New Approaches to Collecting Data From a Respondent-Driven Sample

Session: Computational Demography, Machine Learning, and Algorithmic Bias

Casey F. Breen[1] Dennis M. Feehan[1]

[1]University of California, Berkeley

April 7, 2022

Respondent-Driven Sampling (RDS)

- Leading method for sampling hidden populations

Respondent-Driven Sampling (RDS)

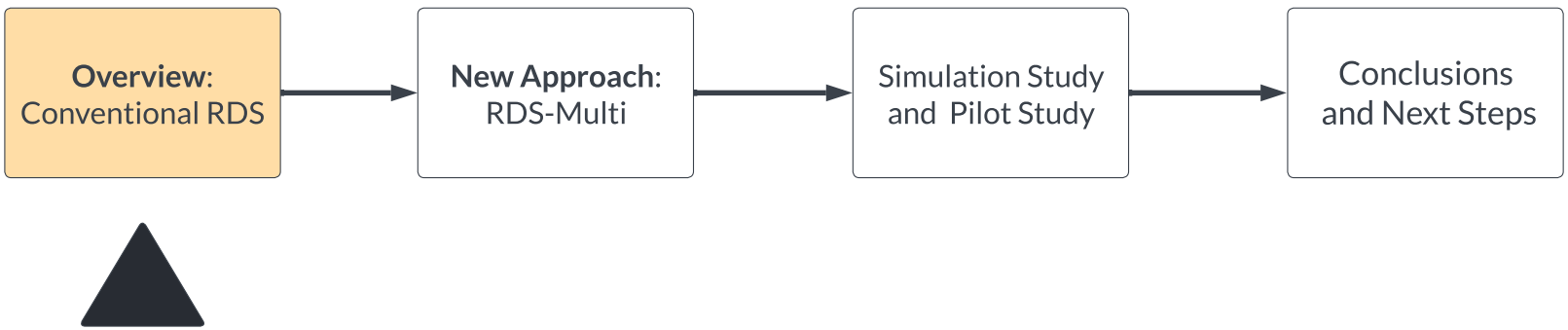
- Leading method for sampling hidden populations
- **Hidden populations:** populations that are hard-to-reach, often due to engaging in stigmatized or illegal behavior (persons who inject drug, commercial sex workers, etc.)

Respondent-Driven Sampling (RDS)

- Leading method for sampling hidden populations
- **Hidden populations:** populations that are hard-to-reach, often due to engaging in stigmatized or illegal behavior (persons who inject drug, commercial sex workers, etc.)
- **RDS Key insight:** Members of a hidden population are often socially connected to each other – and can recruit each other to be interviewed

Presentation Roadmap

- **Goal**: Introduce RDS-Multi, a new approach to collecting data from a respondent-driven sample



Respondent-Driven Sampling – Overview

1. Typical RDS study begins with 3-10 seeds, people known to be in the hidden population (e.g., people who inject drugs)

Respondent-Driven Sampling – Overview

1. Typical RDS study begins with 3-10 seeds, people known to be in the hidden population (e.g., people who inject drugs)

2. Seeds recruit other members of the hidden population to be interviewed

Respondent-Driven Sampling – Overview

1. Typical RDS study begins with 3-10 seeds, people known to be in the hidden population (e.g., people who inject drugs)
2. Seeds recruit other members of the hidden population to be interviewed
3. After being interviewed, respondents recruit next wave of respondents

