How to Analyze RDS Data?
—A Simple Guide

Kin Wang

CONTENT

Intro to RDSAnalyst

PART ONE

PART ONE: Intro to RDSAnalyst

What is it?

- Base: a R package named "RDS"(Cran)
- Built in Java for a user-friendly graphical interface software
- No need to write code by yourself
- Analyze RDS data for sample and population estimations, testing, CIs, sensitivity analysis

Installation

- http://wiki.stat.ucla.edu/hpmrg/index.php/RDS_Analyst_Install
- Windows and Mac

Authors

- Dr. Mark S. Handcock, Dr. Krista J. Gile , Dr. Ian E. Fellows

What you need in data?

PART TWO

PART TWO: What you need in data

- **File Format:**
- RDS Object: *.rdsobj, *.rdsat
- R Object: *.robj
- Others: *.csv, *.txt, *.sav, *.xpt, *bdf, *.dta, *.sys, *.syd, *.arff, *.rec, *.mtp, *.s3
- **Required Variables in Coupon Format Data:**
- Subject ID
- Seed Indicator: 0 is No, 1 is Yes
- Subject's Coupon ID
- Each Coupon ID Given to Subject to Recruit Others
- Network Size: each subject's self-reported degree (number of associations in such population)
- Variables of Interest: demographic info, such as age, race, job, HIV status, education

Logic: find the path of one subject connected to other

- **Required Variables in Recruiter ID Format Data:**
- Subject ID
- Seed Indicator: 0 is No, 1 is Yes
- Recruiter ID: subject ID of the recruiter for each respondent
- Network Size: each subject's self-reported degree (number of associations in such population)
- Variables of Interest: demographic info, such as age, race, job, HIV status, education

Several Estimators
PART THREE

PART THREE: Several estimators

RDS I

- S-H Estimator
- Based Markov chain assumption
- Equating the number of cross-relations between pairs of sub-populations of interest

RDS II

- V-H Estimator
- Estimating inclusion probabilities of sampled units
- Sampling process as random walk

HCG

- Homophily Configuration Graph Estimator
- Based on configuration graph network model
- Added homophily
- Without replacement sampling

We use estimator to estimate population from sample

RDS I (DS)

- Data smoothed version of RDS I
- Averaging over all pairs of groups (averaging degrees in group)

Gile's SS

- Successive Sampling Estimator
- Based on configuration graph network model
- Estimating inclusion probabilities of sampled units
- Without replacement sampling

How to choose?

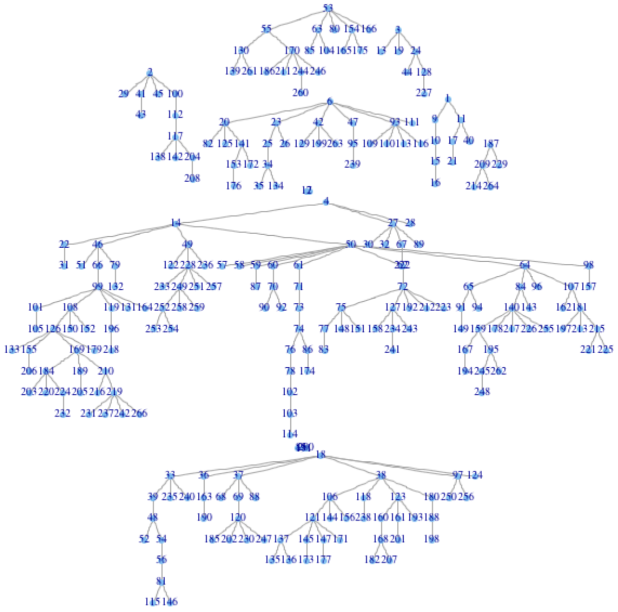
	RDS I	RDS I (DS)	RDS II	SS	HCG
Don't pop size	✅	✅	✅	❌	❌
Large sampling fraction	Biased	Biased	Biased	✅	✅
Don't know recruitment time	✅	✅	✅	✅	❌
Shorter waves	Biased	Biased	Biased	✅	✅
Higher homophily	Somewhat Biased	Somewhat Biased	Biased	Biased	✅
Biased seeds	Somewhat Biased	Somewhat Biased	Biased	Somewhat biased	✅
Highly differential activity	❌	❌	❌	✅	✅
Continuous variable	❌	❌	✅	✅	✅

