Sampling and incomplete network data

567 Statistical analysis of social networks

Peter Hoff

Statistics, University of Washington

Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).

Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).

Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).

Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).

Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).

Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).

Common sampling methods

1. node-induced subgraph sampling
2. edge-induced subgraph sampling
3. egocentric sampling
4. link tracing designs
5. censored nomination schemes

Node-induced subgraph sampling

Procedure:

1. Uniformly sample a set $\mathbf{s} = \{s_1, \ldots, s_{n_s}\}$ of nodes

$$\mathbf{s} \subset \{1, \ldots, n\}.$$

2. Observe relations $\mathbf{y}_\mathbf{s}$ between sampled nodes

$$\mathbf{Y}_\mathbf{s} = \{y_{i,j} : i \in \mathbf{s}, j \in \mathbf{s}\}.$$

Node-induced subgraph sampling

Procedure:

1. Uniformly sample a set $\mathbf{s} = \{s_1, \ldots, s_{n_s}\}$ of nodes

$$\mathbf{s} \subset \{1, \ldots, n\}.$$

2. Observe relations $\mathbf{y_s}$ between sampled nodes

$$\mathbf{Y_s} = \{y_{i,j} : i \in \mathbf{s}, j \in \mathbf{s}\}.$$

Node-induced subgraph sampling

Procedure:

1. Uniformly sample a set $\mathbf{s} = \{s_1, \ldots, s_{n_s}\}$ of nodes

$$\mathbf{s} \subset \{1, \ldots, n\}.$$

2. Observe relations $\mathbf{y_s}$ between sampled nodes

$$\mathbf{Y_s} = \{y_{i,j} : i \in \mathbf{s}, j \in \mathbf{s}\}.$$

Node-induced subgraph sampling

Procedure:

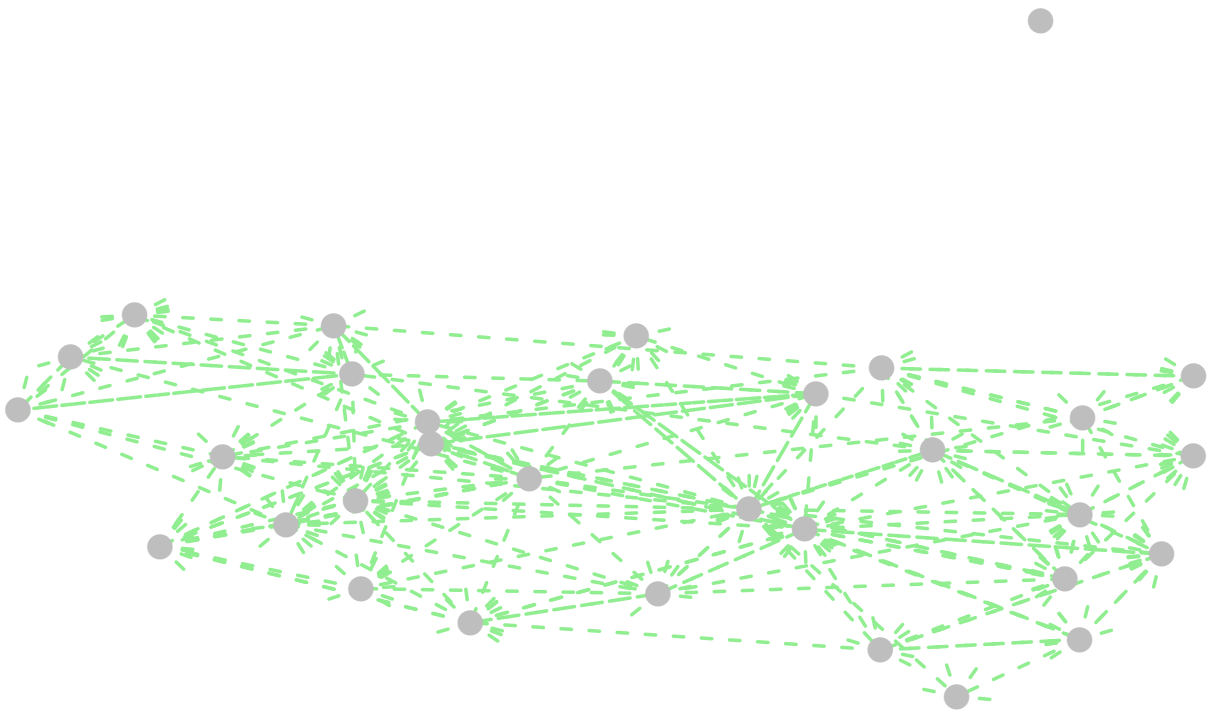
1. Uniformly sample a set $\mathbf{s} = \{s_1, \ldots, s_{n_s}\}$ of nodes

$$\mathbf{s} \subset \{1, \ldots, n\}.$$

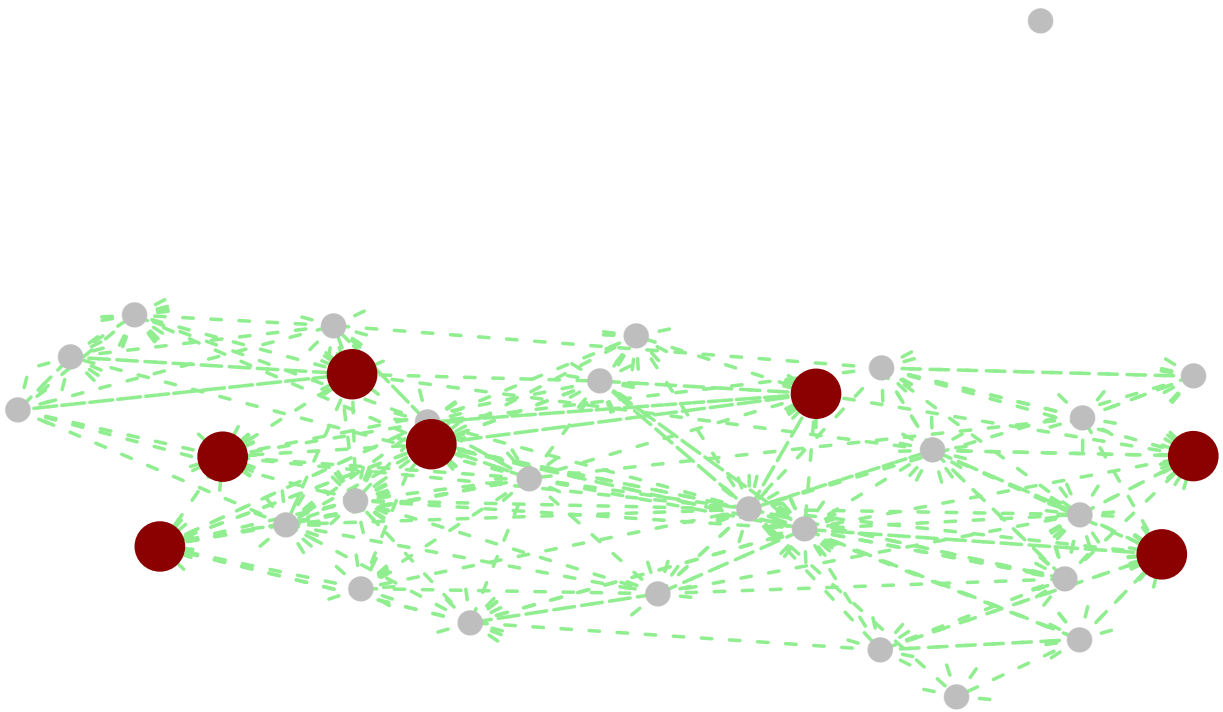
2. Observe relations $\mathbf{y_s}$ between sampled nodes