RDS Recruitment Trees

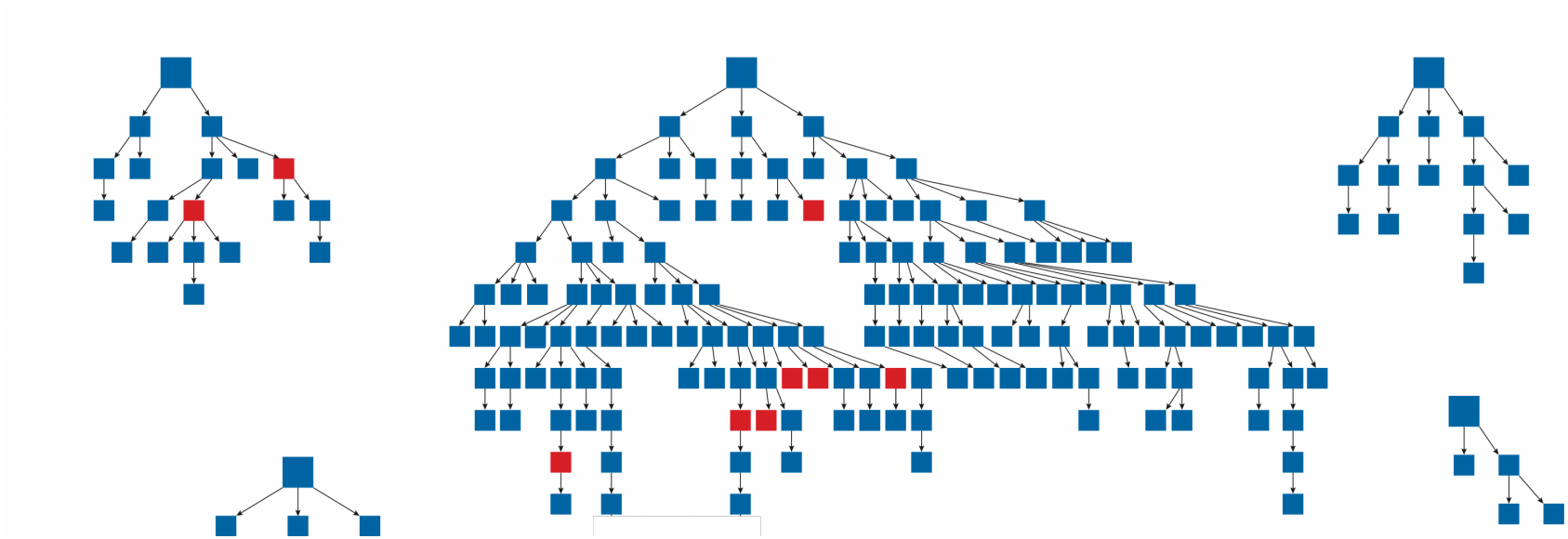


Figure: Recruitment tree plots from Gile et al. (2015)

When Conventional RDS Doesn't Work Well . . .

- **Low Connectivity:** Members of a hidden population don't know other members of hidden population to recruit
- **High Clustering**: Bottlenecks due to extreme homophily make it difficult for RDS to fully traverse network

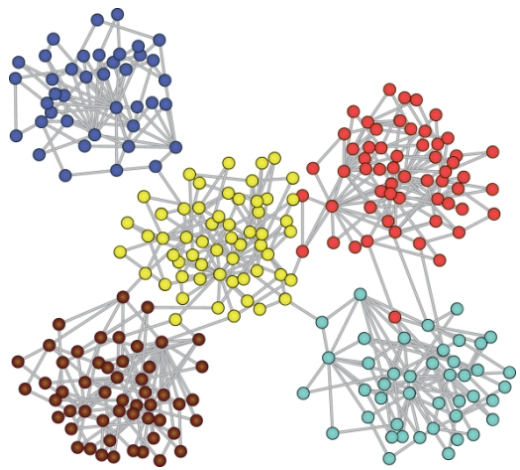


Figure: Clustered Social Network

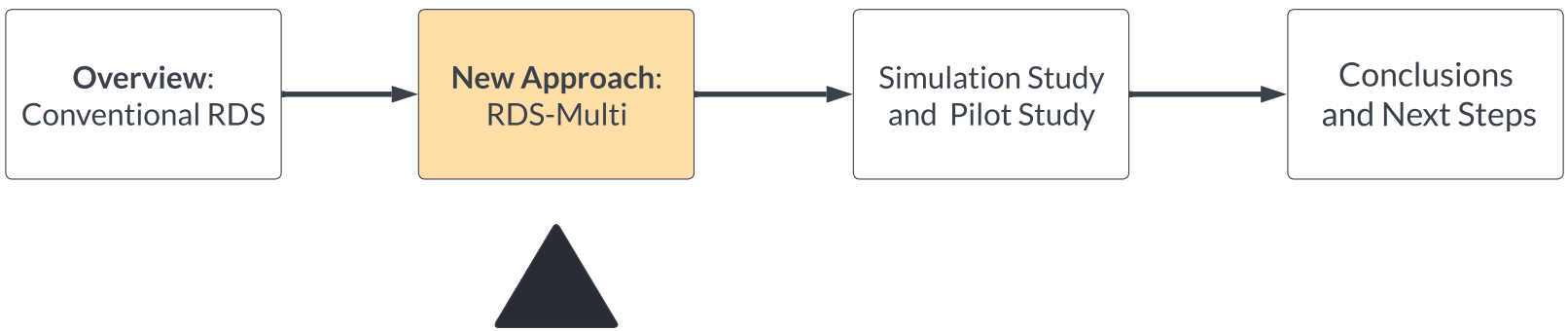
What Do We Do When RDS Doesn't Work Well?