Diagnosis

PART FOUR

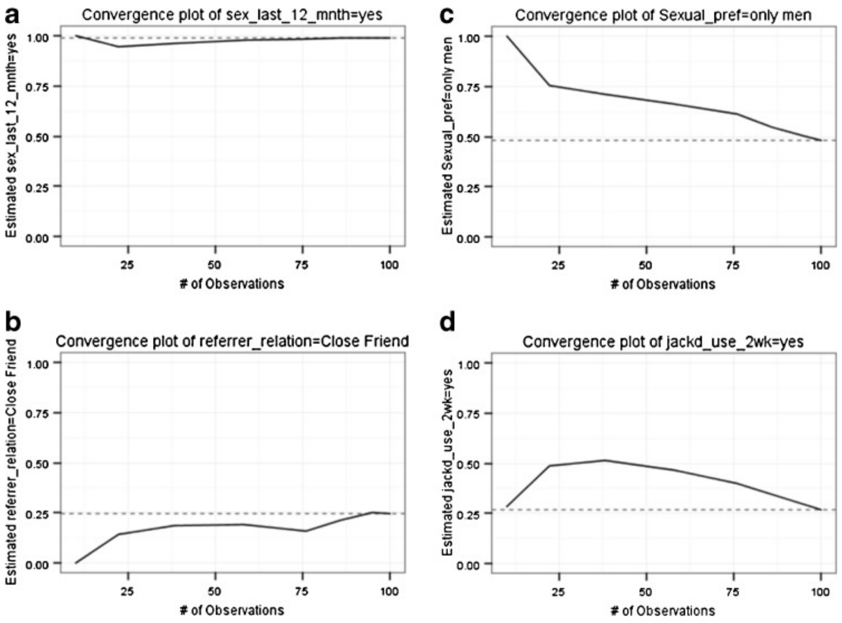
PART FOUR: Diagnosis

Recruitment Tree



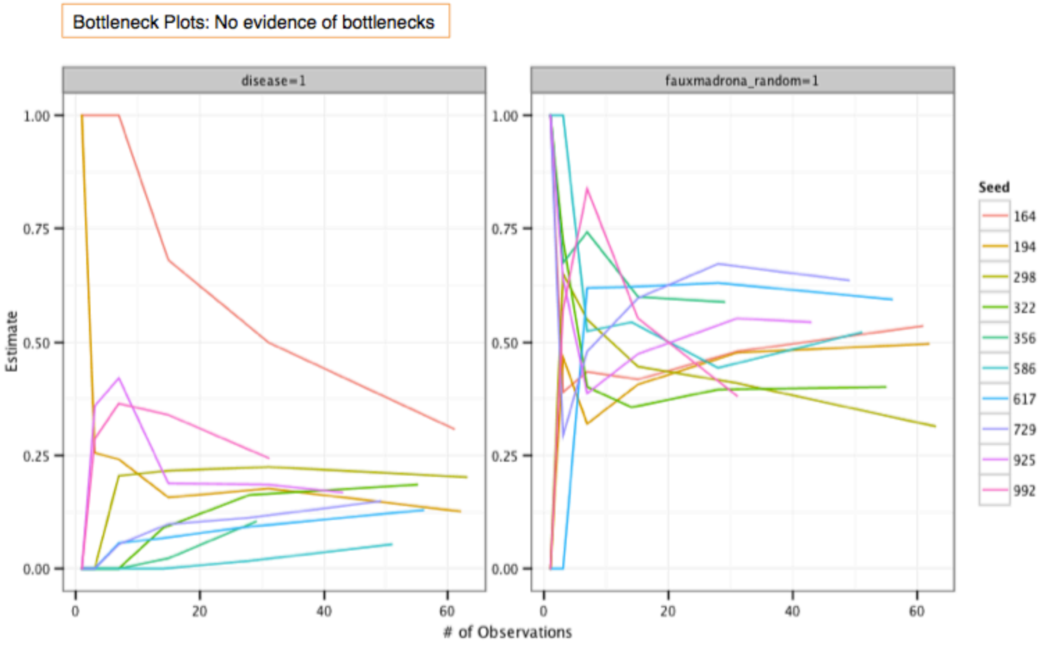
Seeds, waves, network

Convergence Plot



Whether results are converging to constant value and when how many samples

Bottleneck Plot



Whether the population of interest contains distinct sub-communities that could bias the RDS estimate

Homophily: the ratio of number of recruits that have the same disease status as their recruiter to the number we would expect by chance

Differential activity: compare the (weighted) average network size of two classes of the population