$$\mathbf{Y_s} = \{y_{i,j} : i \in \mathbf{s}, j \in \mathbf{s}\}.$$

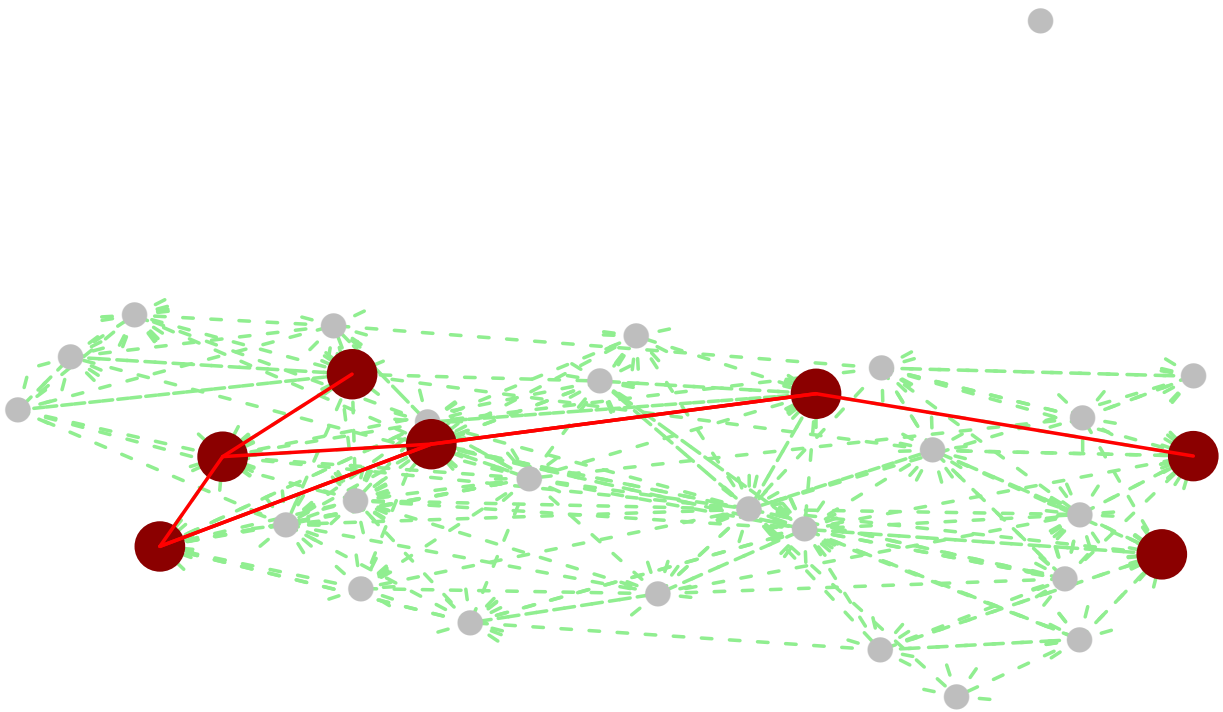
Node-induced subgraph sampling



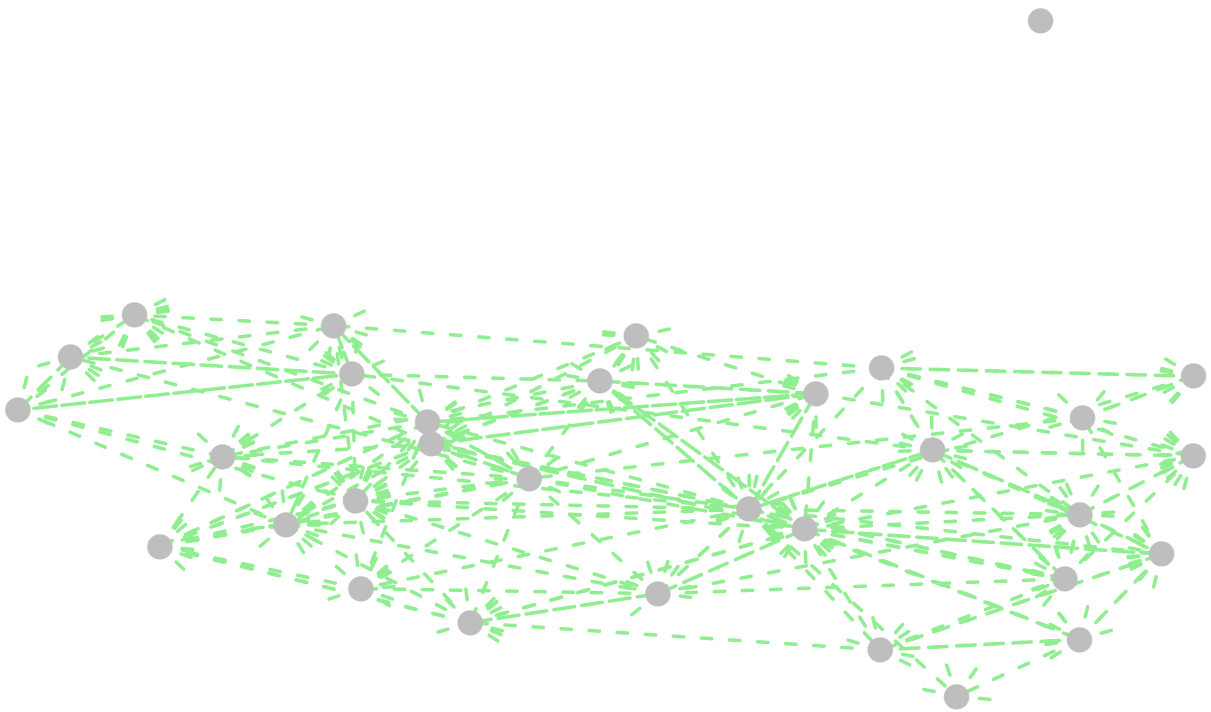
Node-induced subgraph sampling



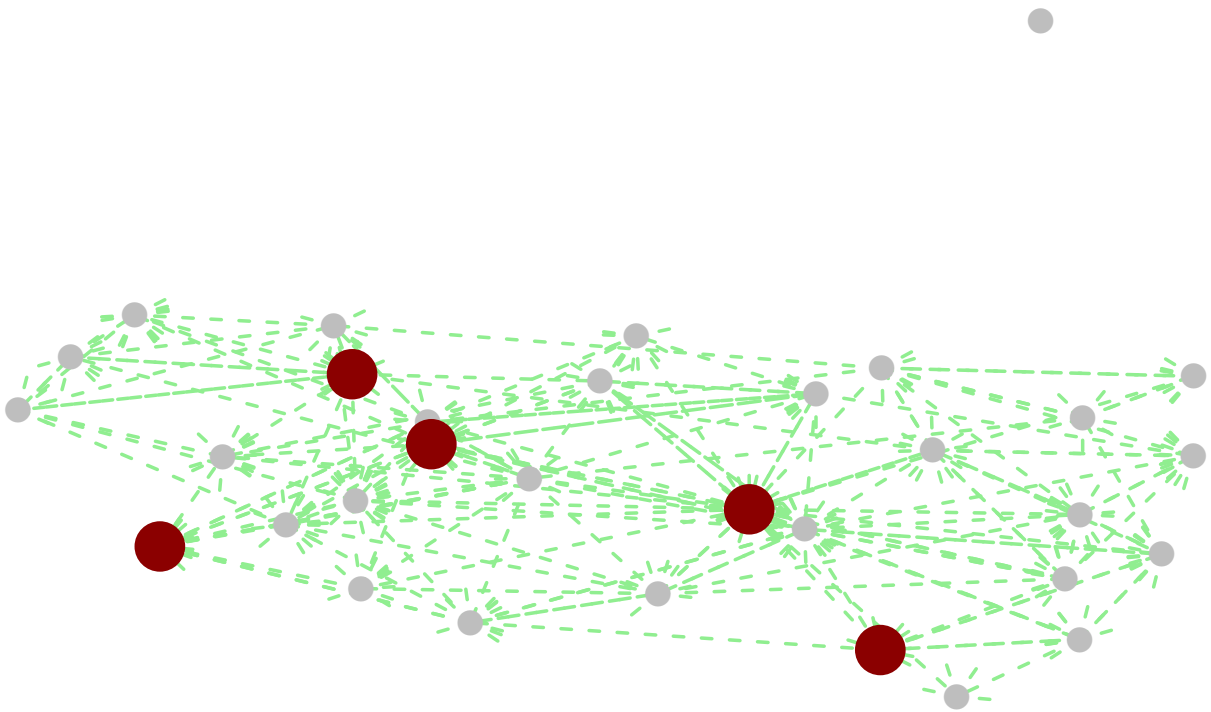
Node-induced subgraph sampling



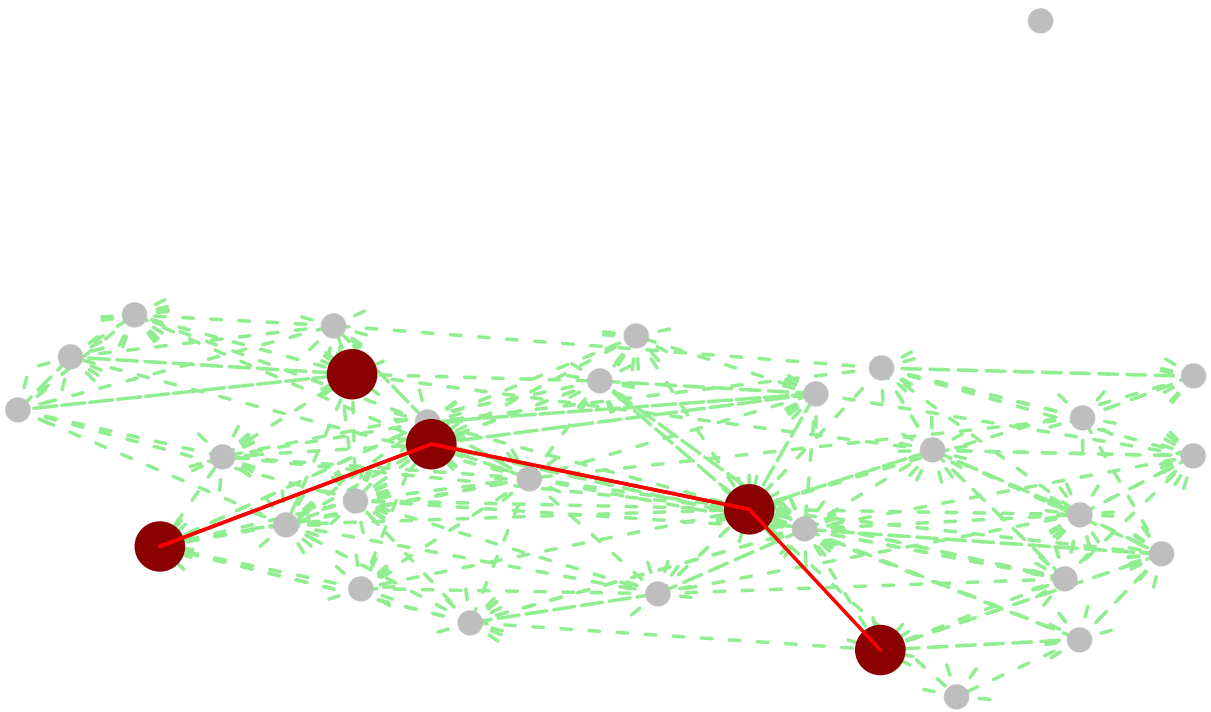
Node-induced subgraph sampling



Node-induced subgraph sampling



Node-induced subgraph sampling



Estimation from sampled data

In what ways does $\mathbf{Y}_s$ resemble $\mathbf{Y}$?

For what functions $g()$ will $g(\mathbf{Y}_s)$ estimate $g(\mathbf{Y})$?

Consider the following setup:

- $n \times n$ sociomatrix $\mathbf{Y}$
- $n \times n$ dyadic covariate $\mathbf{X}_d$
- $n \times 1$ nodal covariate $\mathbf{X}_n$

Can we estimate the following from a sample?

$$\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} \ , \ \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j} \ , \ \bar{x}_n = \frac{1}{n} \sum x_{n,i}$$

$$\overline{yx_d} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} x_{d,i,j} \ , \ \overline{yx_n} = \frac{1}{n(n-1)} \sum_{i} x_{n,i} \bar{y}_{i\cdot}$$

Estimation from sampled data

In what ways does $\mathbf{Y_s}$ resemble $\mathbf{Y}$?

For what functions $g()$ will $g(\mathbf{Y_s})$ estimate $g(\mathbf{Y})$?

Consider the following setup:

- $n \times n$ sociomatrix $\mathbf{Y}$
- $n \times n$ dyadic covariate $\mathbf{X}_d$
- $n \times 1$ nodal covariate $\mathbf{X}_n$

Can we estimate the following from a sample?

$$\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} \ , \ \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j} \ , \ \bar{x}_n = \frac{1}{n} \sum x_{n,i}$$

$$\overline{yx_d} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} x_{d,i,j} \ , \ \overline{yx_n} = \frac{1}{n(n-1)} \sum_{i} x_{n,i} \bar{y}_{i\cdot}$$

Estimation from sampled data

In what ways does $\mathbf{Y_s}$ resemble $\mathbf{Y}$?

For what functions $g()$ will $g(\mathbf{Y_s})$ estimate $g(\mathbf{Y})$?

Consider the following setup:

- $n \times n$ sociomatrix $\mathbf{Y}$
- $n \times n$ dyadic covariate $\mathbf{X}_d$
- $n \times 1$ nodal covariate $\mathbf{X}_n$

Can we estimate the following from a sample?

$$\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} \ , \ \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j} \ , \ \bar{x}_n = \frac{1}{n} \sum x_{n,i}$$

$$\overline{yx_d} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} x_{d,i,j} \ , \ \overline{yx_n} = \frac{1}{n(n-1)} \sum_{i} x_{n,i} \bar{y}_{i\cdot}$$

Estimation from sampled data

In what ways does $\mathbf{Y_s}$ resemble $\mathbf{Y}$?

For what functions $g()$ will $g(\mathbf{Y_s})$ estimate $g(\mathbf{Y})$?

Consider the following setup:

- $n \times n$ sociomatrix $\mathbf{Y}$
- $n \times n$ dyadic covariate $\mathbf{X}_d$
- $n \times 1$ nodal covariate $\mathbf{X}_n$

Can we estimate the following from a sample?

$$\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} \ , \ \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j} \ , \ \bar{x}_n = \frac{1}{n} \sum x_{n,i}$$

$$\overline{yx_d} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} x_{d,i,j} \ , \ \overline{yx_n} = \frac{1}{n(n-1)} \sum_{i} x_{n,i} \bar{y}_{i\cdot}$$

Estimation from sampled data

In what ways does $\mathbf{Y_s}$ resemble $\mathbf{Y}$?

For what functions $g()$ will $g(\mathbf{Y_s})$ estimate $g(\mathbf{Y})$?

Consider the following setup:

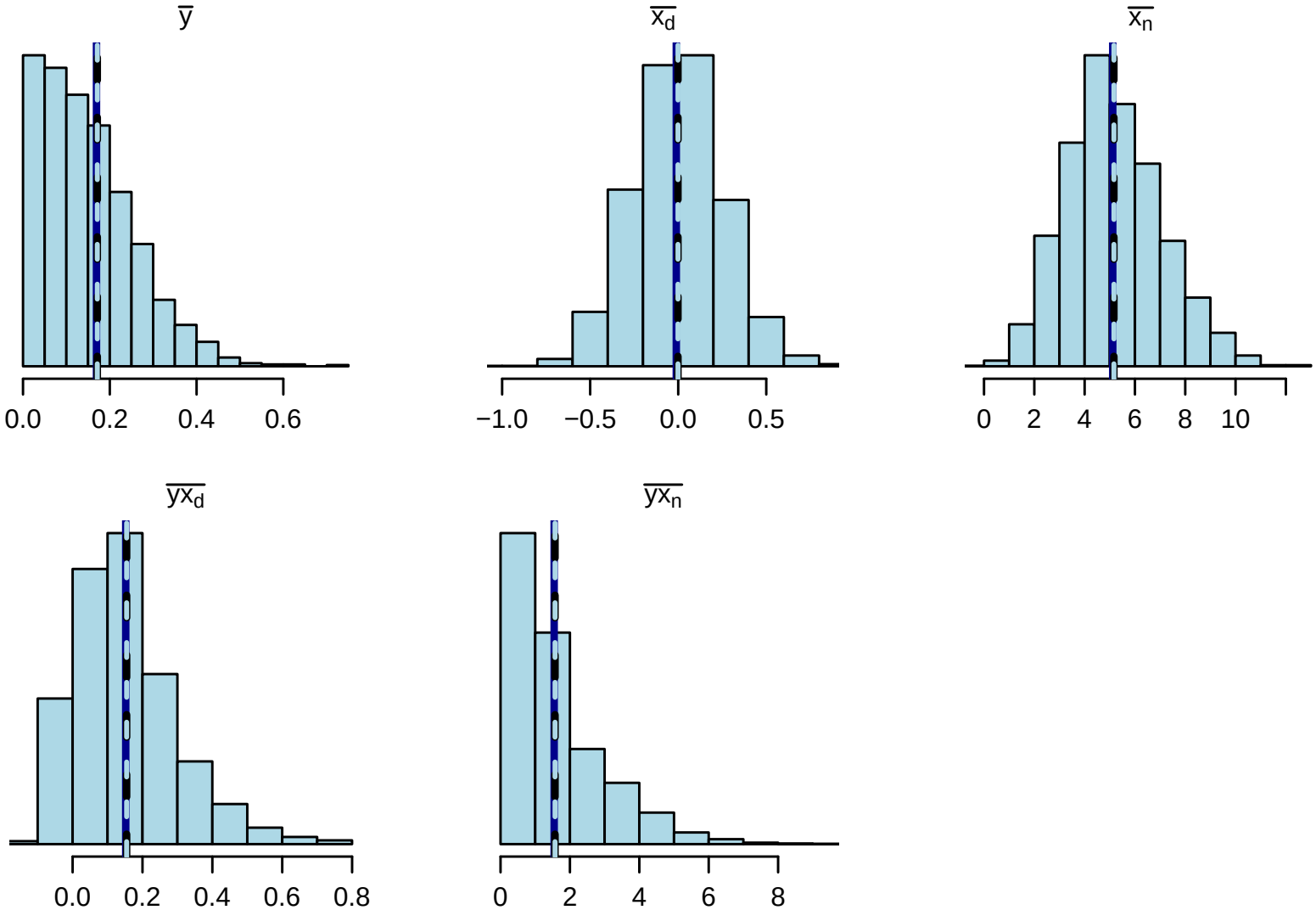
- $n \times n$ sociomatrix $\mathbf{Y}$
- $n \times n$ dyadic covariate $\mathbf{X}_d$
- $n \times 1$ nodal covariate $\mathbf{X}_n$

Can we estimate the following from a sample?

$$\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} \ , \ \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j} \ , \ \bar{x}_n = \frac{1}{n} \sum x_{n,i}$$

$$\overline{yx_d} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} x_{d,i,j} \ , \ \overline{yx_n} = \frac{1}{n(n-1)} \sum_{i} x_{n,i} \bar{y}_{i\cdot}$$

Node-induced subgraph sampling



Node-induced subgraph sampling

For some functions g, the sample value $g(\mathbf{Y}_s)$ is an unbiased estimator of the population value $g(\mathbf{Y})$:

$$g(\mathbf{Y}) = \text{ an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{2}} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \text{ if } n_s \geq 3$$

Why does it work?:
Each subgraph of size k appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:

- in and outdegree distributions;
- geodesics, distances, number of paths, etc.