- Improve statistical methods for analyzing RDS data

What Do We Do When RDS Doesn't Work Well?

- Improve statistical methods for analyzing RDS data
- **Change data collection procedure to give more favorable underlying network structure**

RDS-Multi: Roadmap



Motivating Example

- RDS Study: What is the proportion of people experiencing homelessness in the San Francisco Bay Area are fully vaccinated for COVID-19?

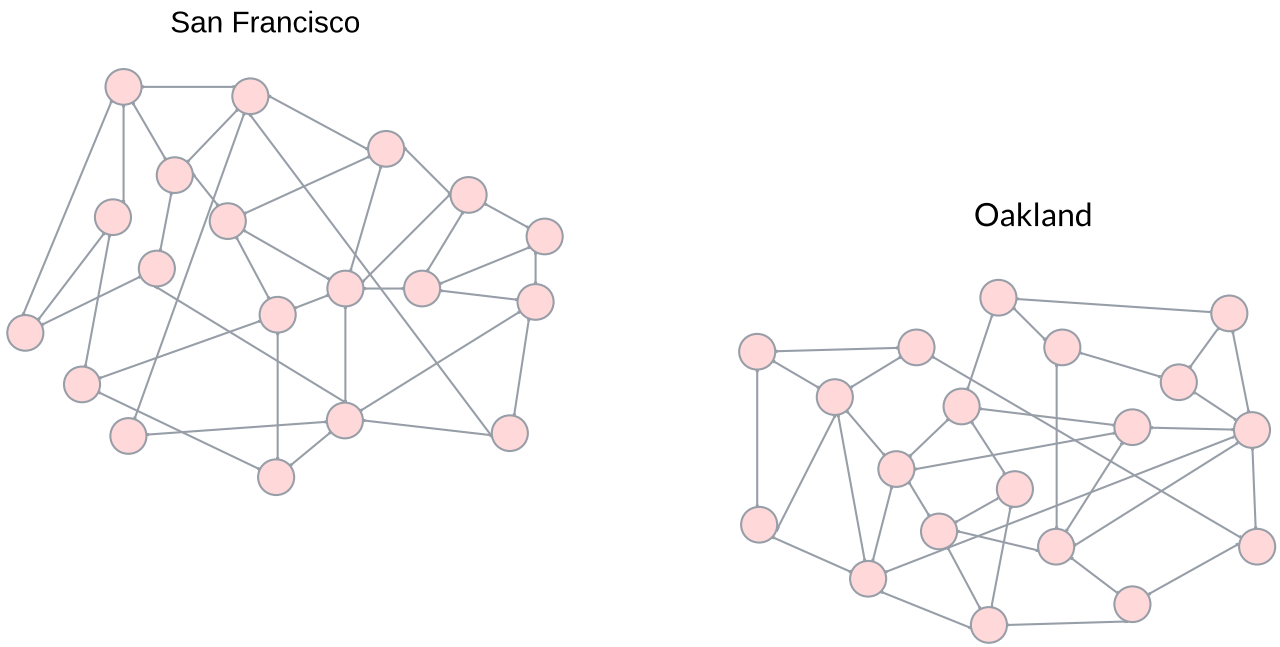


Figure: Bottlenecks between San Francisco and Oakland

New Approach: RDS-Multi

- New referral method: hidden population members refer other hidden population members or **social referents**, people highly connected to – but not in – the hidden population.
- For example:
  - **Hidden population:** People experiencing homelessness in the Bay area
  - **Social Referents**: Social workers specializing in homeless outreach services