A Real Example

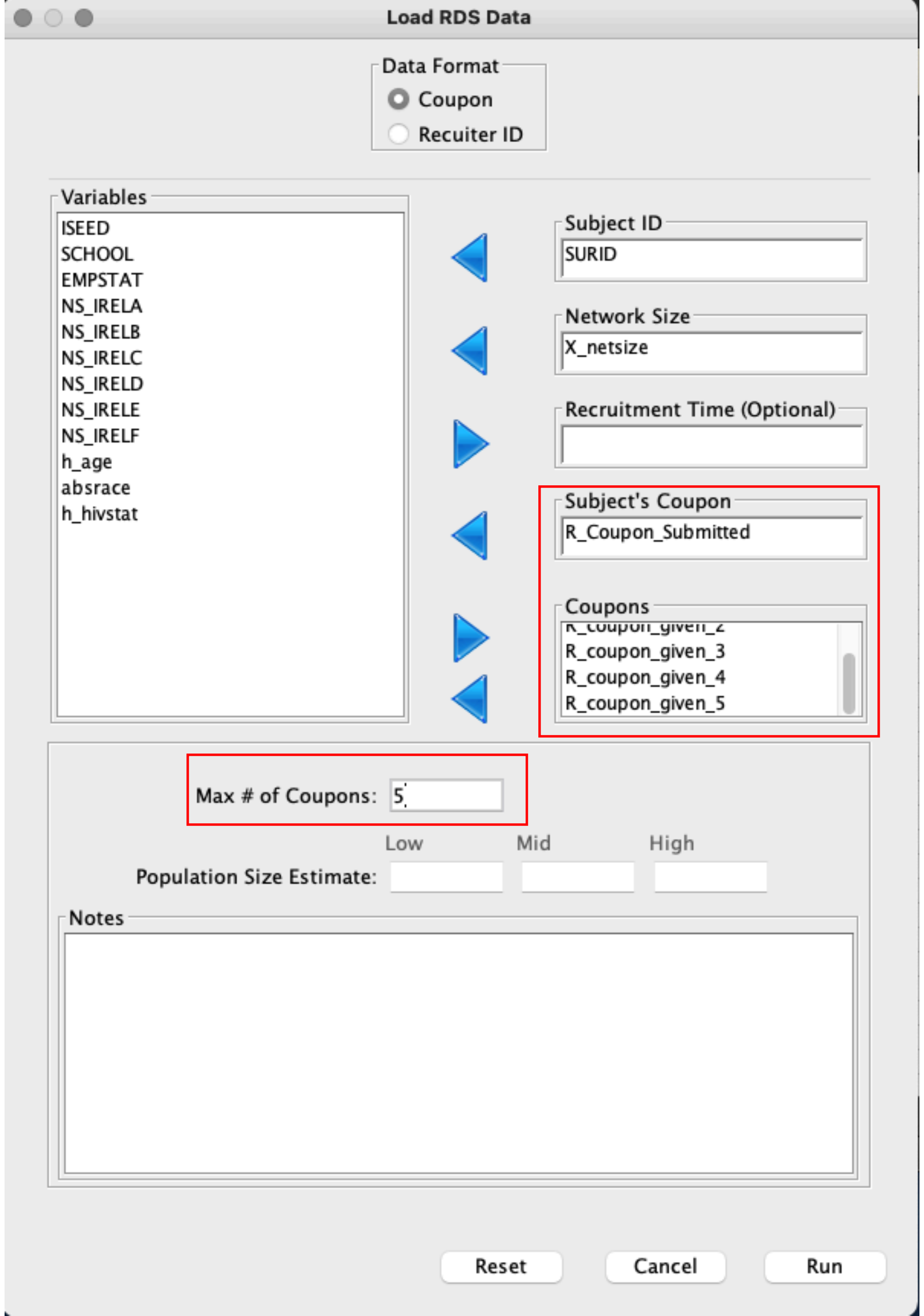
PART FIVE

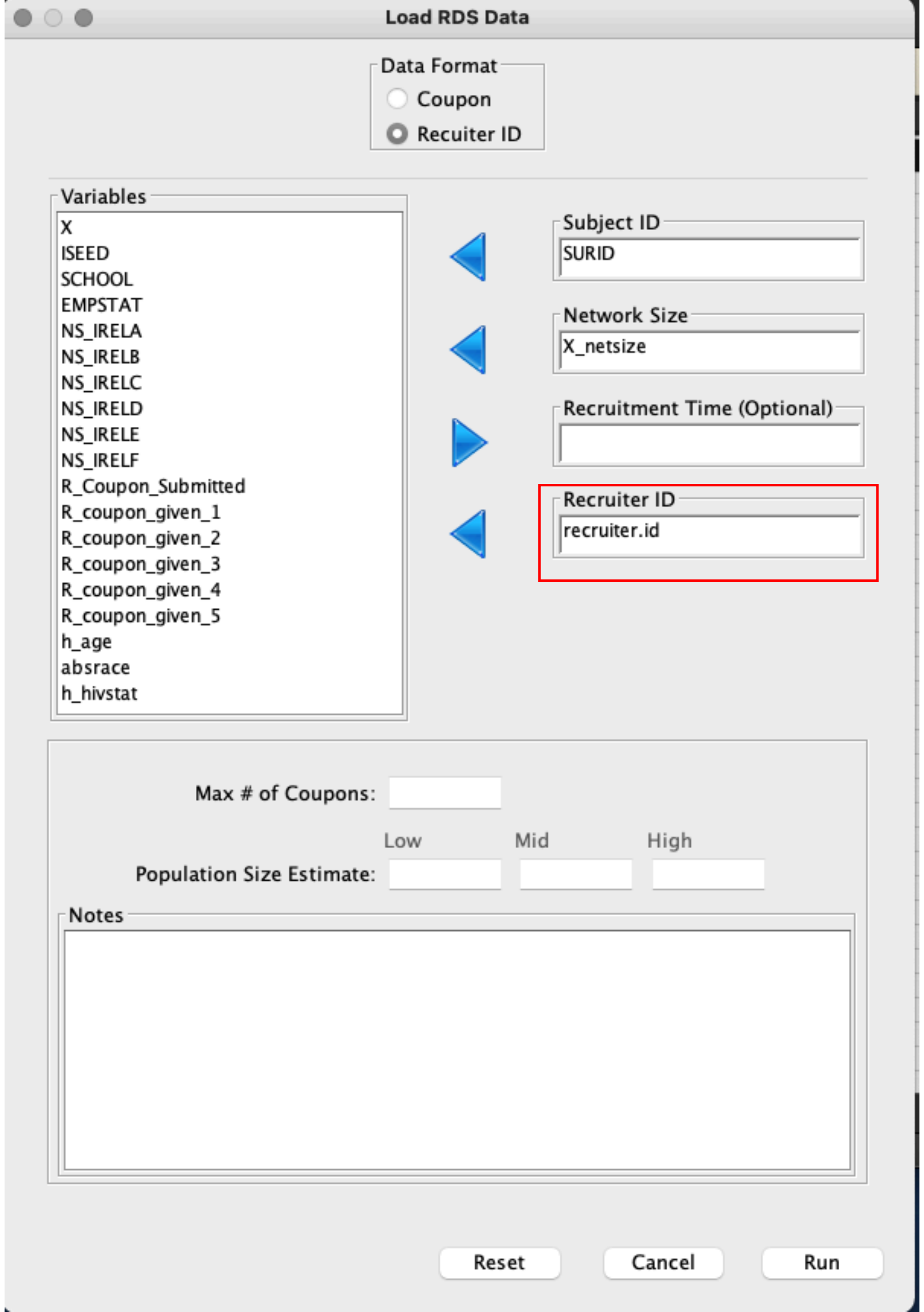
Check your data

Assume: population size is 5000;
Sample size: 524;
Sampling fraction is small;
No continuous variable;
Waves >10;
Lower homophily;

A	B	C	D	E	F	G	H	I	J	K	L	M	N
SURID	ISEED	SCHOOL	EMPSTAT	R_Coupon_Submitted	R_coupon_given_1	R_coupon_given_	R_coupon_giv	R_coupon_g	R_coupon_gi	_netsize	h_age	absrace	h_hivstat
1	1	3	7	-1	1001	1002	1003	1004	1005	17	3	5	3
2	1	1	7	-1	-1	-1	-1	-1	-1	47	2	7	3
3	1	2	8	-1	1017	1018	1019	1020	1021	40	1	5	3
4	1	2	7	-1	1022	1023	1024	1025	1026	60	1	7	3
5	1	3	6	-1	1027	1028	1029	1030	1031	110	2	7	3
6	1	3	7	-1	1517	1518	1519	1520	1521	25	3	5	3
7	1	3	7	-1	1526	1527	1528	1529	1530	23	2	5	3
8	1	3	7	-1	1533	1534	1535	1536	1537	50	3	5	3
9	1	3	1	-1	1538	1539	1540	1541	1542	7	3	5	3