Node-induced subgraph sampling

For some functions g, the sample value $g(\mathbf{Y}_s)$ is an unbiased estimator of the population value $g(\mathbf{Y})$:

$$g(\mathbf{Y}) = \text{ an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{2}} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \text{ if } n_s \geq 3$$

Why does it work?:

Each subgraph of size k appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:

- in and outdegree distributions;
- geodesics, distances, number of paths, etc.

Node-induced subgraph sampling

For some functions g, the sample value $g(\mathbf{Y_s})$ is an unbiased estimator of the population value $g(\mathbf{Y})$:

$$g(\mathbf{Y}) = \text{ an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{2}} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \text{ if } n_s \geq 3$$

Why does it work?:
Each subgraph of size k appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:

- in and outdegree distributions;
- geodesics, distances, number of paths, etc.

Node-induced subgraph sampling

For some functions g, the sample value $g(\mathbf{Y}_s)$ is an unbiased estimator of the population value $g(\mathbf{Y})$:

$$g(\mathbf{Y}) = \text{ an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{2}} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(\mathbf{Y}) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \quad \text{if } n_s \geq 3$$

Why does it work?:
Each subgraph of size k appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:

- in and outdegree distributions;
- geodesics, distances, number of paths, etc.

Edge-induced subgraph sampling

Procedure:

1. Uniformly sample a set $\mathbf{e} = \{e_1, \ldots, e_{n_e}\}$ of edges

$$\mathbf{e} \subset \{(i,j) : y_{i,j} = 1\}$$

2. Let $\mathbf{Y}_s$ be the edge-generated subgraph of $\mathbf{e}$.

Edge-induced subgraph sampling

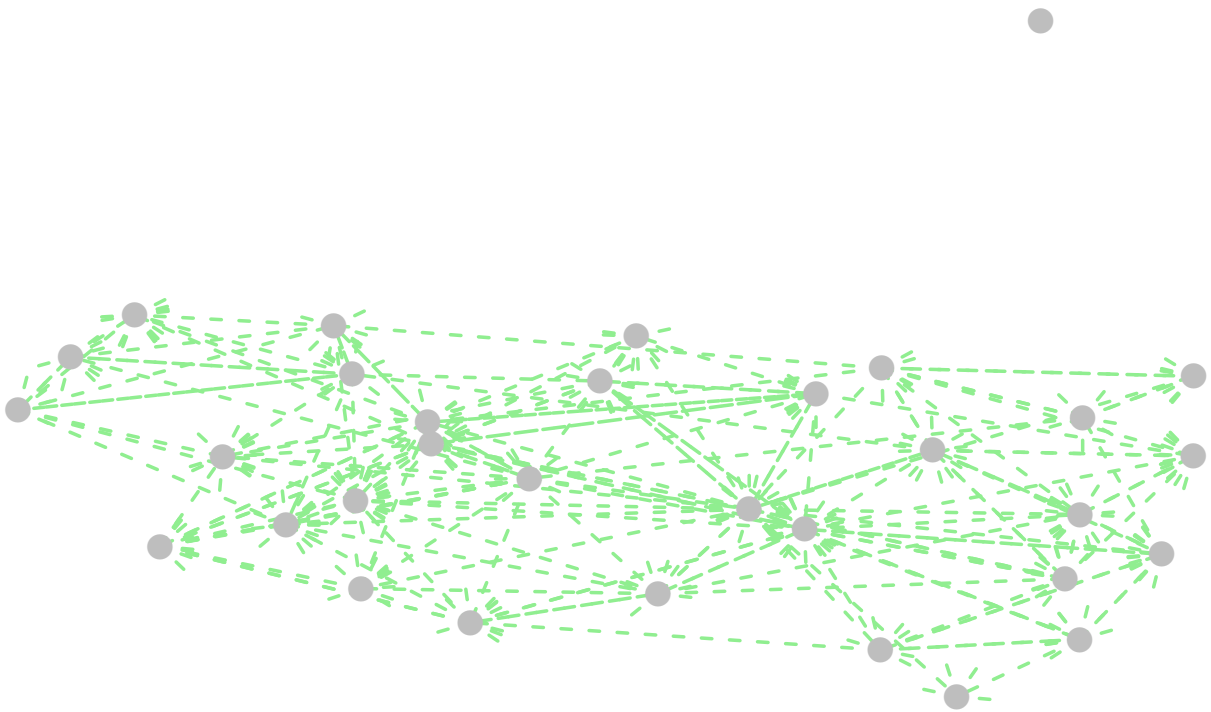
Procedure:

1. Uniformly sample a set $\mathbf{e} = \{e_1, \ldots, e_{n_e}\}$ of edges

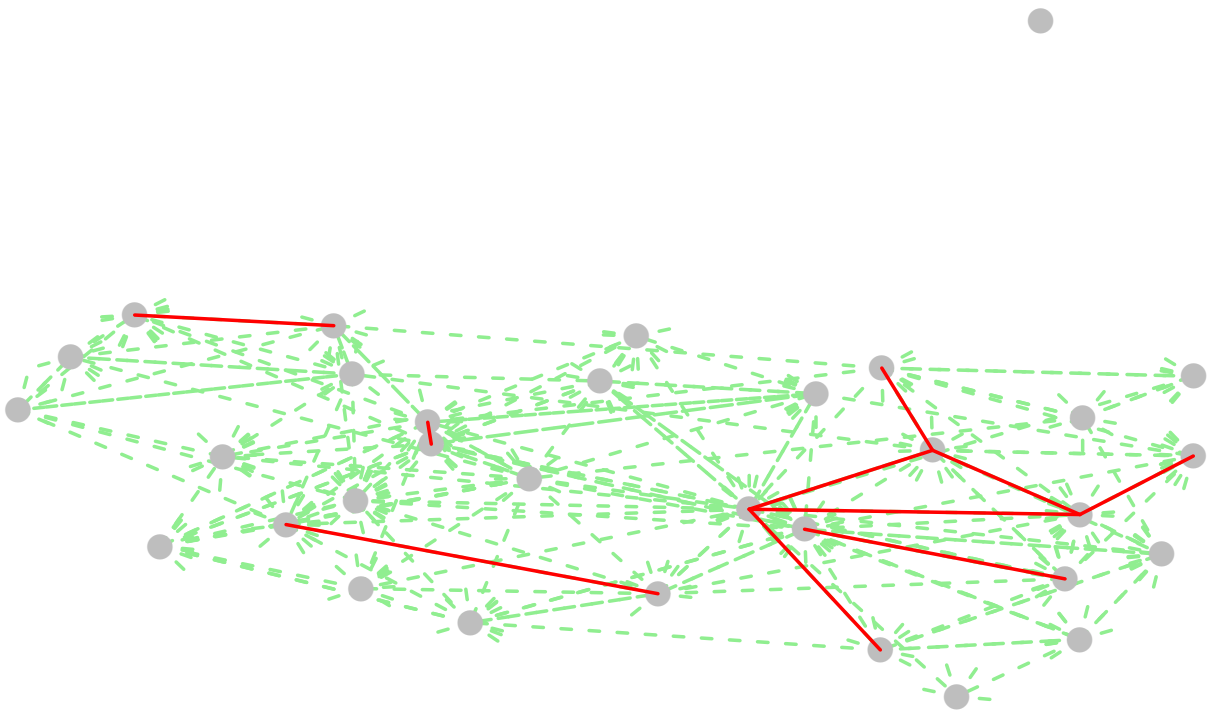
$$\mathbf{e} \subset \{(i,j) : y_{i,j} = 1\}$$