Increase network connectivity

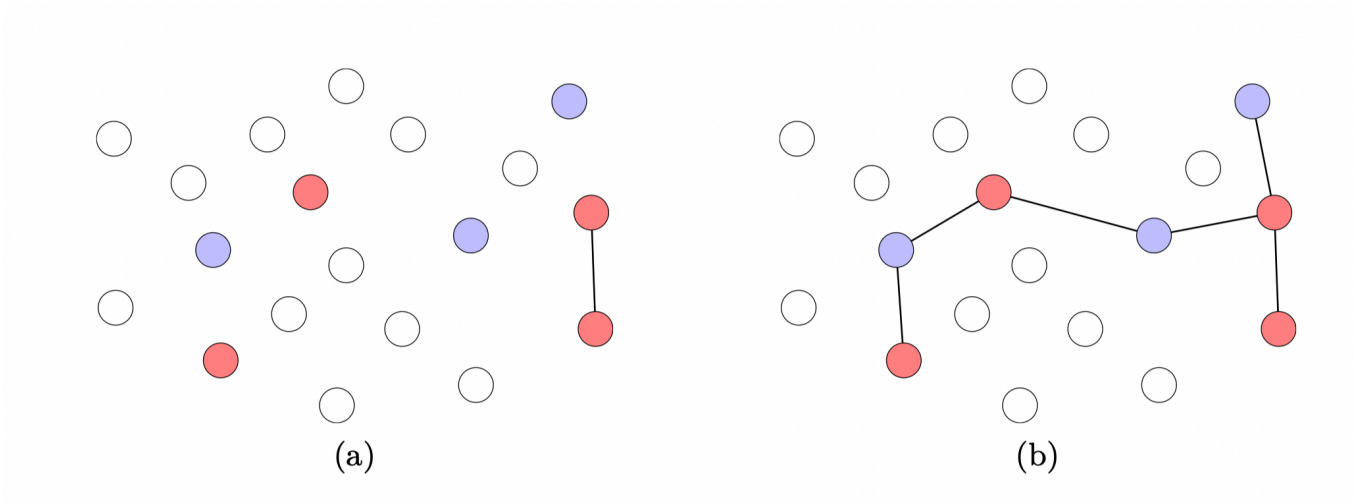
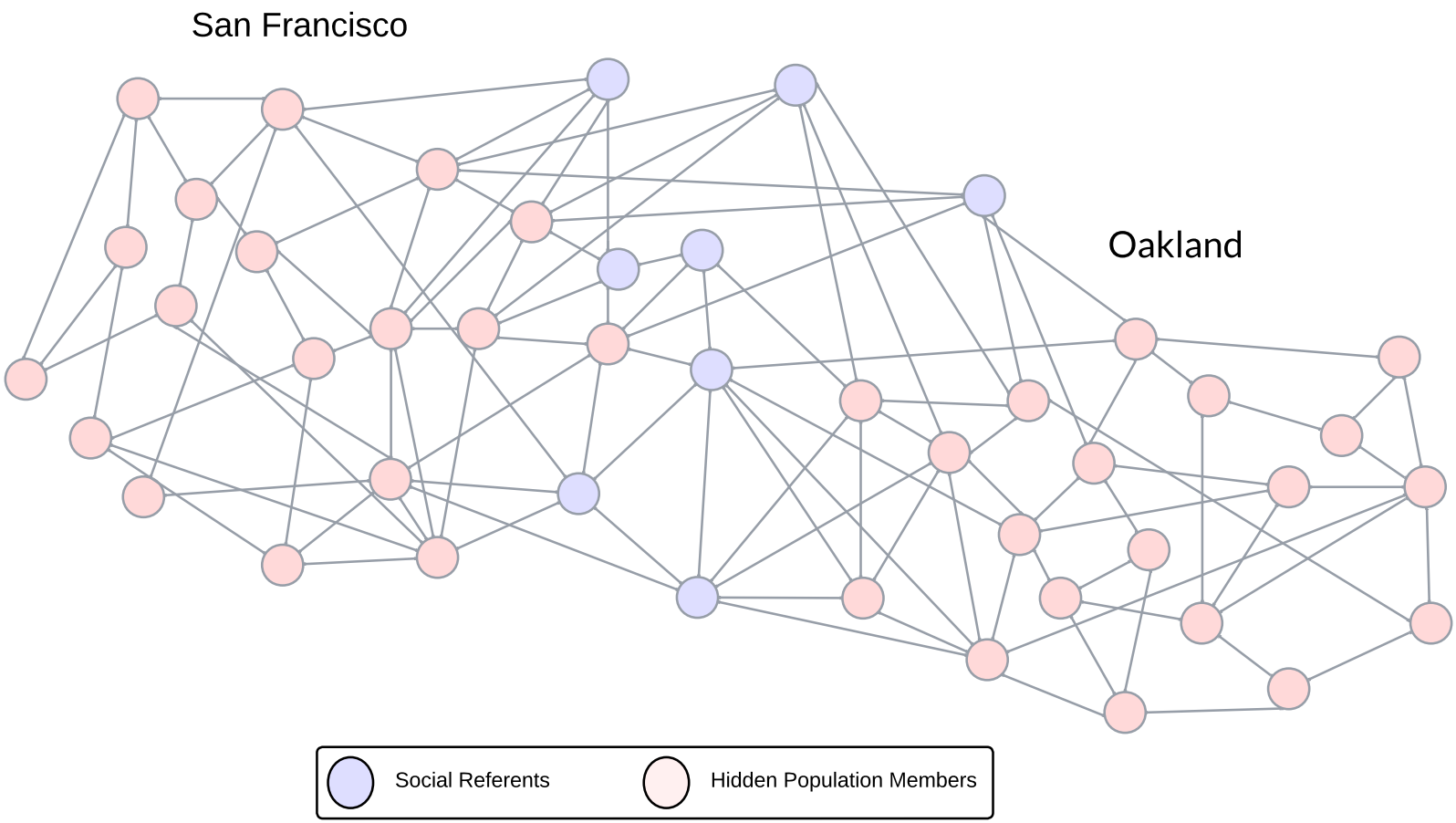