10	1	3	2	-1	1601	1602	1603	1604	1605	15	3	5	3
11	1	4	7	-1	1853	1854	1855	1856	1857	7	3	5	3

0


Load RDS Data
Data Format
Coupon
Recuiter ID
Variables
ISEED
SCHOOL
EMPSTAT
NS_IRELA
NS_IRELB
NS_IRELC
NS_IRELD
NS_IRELE
NS_IRELF
h_age
absrace
h_hivstat
Subject ID
SURID
Network Size
X_netsize
Recruitment Time (Optional)
Subject's Coupon
R_Coupon_Submitted
Coupons
R_coupon_given_3
R_coupon_given_4
R_coupon_given_5
Max # of Coupons:
5
Low
Mid
High
Population Size Estimate:
Notes
Reset
Cancel
Run


Load RDS Data
Data Format
Coupon
Recuiter ID
Variables
X
ISEED
SCHOOL
EMPSTAT
NS_IRELA
NS_IRELB
NS_IRELC
NS_IRELD
NS_IRELE
NS_IRELF
R_Coupon_Submitted
R_coupon_given_1
R_coupon_given_2
R_coupon_given_3
R_coupon_given_4
R_coupon_given_5
h_age
absrace
h_hivstat
Subject ID
SURID
Network Size
X_netsize
Recruitment Time (Optional)
Recruiter ID
recruiter.id
Max # of Coupons:
Low
Mid
High
Population Size Estimate:
Notes
Reset
Cancel
Run

Deleted because of confidentiality

Reference

PART SIX

- Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol. 2004;34(1):193-240.
- Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. J Off Stat. 2008;24(1):79-97.
- Gile KJ, Handcock MS. Respondent-driven sampling: an assessment of current methodology. Sociol Methodol. 2010;40(1):285-327.
- Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. J Am Stat Assoc. 2011;106(493):135-146.
- Gile KJ, Handcock MS. Network model-assisted inference from respondent-driven sampling data. J Royal Stat Soc Ser A (Stat Soc). 2015;178(3):619-639.
- Fellows IE. Respondent-Driven Sampling and the Homophily Configuration Graph Statistics in Medicine. 2019;38:131–150.

THANK YOU FOR LISTENING