2. Let $\mathbf{Y}_s$ be the edge-generated subgraph of $\mathbf{e}$.

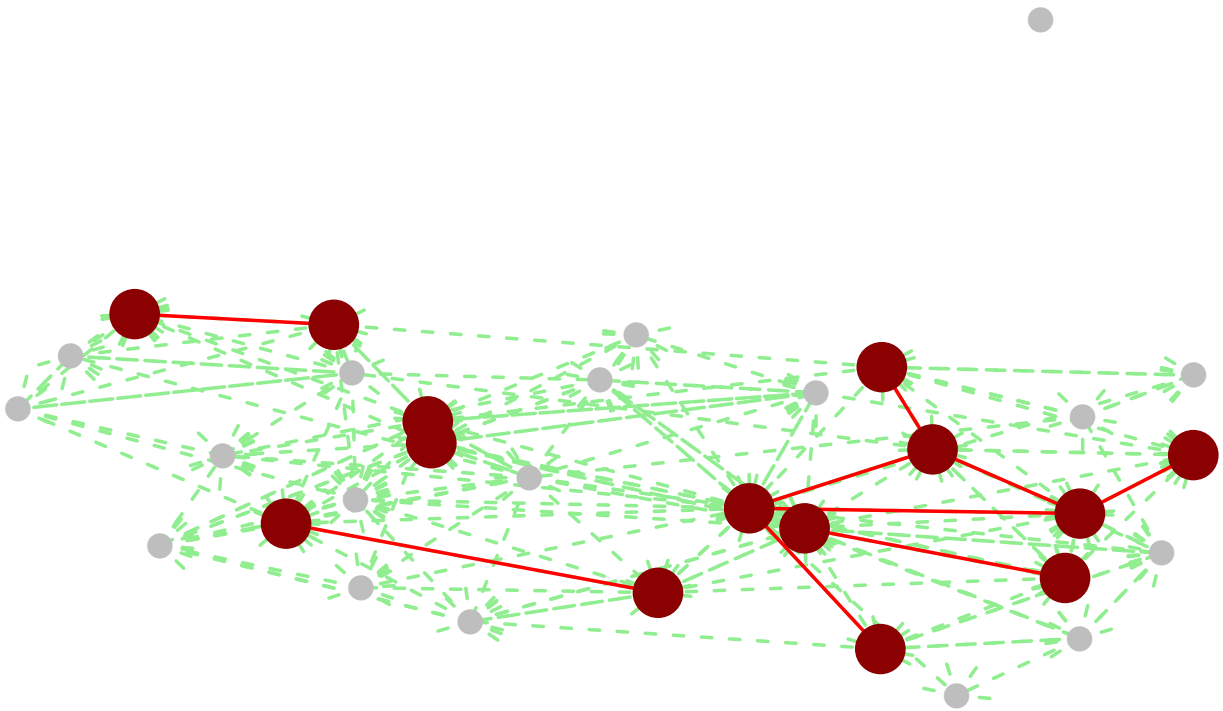
Edge-induced subgraph sampling



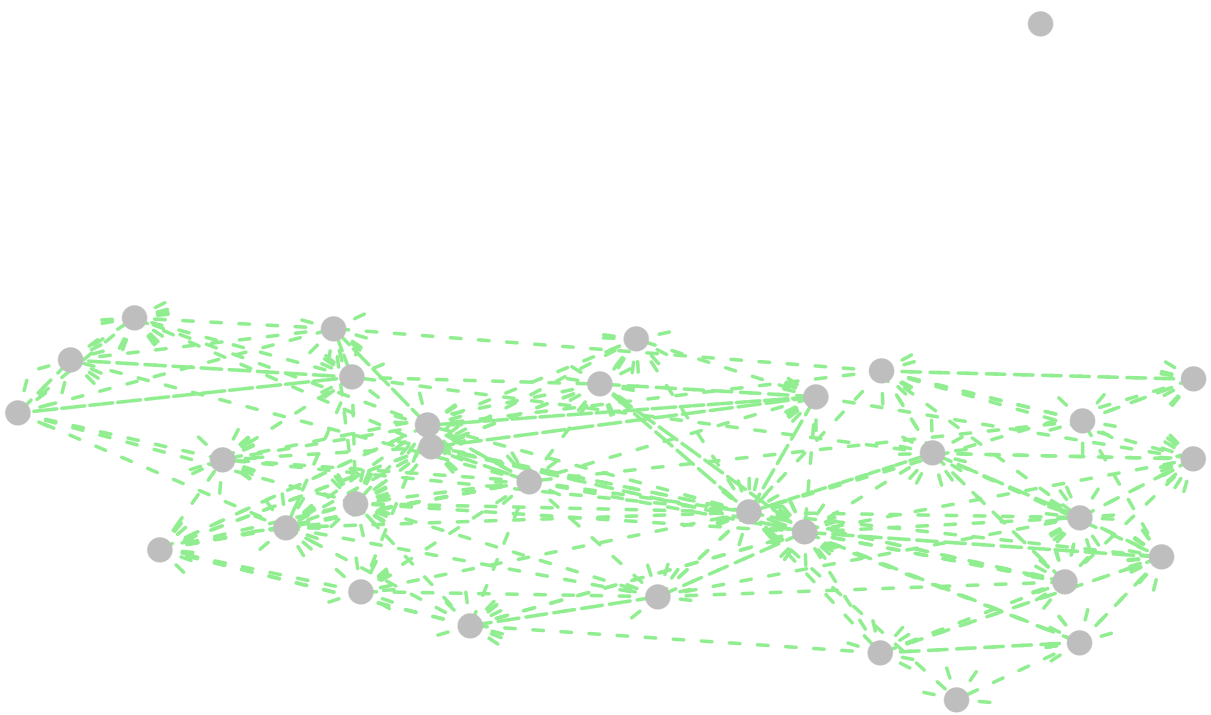
Edge-induced subgraph sampling



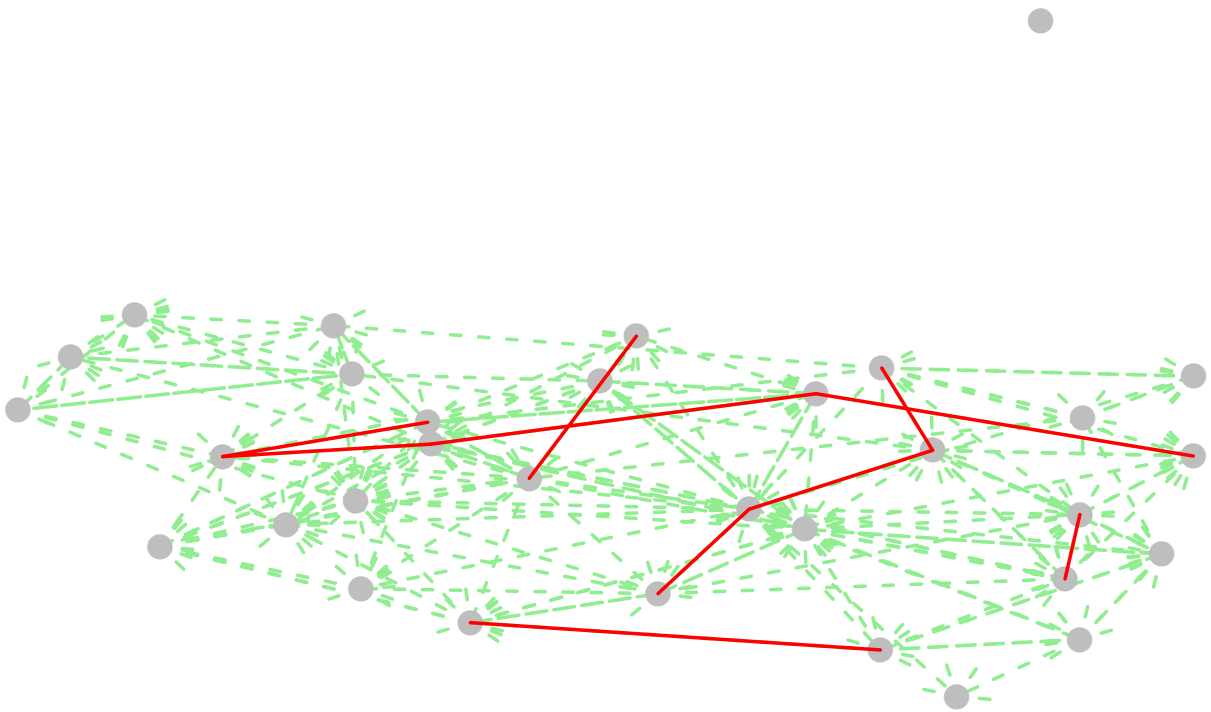
Edge-induced subgraph sampling



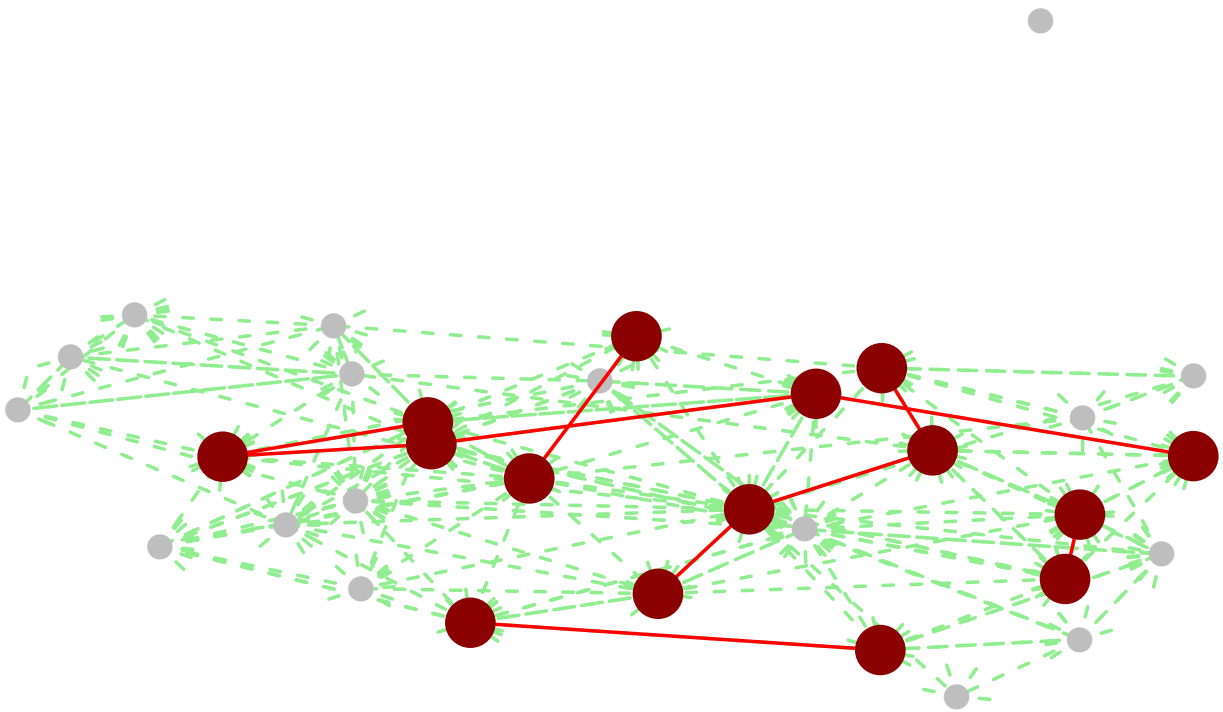
Edge-induced subgraph sampling



Edge-induced subgraph sampling



Edge-induced subgraph sampling



Edge-induced subgraph sampling

How well do these subgraphs represent $\mathbf{Y}$?

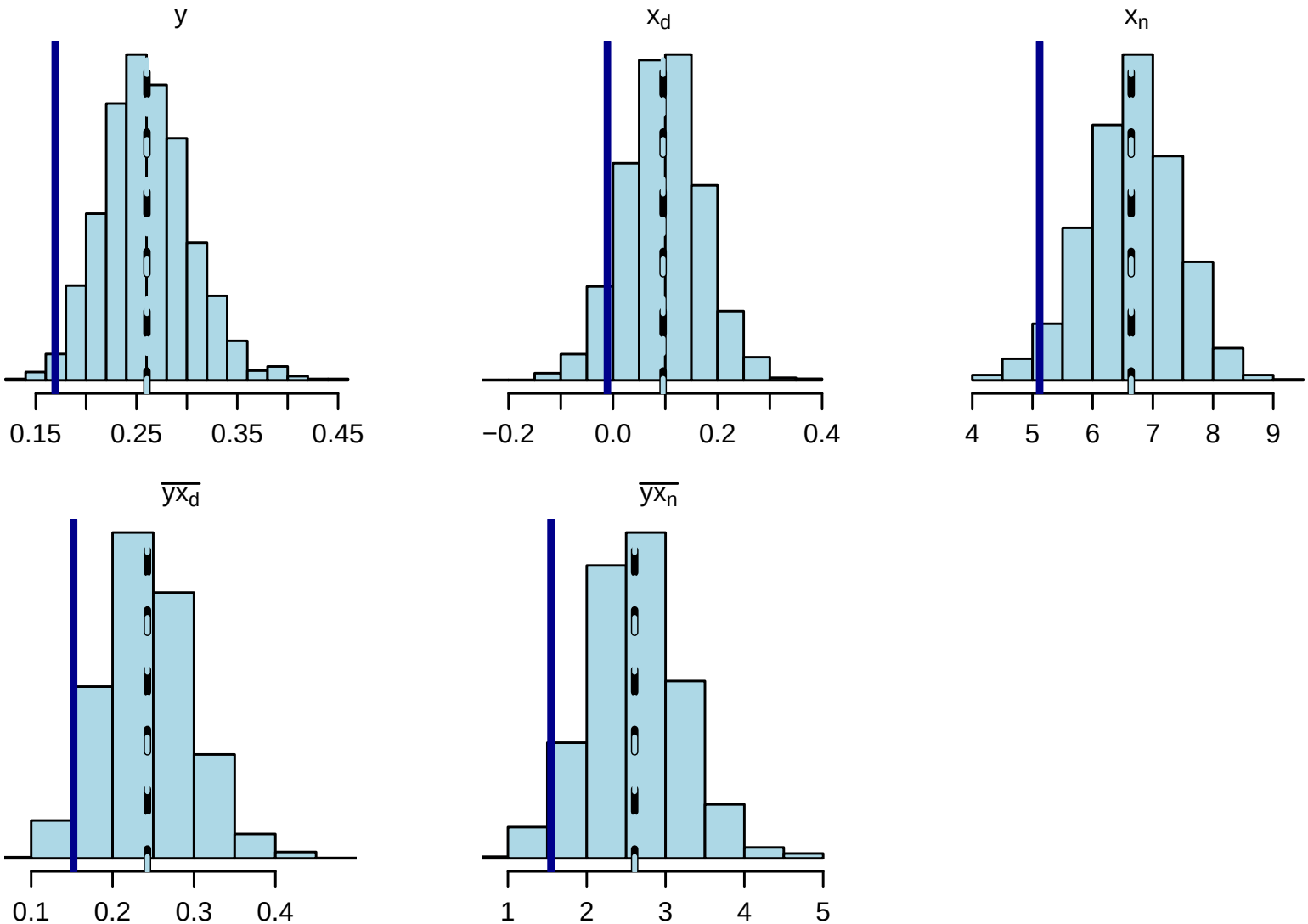
Can you infer anything about $\mathbf{Y}$ from these data?

Edge-induced subgraph sampling

How well do these subgraphs represent $\mathbf{Y}$?

Can you infer anything about $\mathbf{Y}$ from these data?

Edge-induced subgraph sampling



Egocentric sampling

Procedure:

1. Uniformly sample a set $\mathbf{s}_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

$$\mathbf{s}_1 \subset \{1, \ldots, n\}.$$

2. Observe the relations for each $i \in \mathbf{s}_1$, i.e. observe $\{y_{i,1}, \ldots, y_{i,n}\}$.
3. Let $\mathbf{s}_2$ be the set of nodes having a link from anyone in $\mathbf{s}_1$. Observe the relations of anyone in $\mathbf{s}_2$ to anyone in $\mathbf{s}_1 \cup \mathbf{s}_2$.

$$\mathbf{Y}_\mathbf{s} = \{y_{i,j} : i, j \in \mathbf{s}_1 \cup \mathbf{s}_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in \mathbf{s}_1$ the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?

Egocentric sampling

Procedure:

1. Uniformly sample a set $\mathbf{s}_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

$$\mathbf{s}_1 \subset \{1, \ldots, n\}.$$

2. Observe the relations for each $i \in \mathbf{s}_1$, i.e. observe $\{y_{i,1}, \ldots, y_{i,n}\}$.
3. Let $\mathbf{s}_2$ be the set of nodes having a link from anyone in $\mathbf{s}_1$. Observe the relations of anyone in $\mathbf{s}_2$ to anyone in $\mathbf{s}_1 \cup \mathbf{s}_2$.

$$\mathbf{Y}_\mathbf{s} = \{y_{i,j} : i, j \in \mathbf{s}_1 \cup \mathbf{s}_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in \mathbf{s}_1$ the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?

