs in Europe: A systematic review of data for scaling up treatment and prevention. *PLoS One*. 2014;9(7):e103345.

Wirtz A.L., Mehta S.H., Latkin C., et al. Comparison of respondent driven sampling estimators to determine HIV prevalence and population characteristics among men who have sex with men in Moscow, Russia. *PLoS ONE*. 2016.11(6):e0155519.

Wood E, Tyndall MW, Spittal PM, et al. Unsafe injection practices in a cohort of injection drug

users in Vancouver: Could safer injecting rooms help? *CMAJ*. 2001;165(4):405-410.

Woodward J. Safe opioid vending machine closer to reality: BCCDC. CTV News Vancouver. 2018. Available from: <https://bc.ctvnews.ca/safe-opioid-vending-machine-closer-to-reality-bccdc-1.3763945>.

Wylie JL, Jolly AM. Understanding recruitment: Outcomes associated with alternate methods for seed selection in respondent driven sampling. *BMC Med Res Methodol*. 2013.13:93-2288-13-93.

Yamanis TJ, Merli MG, Neely WW, et al. An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociol Methods Res*. 2013.42(3).

Young AM, Jonas AB, Mullins UL, Halgin DS, Havens JR. Network structure and the risk for HIV transmission among rural drug users. *AIDS Behav*. 2013.17(7):2341-2351.

Yu X-, Wang X-, Zhang D-, Liang F, Wu X. Mathematical expressions for epidemics and immunization in small-world networks. *Physica A: Statistical Mechanics and its Applications*. 2008.387(5-6):1421-1430.

Zaidi F, Sallaberry A, Melançon G. Generating artificial social networks with small world and scale free properties. [Research Report] INRIA; LIRMM. 2012.pp.34

Zhang Y, Zhang Z, Zhou S, Guan J. Deterministic weighted scale-free small-world networks. *Physica A: Statistical Mechanics and its Applications*. 2010;389(16):3316-3324.

Zimmermann J-. "Clusters" and "small-worlds":Aproximity issue. *Revue Economique*. 2002.53(3):517-5

10.0 Appendix

10.1 Search Strategy for Systematic Review

Database(s): Ovid **MEDLINE**(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily, Ovid MEDLINE and Versions(R)

Search Strategy:

#	Searches	Results
1	"respondent* driven sampl*".ti,kf.	314
2	"respondent* driven sampl*".ab. /freq=2	54
3	"respondent* driven sampl*".ab. and rds.ab. /freq=2	278
4	1 or 2 or 3	445
5	exp "Bias (Epidemiology)"/ or Evaluation Studies as Topic/ or Validation Studies as Topic/ or exp Reproducibility of Results/ or Statistics as Topic/ or Data Interpretation, Statistical/ or Confidence Intervals/ or Epidemiologic Research Design/ or Practice guidelines as topic/ or Epidemiologic Methods/ or Program Evaluation/	887028
6	(abilit* or accura* or analy* or assess* or assumption* or bias* or capture* or design* or effectiv* or efficac* or efficien* or error* or estimat* or evaluat* or guideline* or improv* or infer* or interpret* or limitation* or methodolog* or quality or reliab* or reproduc* or simulat* or valid* or variance*).ti,kf.	3287678
7	(abilit* or accura* or assess* or assumption* or bias* or capture* or design* or effectiv* or efficac* or efficien* or error* or estimat* or evaluat* or guideline* or improv* or infer* or interpret* or limitation* or methodolog* or quality or reliab* or reproduc* or simulat* or valid* or variance*).ab. /freq=2	5898205
8	5 or 6 or 7	7806149
9	4 and 8	324
10	limit 9 to (english or french)	321

Database(s): **Embase** 1997 to 2017 October

Search Strategy:

#	Searches	Results
1	"respondent* driven sampl*".ti,kw.	360
2	"respondent* driven sampl*".ab. /freq=2	64
3	"respondent* driven sampl*".ab. and rds.ab. /freq=2	297
4	1 or 2 or 3	493
5	exp statistical bias/ or exp evaluation study/ or validation study/ or exp measurement precision/ or exp measurement accuracy/ or confidence interval/ or variance/ or exp reliability/ or exp validity/ or statistics/ or statistical analysis/ or inferential statistics/ or practice guideline/ or exp quality control procedures/ or epidemiology/	3112226
6	(abilit* or accura* or analy* or assess* or assumption* or bias* or capture* or design* or effectiv* or efficac* or efficien* or error* or estimat* or evaluat* or	4016235

	guideline* or improv* or infer* or interpret* or limitation* or methodolog* or quality or reliab* or reproduc* or simulat* or valid* or variance*).ti,kw.	
7	(abilit* or accura* or assess* or assumption* or bias* or capture* or design* or effectiv* or efficac* or efficien* or error* or estimat* or evaluat* or guideline* or improv* or infer* or interpret* or limitation* or methodolog* or quality or reliab* or reproduc* or simulat* or valid* or variance*).ab. /freq=2	7362221
8	5 or 6 or 7	10589313
9	4 and 8	371
10	limit 9 to (english or french)	365

Scopus:

TITLE-ABS-KEY ("respondent* driven sampl*") AND (TITLE (abilit* OR accura* OR analy* OR assess* OR assumption* OR bias* OR capture* OR design* OR effectiv* OR efficac* OR efficien* OR error* OR estimat* OR evaluat* OR guideline* OR improv* OR infer* OR interpret* OR limitation* OR methodolog* OR quality OR reliab* OR reproduc* OR simulat* OR valid* OR variance*) OR KEY (abilit* OR accura* OR analy* OR assess* OR assumption* OR bias* OR capture* OR design* OR effectiv* OR efficac* OR efficien* OR error* OR estimat* OR evaluat* OR guideline* OR improv* OR infer* OR interpret* OR limitation* OR methodolog* OR quality OR reliab* OR reproduc* OR simulat* OR valid* OR variance*) OR ABS ((abilit* OR accura* OR assess* OR assumption* OR bias* OR capture* OR design* OR effectiv* OR efficac* OR efficien* OR error* OR estimat* OR evaluat* OR guideline* OR improv* OR infer* OR interpret* OR limitation* OR methodolog* OR quality OR reliab* OR reproduc* OR simulat* OR valid* OR variance*) W/10 ("respondent* driven sampl*" OR rds))) AND (LIMIT-TO (LANGUAGE , "English ") OR LIMIT-TO (LANGUAGE , " French "))

661 document results

Web of Science:

(TS=("respondent* driven sampl*") AND TS=(abilit* or accura* or assess* or assumption* or bias* or capture* or design* or effectiv* or efficac* or efficien* or error* or estimat* or evaluat* or guideline* or improv* or infer* or interpret* or limitation* or methodolog* or quality or reliab* or reproduc* or simulat* or valid* or variance*)) AND LANGUAGE: (English OR French)

Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

649 results