Figure: New referral method can improve underlying network structure

Decrease clustering and bottlenecks



Conventional RDS: Volz & Heckathorn Point Estimator

- **Core insight:** Not all people have same probability of being recruited into sample

Conventional RDS: Volz & Heckathorn Point Estimator

- **Core insight:** Not all people have same probability of being recruited into sample
- Inclusion probability is proportional to degree

Conventional RDS: Volz & Heckathorn Point Estimator

- **Core insight:** Not all people have same probability of being recruited into sample
- Inclusion probability is proportional to degree

$$\mu_{VH} = \frac{\sum_{i=1}^{n} \frac{z_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}}$$

where d_i is respondent i's degree and z_i is a binary covariate

RDS-Multi: Adapted Volz & Heckathorn Point Estimator

Adapt Volz-Heckathorn estimator to account for the new referral pattern:

$$\widehat{\mu}'_{VH} = \frac{\sum_{i \in s' \cap H} \frac{q_i}{d_{i,R}}}{\sum_{i \in s' \cap H} \frac{1}{d_{i,R}}},$$

where

- z_i is a binary covariate
- $s' \cap H$ is the subset of the sample that consists of hidden population members;
- and $d_{i,R}$ is the number of connections between i and the set R of social referents.

Variance Estimator

Estimating uncertainty in respondent-driven sampling using a tree bootstrap method

Aaron J. Baraff, Tyler H. McCormick, and Adrian E. Raftery

+ See all authors and affiliations

PNAS December 20, 2016 113 (51) 14668-14673; first published December 7, 2016;
https://doi.org/10.1073/pnas.1617258113

Contributed by Adrian E. Raftery, October 27, 2016 (sent for review November 24, 2015; reviewed by Sharad Goel and Matthew J. Salganik)