Egocentric sampling

Procedure:

1. Uniformly sample a set $\mathbf{s}_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

$$\mathbf{s}_1 \subset \{1, \ldots, n\}.$$

2. Observe the relations for each $i \in \mathbf{s}_1$, i.e. observe $\{y_{i,1}, \ldots, y_{i,n}\}$.
3. Let $\mathbf{s}_2$ be the set of nodes having a link from anyone in $\mathbf{s}_1$. Observe the relations of anyone in $\mathbf{s}_2$ to anyone in $\mathbf{s}_1 \cup \mathbf{s}_2$.

$$\mathbf{Y}_\mathbf{s} = \{y_{i,j} : i, j \in \mathbf{s}_1 \cup \mathbf{s}_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in \mathbf{s}_1$ the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?

Egocentric sampling

Procedure:

1. Uniformly sample a set $\mathbf{s}_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

$$\mathbf{s}_1 \subset \{1, \ldots, n\}.$$

2. Observe the relations for each $i \in \mathbf{s}_1$, i.e. observe $\{y_{i,1}, \ldots, y_{i,n}\}$.
3. Let $\mathbf{s}_2$ be the set of nodes having a link from anyone in $\mathbf{s}_1$. Observe the relations of anyone in $\mathbf{s}_2$ to anyone in $\mathbf{s}_1 \cup \mathbf{s}_2$.

$$\mathbf{Y}_\mathbf{s} = \{y_{i,j} : i, j \in \mathbf{s}_1 \cup \mathbf{s}_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in \mathbf{s}_1$ the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?

Egocentric sampling

Procedure:

1. Uniformly sample a set $\mathbf{s}_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

$$\mathbf{s}_1 \subset \{1, \ldots, n\}.$$

2. Observe the relations for each $i \in \mathbf{s}_1$, i.e. observe $\{y_{i,1}, \ldots, y_{i,n}\}$.
3. Let $\mathbf{s}_2$ be the set of nodes having a link from anyone in $\mathbf{s}_1$. Observe the relations of anyone in $\mathbf{s}_2$ to anyone in $\mathbf{s}_1 \cup \mathbf{s}_2$.

$$\mathbf{Y}_\mathbf{s} = \{y_{i,j} : i, j \in \mathbf{s}_1 \cup \mathbf{s}_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in \mathbf{s}_1$ the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?

Link-tracing designs

Snowball sampling: Iteratively repeat the egocentric sampler, obtaining the stage-k nodes $\mathbf{s}_k$ from the links of $\mathbf{s}_{k-1}$.

This is a type of link-tracing design. The links of the current nodes determine who is next to be included in the sample.

How will such subgraphs $\mathbf{Y}_s$ be similar to $\mathbf{Y}$?
How will they differ?

Link-tracing designs

Snowball sampling: Iteratively repeat the egocentric sampler, obtaining the stage-k nodes $\mathbf{s}_k$ from the links of $\mathbf{s}_{k-1}$.

This is a type of link-tracing design. The links of the current nodes determine who is next to be included in the sample.

How will such subgraphs $\mathbf{Y}_s$ be similar to $\mathbf{Y}$?
How will they differ?

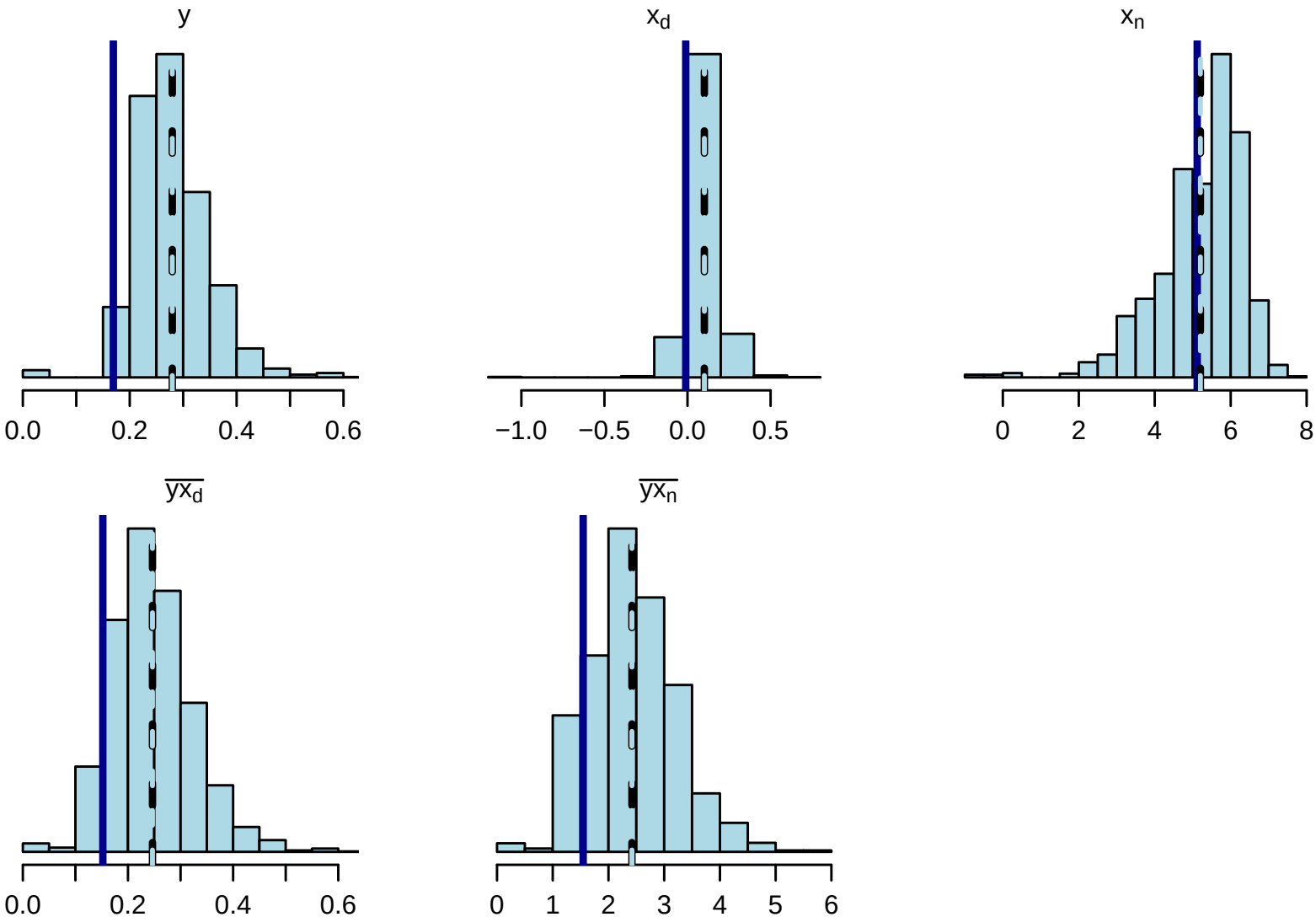
Link-tracing designs

Snowball sampling: Iteratively repeat the egocentric sampler, obtaining the stage-k nodes $\mathbf{s}_k$ from the links of $\mathbf{s}_{k-1}$.

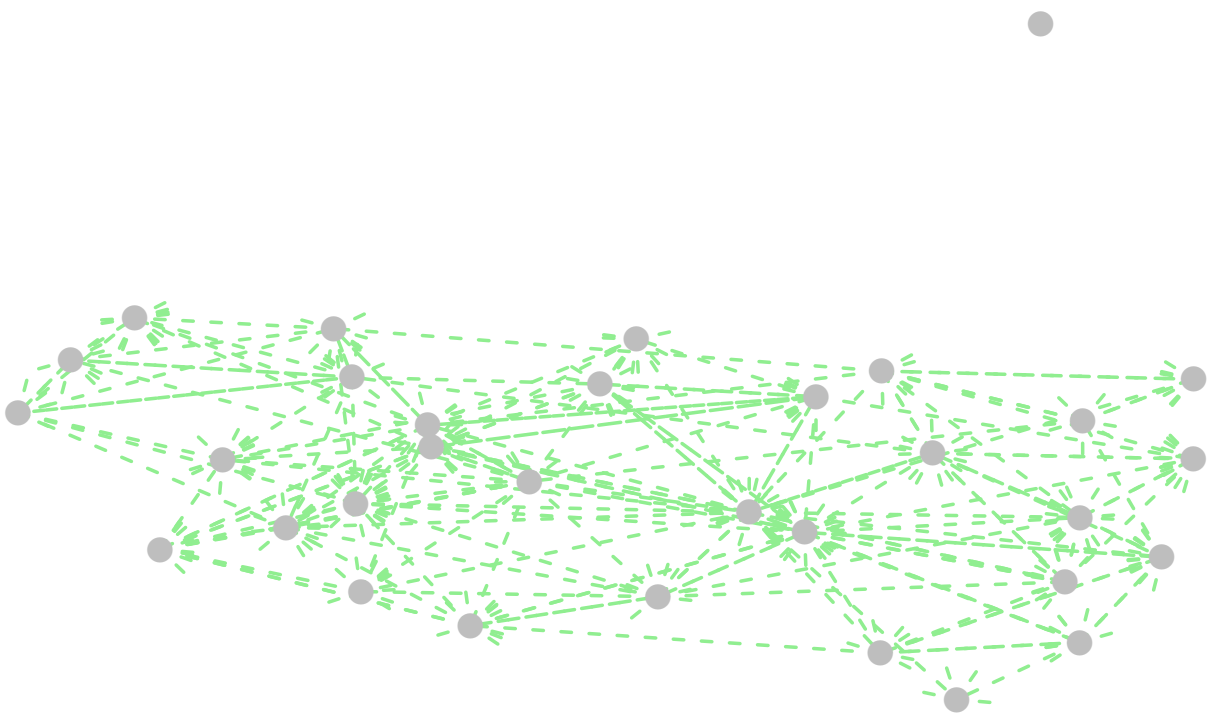
This is a type of link-tracing design. The links of the current nodes determine who is next to be included in the sample.

How will such subgraphs $\mathbf{Y_s}$ be similar to $\mathbf{Y}$?
How will they differ?

