Assessing and Analyzing the Representativeness of Respondent-driven sampling (RDS):the HIV Transmission Network Metastudy Example

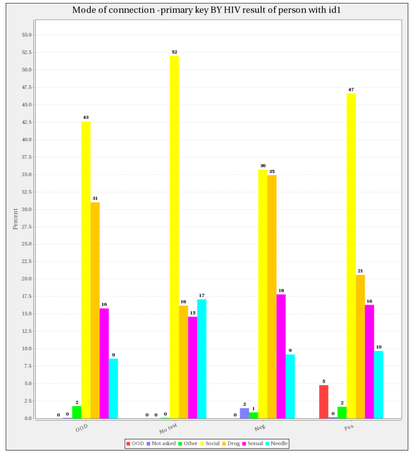
Chenxi Liu, Hao (Frank) Yang
Department of Civil and Environmental Engineering, University of Washington, Seattle WA.

Stat/CSSS 567, Social Network Spring 2019

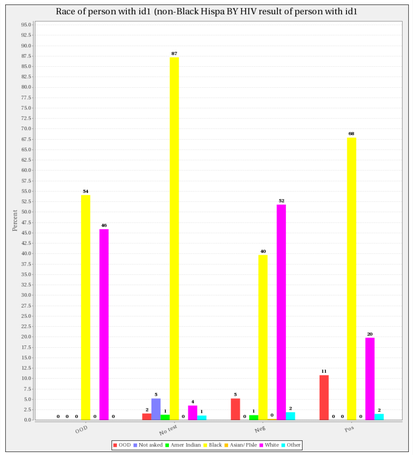
1. Introduction

- The **PURPOSES** of this study are:
  1) Illustrating and analyzing of RDS sample corresponding to the HIV transmission network representativeness, more specifically, about these three problems: seed selection impact, wave selection impact, and the features selection impact.
  2) Providing a details comparison and analysis and developing associated methods to address the impacts.

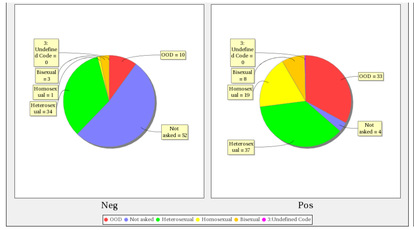
2. Metadata Overview



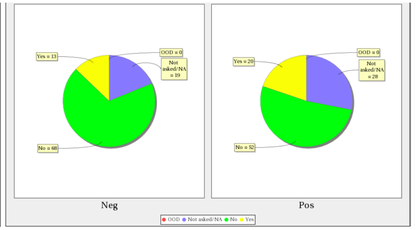
Ground truth data mode of connection related with the HIV results



Ground truth data: race related with the HIV results



Ground truth data sexual orientation related with the HIV results



Ground truth data homeless related with the HIV results

Regression Coefficients					Test That Each Coefficient = 0	
	B	SE(B)	Beta	SE(Beta)	T-statistic	Probability
sex1	.641	.009	.330	.005	72.781	.000
race1	.675	.006	.475	.004	107.843	.000
orient1	.343	.003	.336	.003	128.299	.000
streets1	-.126	.004	-.060	.003	-29.269	.000
educ1	-.220	.001	-.444	.003	-162.560	.000
Constant	-1.196	.028			-42.317	.000

Color coding: | <-0.5 | <-0.3 | <0.3 | <0.5 | T
Effect of each variable: | Negative | Positive |

Multiple R = .658 R-Squared = .433 Adjusted R-Squared = .433 SE of Estimate (Root MSE) = 3.108

Global Tests for Groups of Variables					
Group	Wald Chi-sq	Numerator	df Denominator	Adjusted Wald F	P
All independent variables	65,533.926	5	85818	13,106.174	.000

P = Probability that ALL B's in the group equal 0

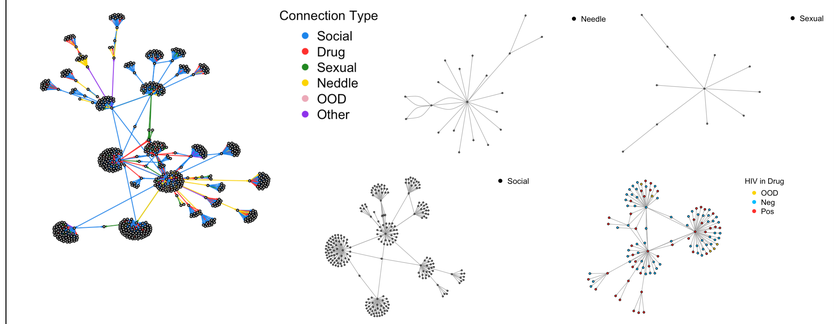
Allocation of cases	
Valid cases	85,823
Cases with invalid codes on variables in the analysis	67
Total cases	85,890