Variance Estimator – Tree Bootstrap

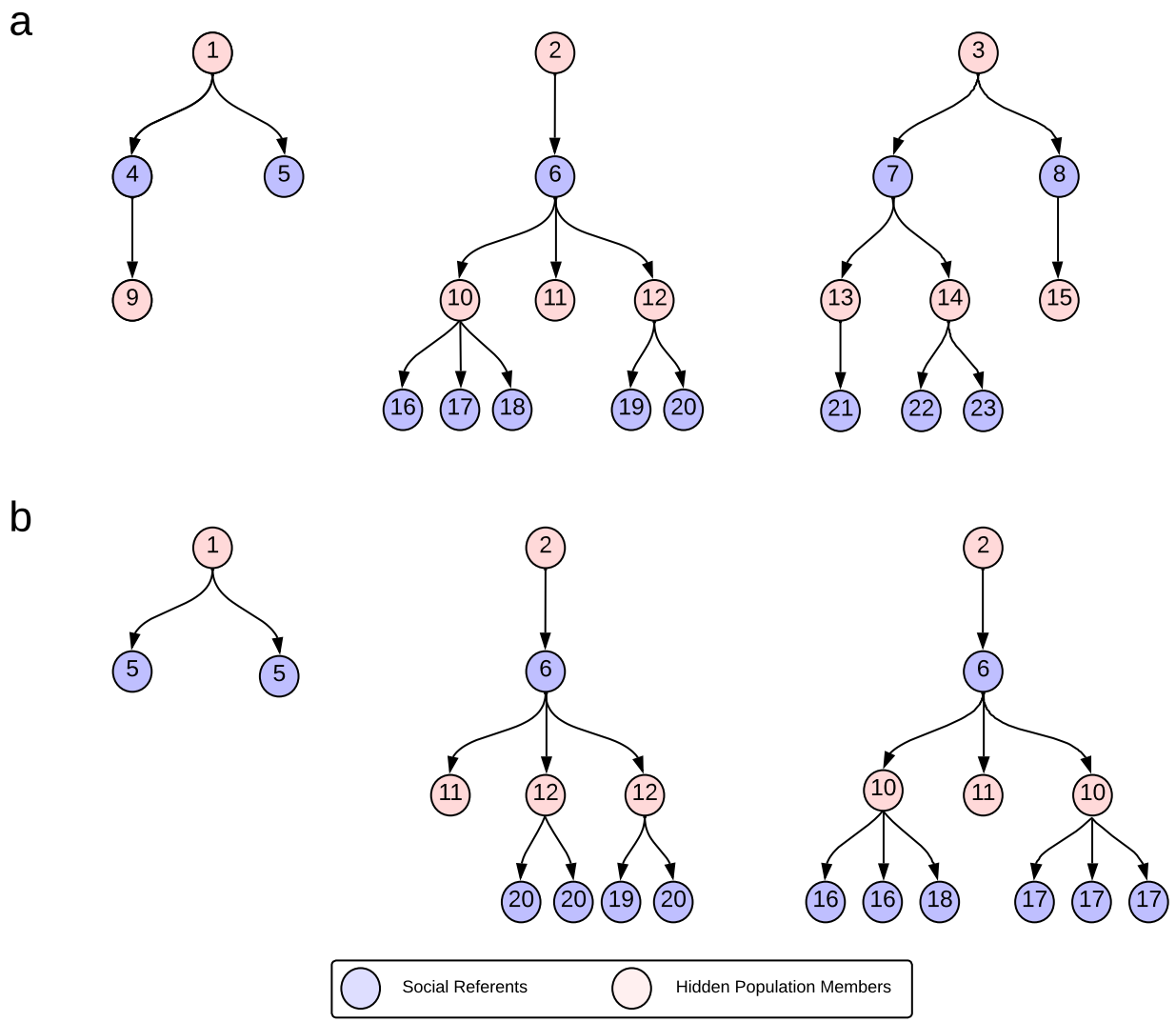


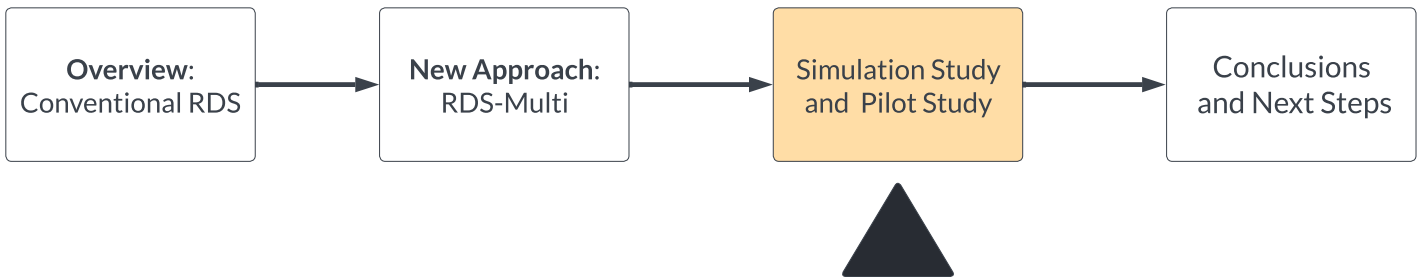
Figure: Adapted Tree Bootstrap Estimator (Baraff, McCormick and Raftery, 2016)