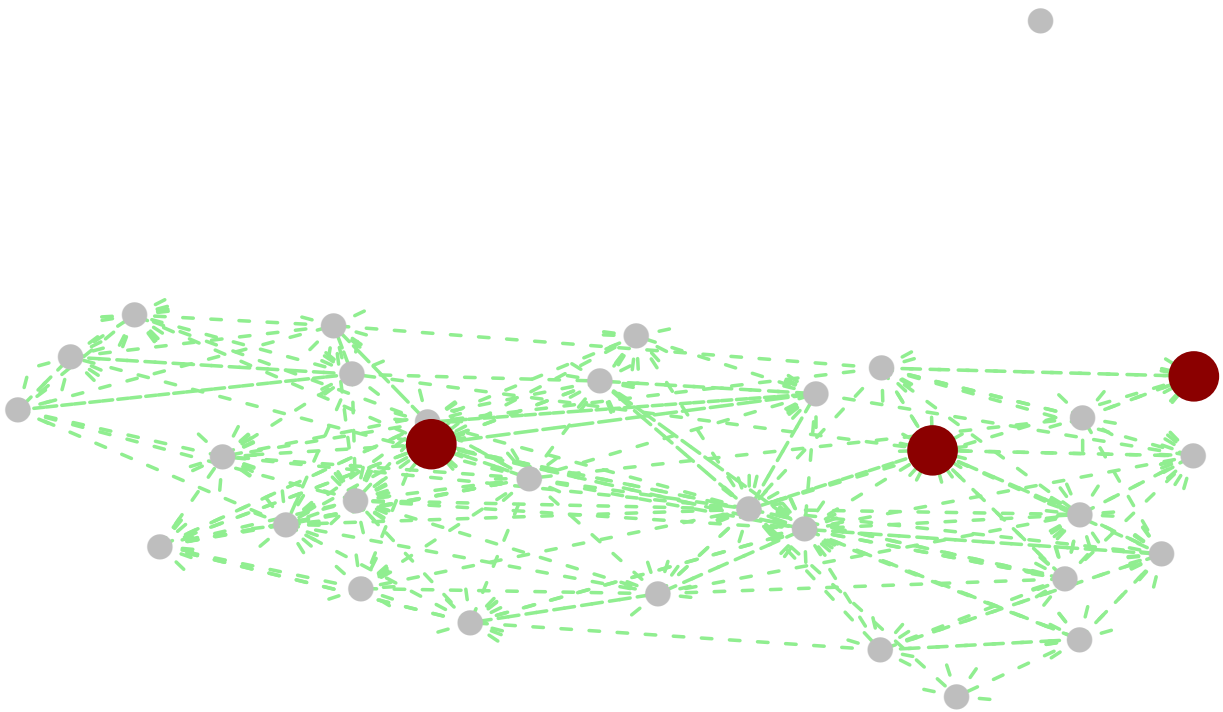
Egocentric sampling



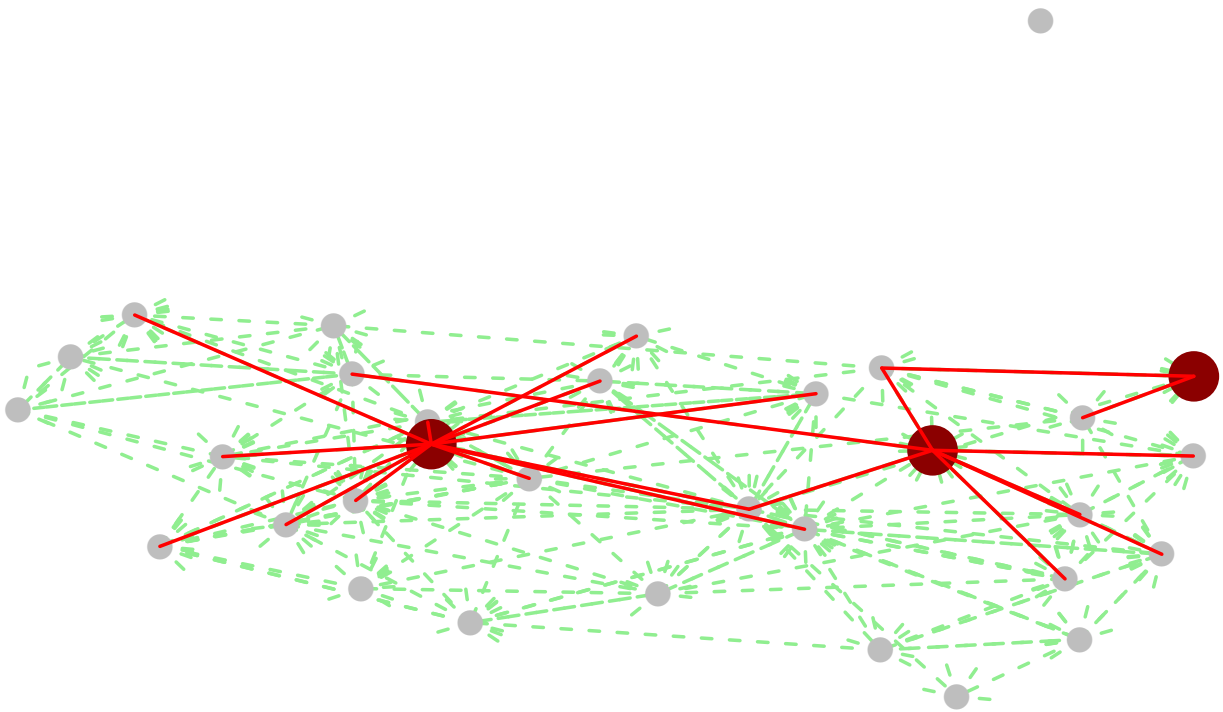
Egocentric sampling



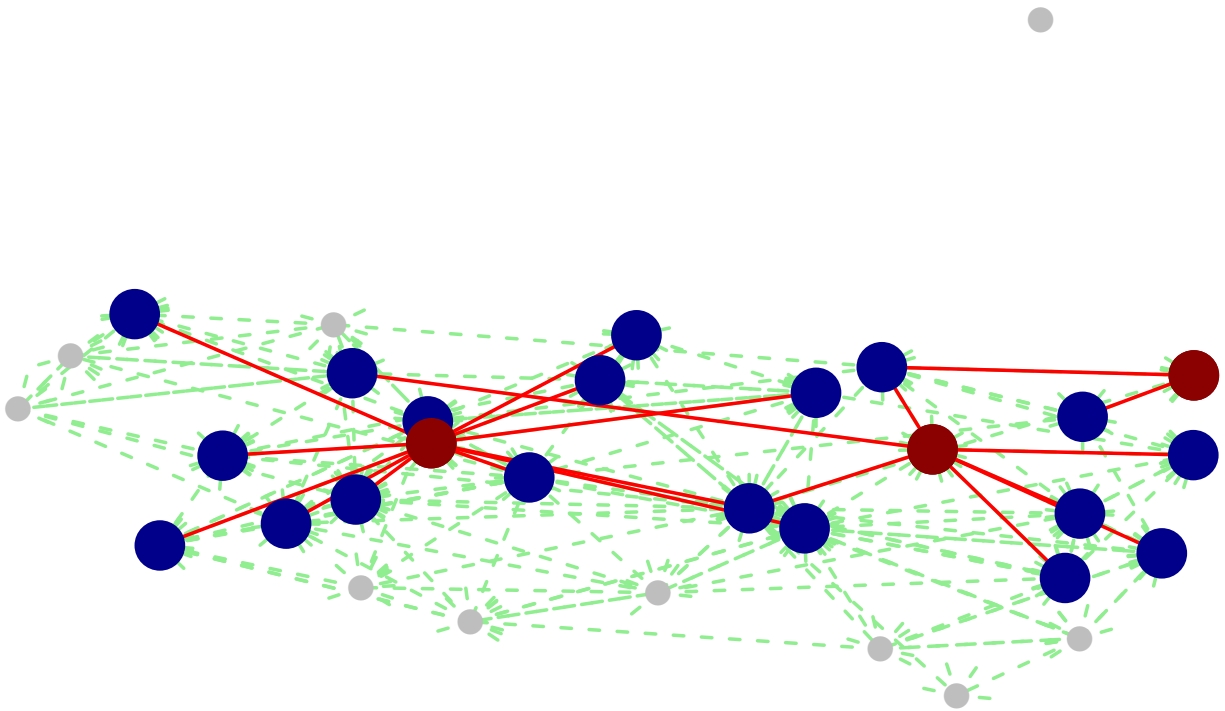
Egocentric sampling



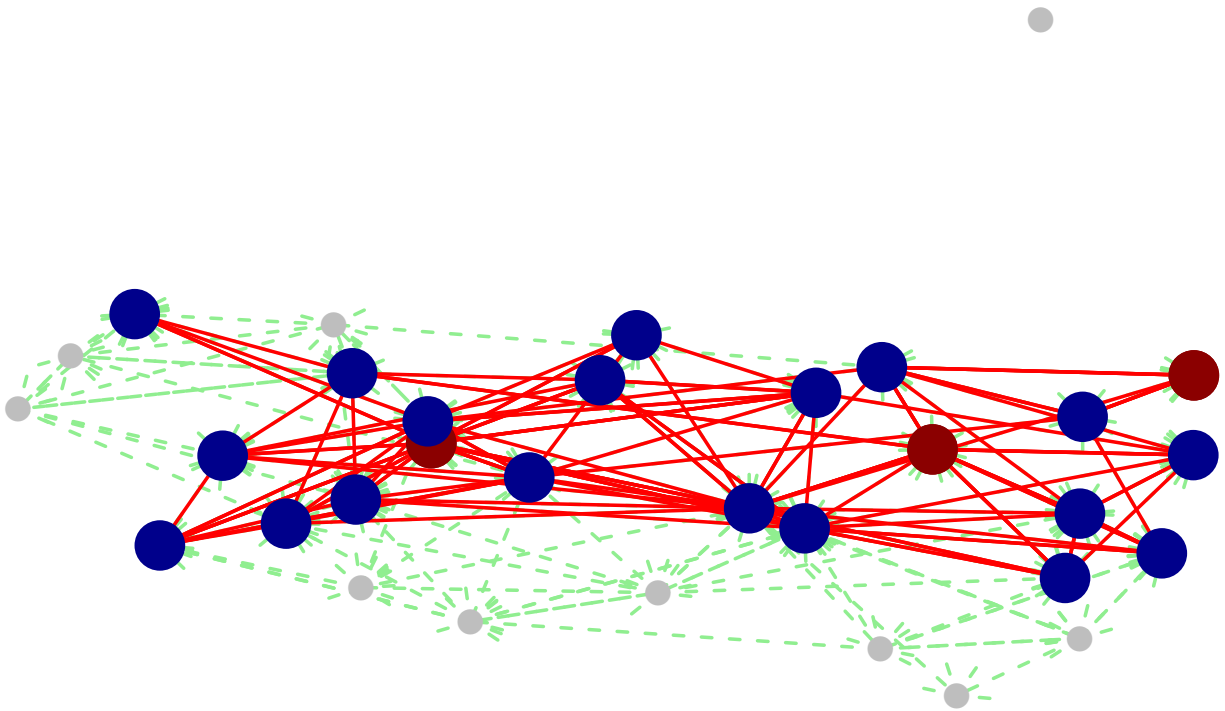
Egocentric sampling



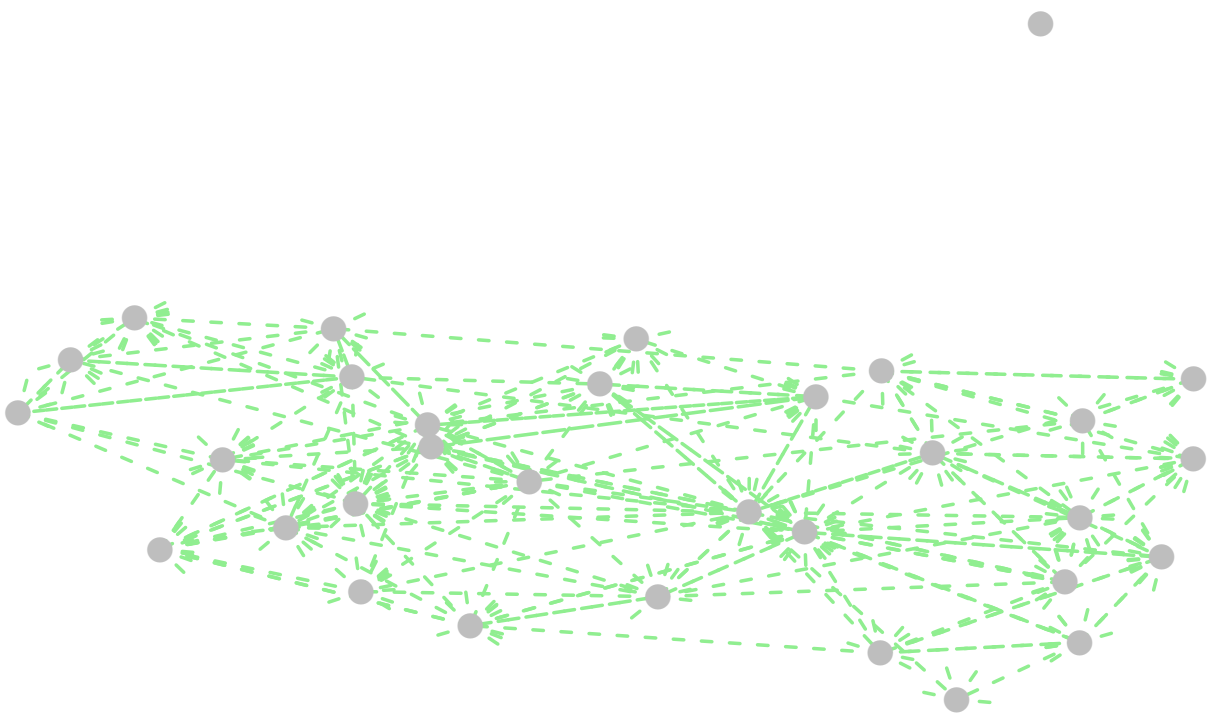
Egocentric sampling



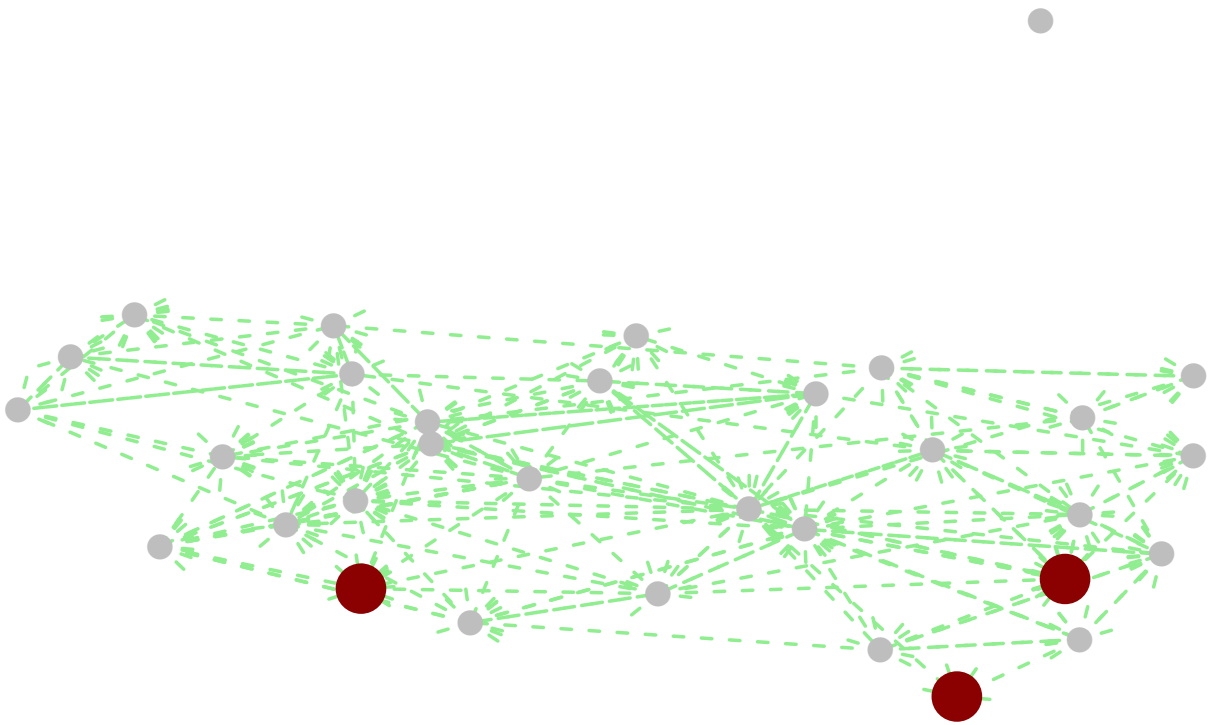
Egocentric sampling



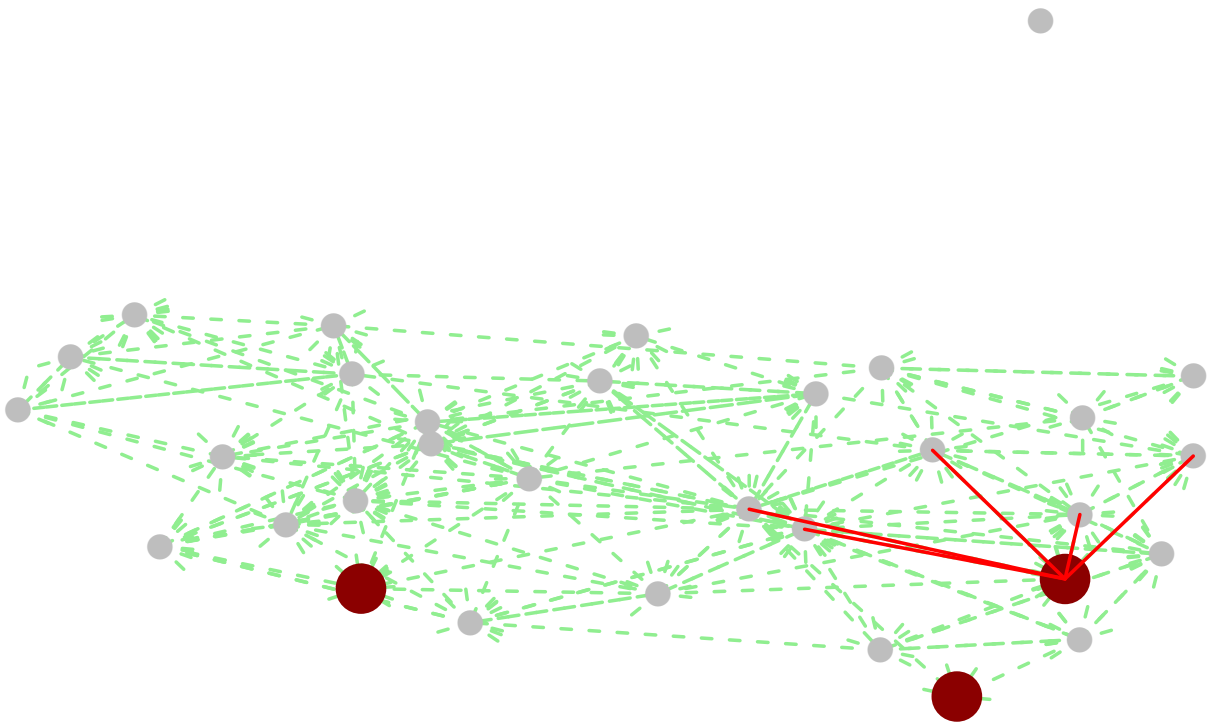
Egocentric sampling



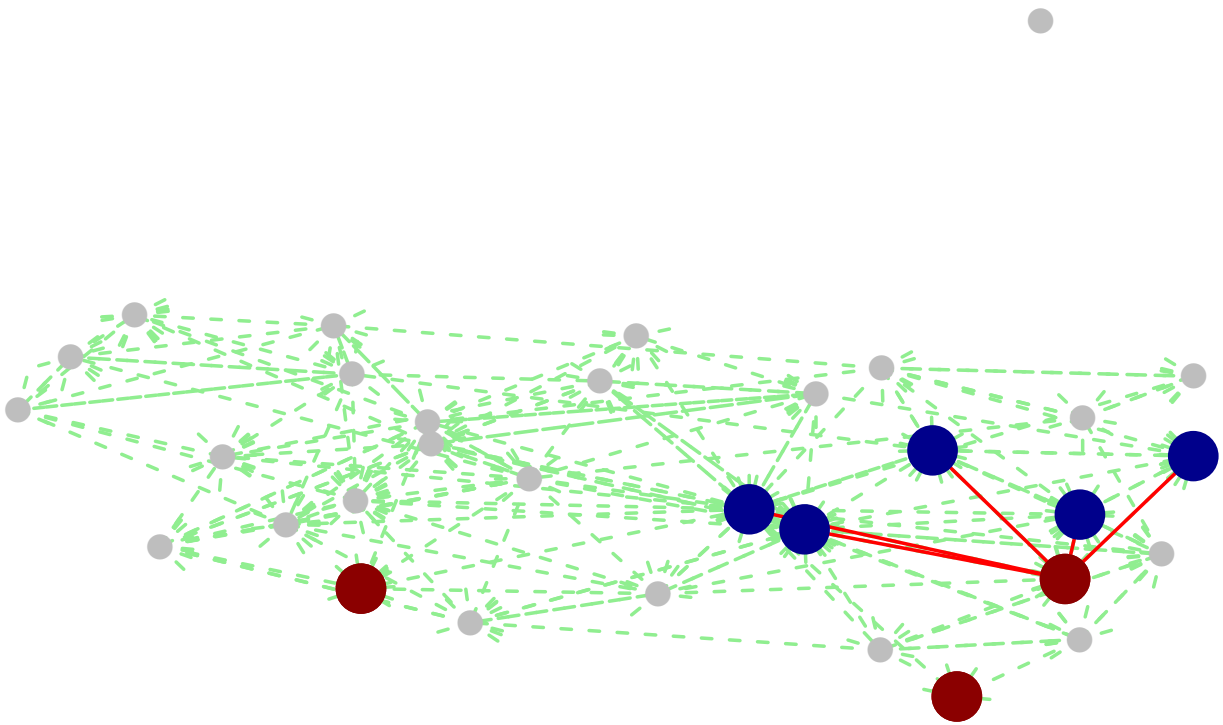
Egocentric sampling



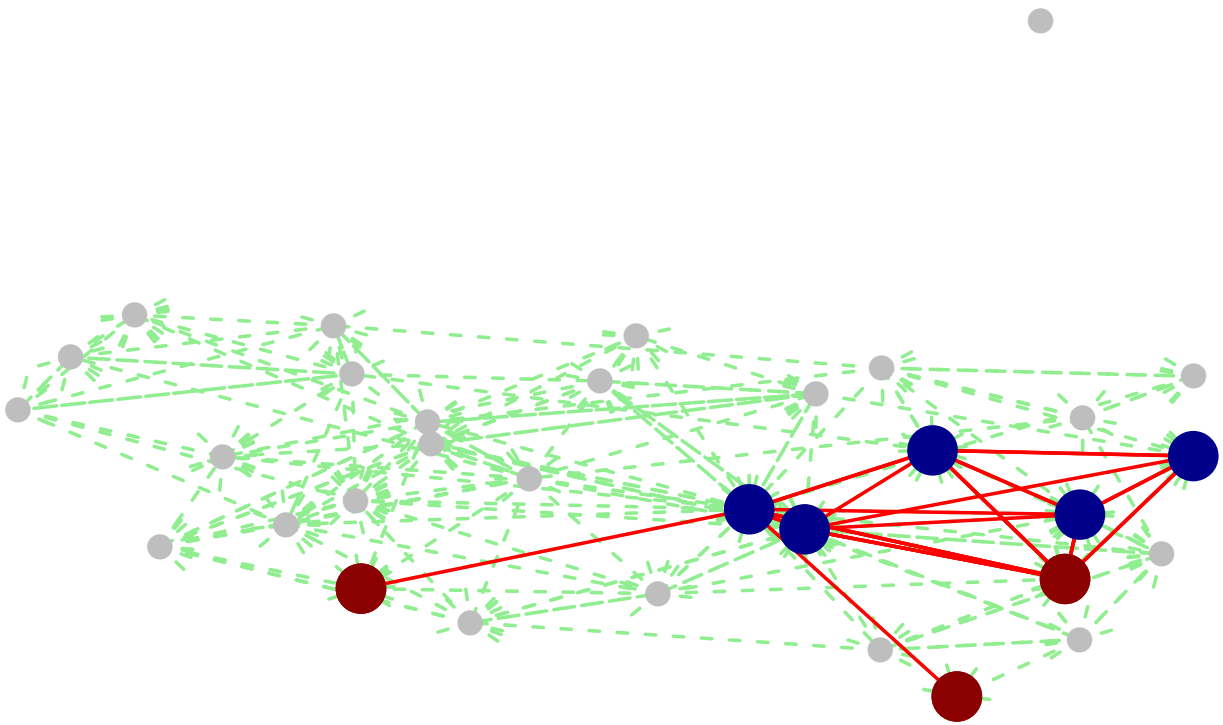
Egocentric sampling



Egocentric sampling



Egocentric sampling



Inference with egocentric samples

- $\mathbf{Y}_s$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y}_s$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y}_s$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

  However, use of data from s_2 generally requires a reweighting scheme.

Inference with egocentric samples

- $\mathbf{Y}_s$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y}_s$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y}_s$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

  However, use of data from s_2 generally requires a reweighting scheme.

Inference with egocentric samples

- $\mathbf{Y_s}$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y_s}$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y_s}$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

  However, use of data from s_2 generally requires a reweighting scheme.

Inference with egocentric samples

- $\mathbf{Y_s}$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y_s}$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y_s}$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

However, use of data from s_2 generally requires a reweighting scheme.

Inference with egocentric samples

- $\mathbf{Y_s}$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y_s}$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y_s}$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

However, use of data from s_2 generally requires a reweighting scheme.