HIV results related with the race. Orient, homeless or not, education situation primary regression analysis

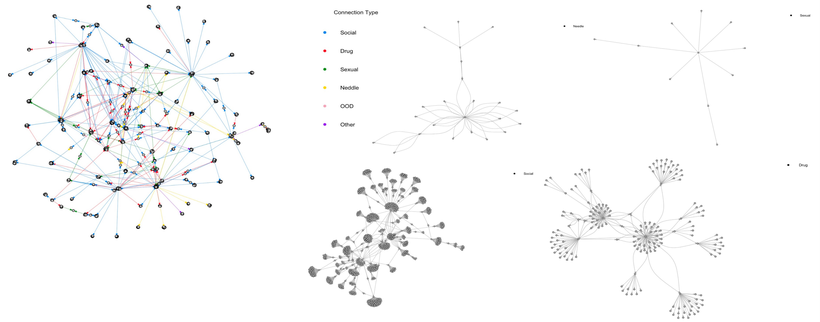
4. Methodology

1. We summarize the HIV transmission network growth features.
2. We use different seed selection, wave selection and feature selection to do the RDS sampling based on one data set.
3. We compare sampling results based on four aspects: degree, nodes betweenness, network growth and contribution to the HIV transmission.
4. We compare the sampling results with the ground truth data and summary the difference among different sampling strategies and the sampling representativeness.

4. Results (Wave impacts)

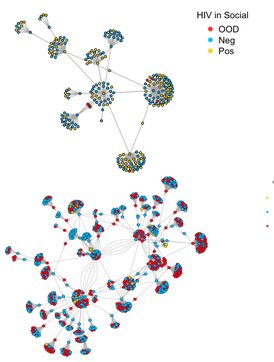


Here is the wave=3 RDS sampling graph based on different tie types. We can see the different kinds of connection types contribute to the network growth situation is different.

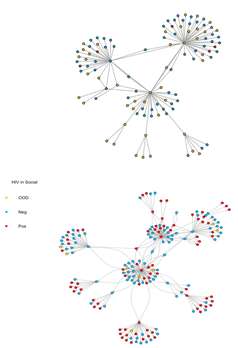


Here is the wave=4 RDS sampling graph based on different tie types. The main contributor of the network growth is social and drugs.

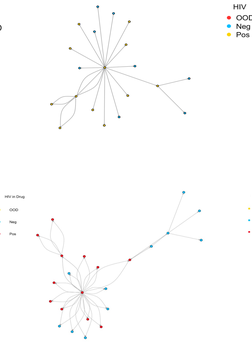
4. Results (Transmission impacts)



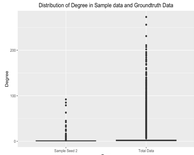
HIV growth graph of Social connections

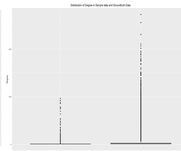


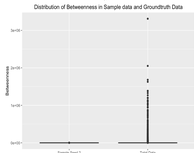
HIV growth graph of drug connections

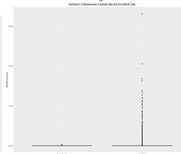


HIV growth graph of needle connections

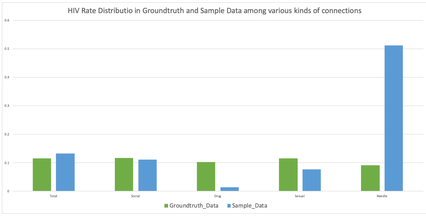








	Total	Social	Drug	Sexual	Needle
Ground truth Data	0.1146	0.116092	0.101456	0.115174	0.091051
Sample Data	0.132122	0.11051	0.013699	0.076471	0.512111



The summaries of representativeness features of RDS sampling comparing with the ground truth data. From the total HIV rate we can find that the RDS sampling can showing the over all HIV distribution. However, for each of the connection types, RDS sampling still shows bias, and some of them even with pretty high bias.

5. Conclusions

1. We Illustrating and analyzing of RDS sample corresponding to the HIV transmission network representativeness, find that the RDS sampling can show the overall distribution of the HIV transmission network.
2. RDS sampling still exists bias on some features. The biases introduces and expands mainly from three parts, seed selection, wave growth and features ununiform distribution.
3. The bias of the RDS sampling could be control and reduce through these process.