RDS-Multi: Key Considerations

1. Population of interest is the hidden population – we'll drop all social referents, resulting in a smaller effective sample size (or will need to conduct more interviews)

2. We need a sufficiently large and well-connected set of social referent nodes

RDS–Multi: Roadmap



Simulation Study

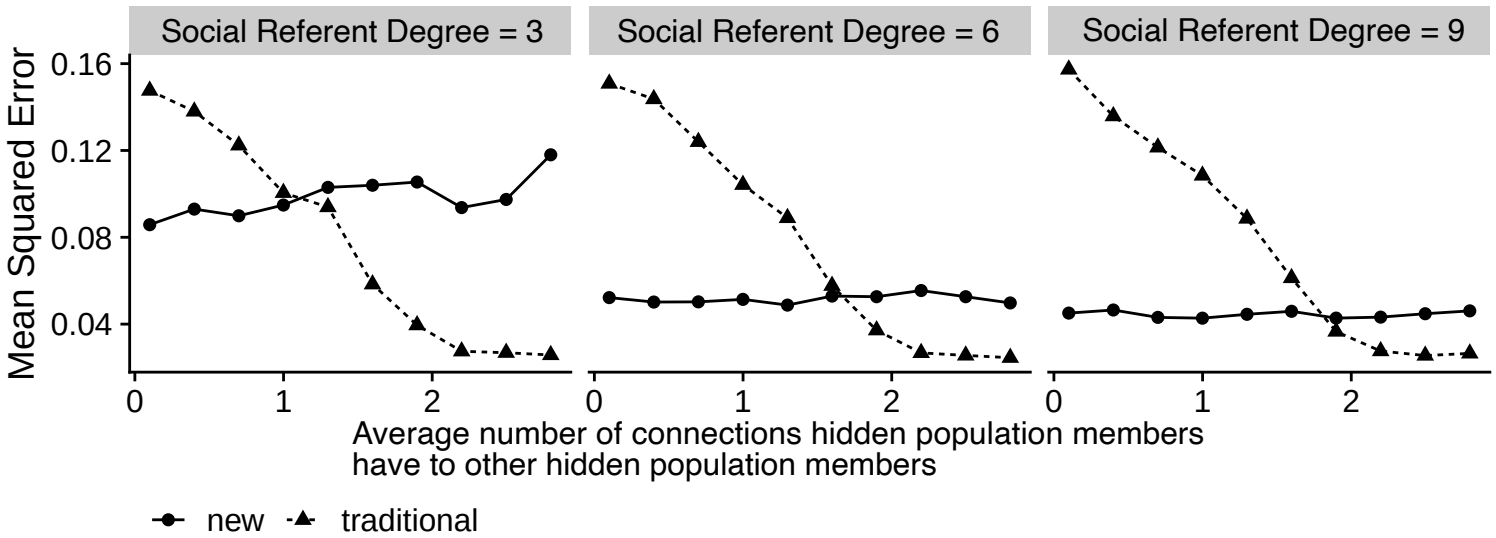


Figure: When connectivity is low, RDS-Multi performs better than the conventional RDS