Inference with egocentric samples

- $\mathbf{Y_s}$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y_s}$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y_s}$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

However, use of data from s_2 generally requires a reweighting scheme.

Inference with egocentric samples

- $\mathbf{Y_s}$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y_s}$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y_s}$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

  However, use of data from $\mathbf{s}_2$ generally requires a reweighting scheme.

Inference with egocentric samples

- $\mathbf{Y_s}$ is not generally representative of $\mathbf{Y}$.
- For some statistics, weighted averages based on $\mathbf{Y_s}$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, *part* of $\mathbf{Y_s}$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

  However, use of data from $\mathbf{s}_2$ generally requires a reweighting scheme.

References

- Snijders (1992), "Estimation on the basis of snowball samples: How to Weight?"
- Kolaczyk (2009) "Sampling and Estimation in Network Graphs," chapter 5 of *Statistical Analysis of Network Data*.

Parameter estimation with incomplete sampled data

Model: $\Pr(\mathbf{Y} = \mathbf{y}|\theta), \theta \in \Theta$.

Complete data: $\mathbf{Y}$

Observed data: $\mathbf{Y}[\mathbf{O}]$, where $\mathbf{O}$ is a set of pairs of indices

$$\mathbf{O} = \begin{pmatrix} i_1 & j_1 \\ i_2 & j_2 \\ i_3 & j_3 \\ \vdots & \vdots \\ i_s & j_s \end{pmatrix}$$

How can we make inference for θ based on $\mathbf{Y}[\mathbf{O}]$?

Parameter estimation with incomplete sampled data

Model: $\Pr(\mathbf{Y} = \mathbf{y}|\theta), \theta \in \Theta$.

Complete data: $\mathbf{Y}$

Observed data: $\mathbf{Y}[\mathbf{O}]$, where $\mathbf{O}$ is a set of pairs of indices

$$\mathbf{O} = \begin{pmatrix} i_1 & j_1 \\ i_2 & j_2 \\ i_3 & j_3 \\ \vdots & \vdots \\ i_s & j_s \end{pmatrix}$$

How can we make inference for θ based on $\mathbf{Y}[\mathbf{O}]$?

