# Improved Inference for RDS Data: Inspirations and Implications for Data Collecting Process

Lingyu (Jack) Fuca

Note: All the questions are colored as red.

# Why Do We Trace Links to Collect Data?”

- What does the data generating process look like…?
- “Observability”?
- What is the relationship between the true population and the observable sample?
- What is a “probability sample?”

# RDS: Respondent-Driven Sampling

- A sampling design…
  - Respondents at each wave to select the next wave…
  - Exploits the network of social relations…
- A correspondent approach to estimation…
  - Why does adding more sampling waves make the final sample less dependent on the starting seeds?

- RDS works as… Initial conditions only matter when the network/system is small, fragile, or poorly connected. But once the network/system reaches a critical size, connectivity, or self-reinforcement process, it starts "forgetting" its starting point/status. The structure becomes more important as the *observed* network grows…

# RDS Working in the Real World…

- If we were designing an ideal case for RDS, what would the network or recruitment data look like?
  - Strong contagion (via recruitment)… what is recruitment in various domains…?
  - Network structure dominates over initial seeds/status… (What does structure mean in different real-world settings?)

- In summary… An ideal network for RDS application should be…
  - Dense (not sparse) internal network (why?)
  - Clear boundary (why?)
  - Well-connected structure (short paths) (why?)
  - Moderate degree heterogeneity (not too flat, not too extreme) (why?)
  - Temporal stability during recruitment (why?)

# Gile's Improvement to the RDS Framework

Existing RDS framework:

- Inclusion probability ∝ each node's degree (Your chance of being in the sample is proportional to your degree )
- Assumes sampling with replacement
- Ignores how network structure changes during sampling

Gile's Successive Sampling (Structure-Aware View)

- Models sampling without replacement
- Recruitment changes local network structure
- Probability of inclusion = Degree + Sample Size + Remaining Degree Distribution

Recruiting someone does not only affect them. It indirectly affects their neighbors by limiting their recruitment pathways.

# Example: How Recruitment Changes the Network

Imagine this observed partial network A — B — C — D — E

- The traditional RDS:
  - B, C, and D have the equal probability of being recruited because they have the same degree (local maxima = 2).
  - It doesn’t matter where they are located in the network.
  - It doesn’t care about local recruitment dynamics.

- Gile’s framework:
  - Start with B and B recruits A first…
  - Then B only has C left to recruit.
  - D is still far away, because D depends on whether C recruits them.
  - C has a higher probability of being recruited soon (because they are directly connected to B); but D’s chance depends on multiple recruitment steps happening first.

# Back to Data Collection…

- If we were designing an ideal case for Gile's approach, what would the network or recruitment data look like?
  - Good data for Gile's method means respecting the network: Dense inside, closed outside, dynamic, (moderately) heterogenous, but stable enough (during the recruitment) to let structure emerge.

- Inspirations for data collecting process…?
  - Recruitment follows true social ties (not random, not forced).
  - Recruitment chains are long enough (many waves) to reduce (initial) seed dependence.
  - Accurate reporting of degree (how many people you know in the target population).
  - Accurate recording of who recruited whom (so the structure can be reconstructed).
  - Key: Let the network structure speak for itself through data.