Note: Sample sizes = 500, including social referents

Pilot Study in Kaya, Burkina Faso

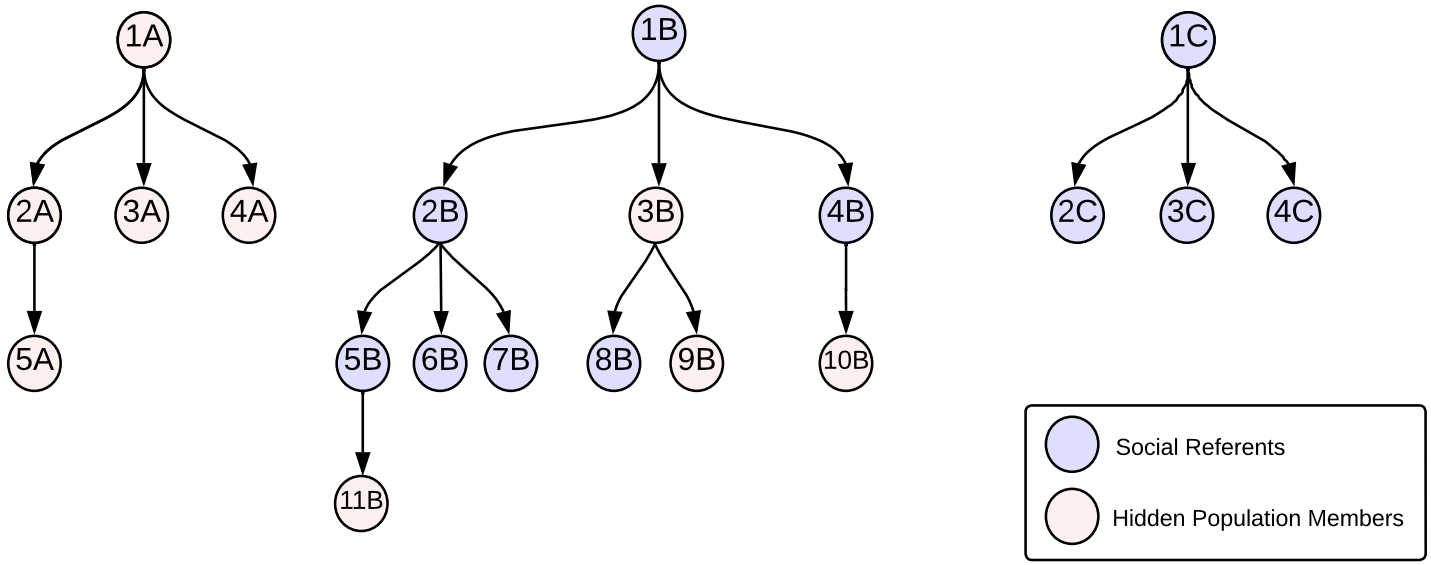
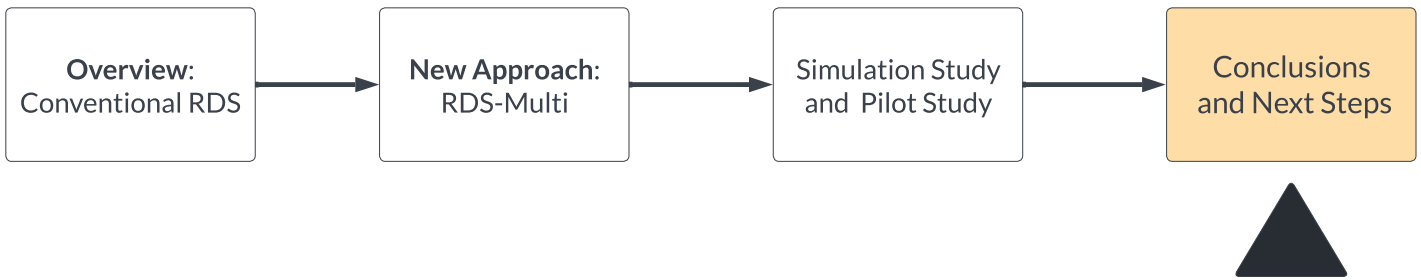


Figure: RDS-Multi recruitment trees from pilot study[1]

[1] Zan, Owolabi, Baguiya, Oduor, Bangha, Kim and Rossier (2022)

RDS–Multi: Roadmap



Conclusion

- RDS-Multi is a novel approach to collecting RDS data using **social referents**
- **Key advantages:**
  - Enables RDS for weakly connected hidden populations
  - Enables RDS for highly clustered networks
- **Key consideration:** RDS-Multi requires the availability of a sufficiently large and well-connected set of **social referents**

Next Steps

- More formal mathematical and empirical understanding of the trade-offs between RDS and RDS-Multi
- More empirical evidence from real world RDS-Multi studies

Thank You

- Questions?

caseyfbreen
caseybreen@berkeley.edu

References

Baraff, Aaron J., Tyler H. McCormick and Adrian E. Raftery. 2016. "Estimating Uncertainty in Respondent-Driven Sampling Using a Tree Bootstrap Method." *Proceedings of the National Academy of Sciences* 113(51):14668–14673.

Zan, Moussa L., Onikepe Owolabi, Adama Baguiya, Clement Oduor, Martin Bangha, Caron Kim and Clémentine Rossier. 2022. "Using Respondent Driven Sampling to Measure Abortion Safety in Restrictive Contexts: Results from Kaya (Burkina Faso) and Nairobi (Kenya).".