Parameter estimation with incomplete sampled data

Model: $\Pr(\mathbf{Y} = \mathbf{y}|\theta), \theta \in \Theta$.

Complete data: $\mathbf{Y}$

Observed data: $\mathbf{Y}[\mathbf{O}]$, where $\mathbf{O}$ is a set of pairs of indices

$$\mathbf{O} = \begin{pmatrix} i_1 & j_1 \\ i_2 & j_2 \\ i_3 & j_3 \\ \vdots & \vdots \\ i_s & j_s \end{pmatrix}$$

How can we make inference for θ based on $\mathbf{Y}[\mathbf{O}]$?

Parameter estimation with incomplete sampled data

Model: $\Pr(\mathbf{Y} = \mathbf{y}|\theta), \theta \in \Theta$.

Complete data: $\mathbf{Y}$

Observed data: $\mathbf{Y}[\mathbf{O}]$, where $\mathbf{O}$ is a set of pairs of indices

$$\mathbf{O} = \begin{pmatrix} i_1 & j_1 \\ i_2 & j_2 \\ i_3 & j_3 \\ \vdots & \vdots \\ i_s & j_s \end{pmatrix}$$

How can we make inference for θ based on $\mathbf{Y}[\mathbf{O}]$?

Node-induced subgraph sampling

	1	2	3	4	5	6
1	*NA*	0	1	0	0	1
2	0	*NA*	1	0	0	1
3	1	0	*NA*	1	0	0
4	0	1	0	*NA*	0	0
5	0	0	1	0	*NA*	1
6	1	0	1	0	0	*NA*

Node-induced subgraph sampling

	1	**2**	**3**	4	**5**	6
1	*NA*	0	1	0	0	1
2	0	*NA*	1	0	0	1
3	1	0	*NA*	1	0	0
4	0	1	0	*NA*	0	0
5	0	0	1	0	*NA*	1
6	1	0	1	0	0	*NA*

Node-induced subgraph sampling

	1	**2**	**3**	4	**5**	6
1	*NA*	0	1	0	0	1
2	0	**NA**	**1**	0	**0**	1
3	1	**0**	**NA**	1	**0**	0
4	0	1	0	*NA*	0	0
5	0	**0**	**1**	0	**NA**	1
6	1	0	1	0	0	*NA*

Node-induced subgraph sampling: Observed data

	1	**2**	**3**	4	**5**	6
1	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2	*NA*	**NA**	**1**	*NA*	**0**	*NA*
3	*NA*	**0**	**NA**	*NA*	**0**	*NA*
4	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
5	*NA*	**0**	**1**	*NA*	**NA**	*NA*
6	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

Edge-induced subgraph sampling

	1	2	3	4	5	6
1	*NA*	0	1	0	0	1
2	0	*NA*	1	0	0	1
3	1	0	*NA*	1	0	0
4	0	1	0	*NA*	0	0
5	0	0	1	0	*NA*	1
6	1	0	1	0	0	*NA*

Edge-induced subgraph sampling

	1	2	3	4	5	6
1	*NA*	0	1	0	0	**1**
2	0	*NA*	**1**	0	0	1
3	**1**	0	*NA*	1	0	0
4	0	1	0	*NA*	0	0
5	0	0	**1**	0	*NA*	1
6	1	0	**1**	0	0	*NA*

Edge-induced subgraph sampling: Observed data

	1	2	3	4	5	6
1	*NA*	*NA*	*NA*	*NA*	*NA*	**1**
2	*NA*	*NA*	**1**	*NA*	*NA*	*NA*
3	**1**	*NA*	*NA*	*NA*	*NA*	*NA*
4	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
5	*NA*	*NA*	**1**	*NA*	*NA*	*NA*
6	*NA*	*NA*	**1**	*NA*	*NA*	*NA*

Egocentric sampling

	1	2	3	4	5	6
1	*NA*	0	1	0	0	1
2	0	*NA*	1	0	0	1
3	1	0	*NA*	1	0	0
4	0	1	0	*NA*	0	0
5	0	0	1	0	*NA*	1
6	1	0	1	0	0	*NA*

Egocentric sampling

	1	2	3	4	5	6
1	*NA*	0	1	0	0	1
2	0	*NA*	1	0	0	1
3	1	0	*NA*	1	0	0
4	0	1	0	*NA*	0	0
5	0	0	1	0	*NA*	1
6	1	0	1	0	0	*NA*

Egocentric sampling

	1	2	3	4	5	6
1	*NA*	0	1	0	0	1
2	**0**	**NA**	**1**	**0**	**0**	**1**
3	1	0	*NA*	1	0	0
4	0	1	0	*NA*	0	0
5	0	0	1	0	*NA*	1
6	1	0	1	0	0	*NA*

Egocentric sampling

	1	2	3	4	5	6
1	*NA*	0	1	0	0	1
2	**0**	**NA**	**1**	**0**	**0**	**1**
3	**1**	**0**	**NA**	**1**	**0**	**0**
4	0	1	0	*NA*	0	0
5	0	0	1	0	*NA*	1
6	1	**0**	**1**	0	0	**NA**

Egocentric sampling

	1	2	3	4	5	6
1	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2	**0**	**NA**	**1**	**0**	**0**	**1**
3	*NA*	**0**	*NA*	*NA*	*NA*	**0**
4	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
5	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
6	*NA*	**0**	**1**	*NA*	*NA*	**NA**

Parameter estimation with missing data

If the data are missing at random, i.e. the value of $\mathbf{o}$, what you get to observe,

- doesn't depend on θ
- doesn't depend on values of $\mathbf{Y}$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$\begin{aligned} l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}] : \theta) \\ &= \sum_{\mathbf{y}[\mathbf{o}^C]} \Pr(\mathbf{Y} = \mathbf{y} : \theta) \end{aligned}$$

Inference based on $l(\theta : \mathbf{y}[\mathbf{o}])$ is provided in `amen`:

- put NA's in place of any non-observed relations.

Parameter estimation with missing data

If the data are missing at random, i.e. the value of $\mathbf{o}$, what you get to observe,

- doesn't depend on θ
- doesn't depend on values of $\mathbf{Y}$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$\begin{aligned} l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}] : \theta) \\ &= \sum_{\mathbf{y}[\mathbf{o}^C]} \Pr(\mathbf{Y} = \mathbf{y} : \theta) \end{aligned}$$

Inference based on $l(\theta : \mathbf{y}[\mathbf{o}])$ is provided in `amen`:

- put NA's in place of any non-observed relations.

Parameter estimation with missing data

If the data are missing at random, i.e. the value of $\mathbf{o}$, what you get to observe,

- doesn't depend on θ
- doesn't depend on values of $\mathbf{Y}$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$\begin{aligned} l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}] : \theta) \\ &= \sum_{\mathbf{y}[\mathbf{o}^C]} \Pr(\mathbf{Y} = \mathbf{y} : \theta) \end{aligned}$$

Inference based on $l(\theta : \mathbf{y}[\mathbf{o}])$ is provided in `amen`:

- put NA's in place of any non-observed relations.

Parameter estimation with missing data

If the data are missing at random, i.e. the value of $\mathbf{o}$, what you get to observe,

- doesn't depend on θ
- doesn't depend on values of $\mathbf{Y}$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$\begin{aligned} l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}] : \theta) \\ &= \sum_{\mathbf{y}[\mathbf{o}^C]} \Pr(\mathbf{Y} = \mathbf{y} : \theta) \end{aligned}$$

Inference based on $l(\theta : \mathbf{y}[\mathbf{o}])$ is provided in `amen`:

- put NA's in place of any non-observed relations.

Parameter estimation with missing data

If the data are missing at random, i.e. the value of $\mathbf{o}$, what you get to observe,

- doesn't depend on θ
- doesn't depend on values of $\mathbf{Y}$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$\begin{aligned} l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}] : \theta) \\ &= \sum_{\mathbf{y}[\mathbf{o}^c]} \Pr(\mathbf{Y} = \mathbf{y} : \theta) \end{aligned}$$

Inference based on $l(\theta : \mathbf{y}[\mathbf{o}])$ is provided in `amen`:

- put NA's in place of any non-observed relations.

Missing at random designs

Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?

Missing at random designs

Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?

Missing at random designs

Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?

Missing at random designs

Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?

Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The "data" include

- $\mathbf{O} = \mathbf{o}$, the determination of which relations you get to see;
- $\mathbf{Y}[\mathbf{O}] = \mathbf{y}[\mathbf{o}]$, the relationship values for the observable relations.

The likelihood is then

$$\begin{aligned} l(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}], \mathbf{O} = \mathbf{o}|\theta) \\ &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]|\theta) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \\ &= l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \end{aligned}$$

If the design part doesn't depend on θ, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.

Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The "data" include

- $\mathbf{O} = \mathbf{o}$, the determination of which relations you get to see;
- $\mathbf{Y}[\mathbf{O}] = \mathbf{y}[\mathbf{o}]$, the relationship values for the observable relations.

The likelihood is then

$$\begin{aligned} l(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}], \mathbf{O} = \mathbf{o}|\theta) \\ &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]|\theta) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \\ &= l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \end{aligned}$$

If the design part doesn't depend on θ, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.

Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The "data" include

- $\mathbf{O} = \mathbf{o}$, the determination of which relations you get to see;
- $\mathbf{Y}[\mathbf{O}] = \mathbf{y}[\mathbf{o}]$, the relationship values for the observable relations.

The likelihood is then

$$\begin{aligned} l(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}], \mathbf{O} = \mathbf{o}|\theta) \\ &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]|\theta) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \\ &= l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \end{aligned}$$

If the design part doesn't depend on θ, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.

Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The "data" include

- $\mathbf{O} = \mathbf{o}$, the determination of which relations you get to see;
- $\mathbf{Y}[\mathbf{O}] = \mathbf{y}[\mathbf{o}]$, the relationship values for the observable relations.

The likelihood is then

$$\begin{aligned} l(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}], \mathbf{O} = \mathbf{o}|\theta) \\ &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]|\theta) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \\ &= l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \end{aligned}$$

If the design part doesn't depend on θ, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.

Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The "data" include

- $\mathbf{O} = \mathbf{o}$, the determination of which relations you get to see;
- $\mathbf{Y}[\mathbf{O}] = \mathbf{y}[\mathbf{o}]$, the relationship values for the observable relations.

The likelihood is then

$$\begin{aligned} l(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}], \mathbf{O} = \mathbf{o}|\theta) \\ &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]|\theta) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \\ &= l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \end{aligned}$$

If the design part doesn't depend on θ, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.

Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The "data" include

- $\mathbf{O} = \mathbf{o}$, the determination of which relations you get to see;
- $\mathbf{Y}[\mathbf{O}] = \mathbf{y}[\mathbf{o}]$, the relationship values for the observable relations.

The likelihood is then

$$\begin{aligned} l(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}], \mathbf{O} = \mathbf{o}|\theta) \\ &= \Pr(\mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]|\theta) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \\ &= l_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o}|\theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}]) \end{aligned}$$

If the design part doesn't depend on θ, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

Ignorable designs

$$I(\theta : \mathbf{o}, \mathbf{y}[\mathbf{o}]) = I_{MAR}(\theta : \mathbf{y}[\mathbf{o}]) \times \Pr(\mathbf{O} = \mathbf{o} | \theta, \mathbf{Y}[\mathbf{o}] = \mathbf{y}[\mathbf{o}])$$

When is the design ignorable?

(MAR) If the probability that $\mathbf{O}$ equals $\mathbf{o}$ doesn't depend on θ or $\mathbf{Y}$ (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that $\mathbf{O}$ equals $\mathbf{o}$

- doesn't depend on θ
- only depends on $\mathbf{Y}$ through $\mathbf{Y}[\mathbf{o}]$.

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.

References

- Thompson and Frank (2000) "Model-based estimation with link-tracing sampling designs"
- Heitjan and Basu (1996) "Distinguishing 'Missing at Random' and 'Missing Completely at Random'

Simulation study - ID likelihoods

$$y_{i,j} = \beta_0 + \beta_r x_{n,i} + \beta_c x_{n,j} + \beta_{d,i,j} + a_i + b_j + \epsilon_{i,j}$$

`fit.pop` = fitted model based on complete network data

`fit.samp` = fitted model based on sampled network data

How do the parameter estimates of `fit.samp` compare to those of `fit.pop`?

Simulation study - ID likelihoods

$$y_{i,j} = \beta_0 + \beta_r x_{n,i} + \beta_c x_{n,j} + \beta_{d,i,j} + a_i + b_j + \epsilon_{i,j}$$

`fit.pop` = fitted model based on complete network data

`fit.samp` = fitted model based on sampled network data

How do the parameter estimates of `fit.samp` compare to those of `fit.pop`?

Simulation study - ID likelihoods

$$y_{i,j} = \beta_0 + \beta_r x_{n,i} + \beta_c x_{n,j} + \beta_{d,i,j} + a_i + b_j + \epsilon_{i,j}$$

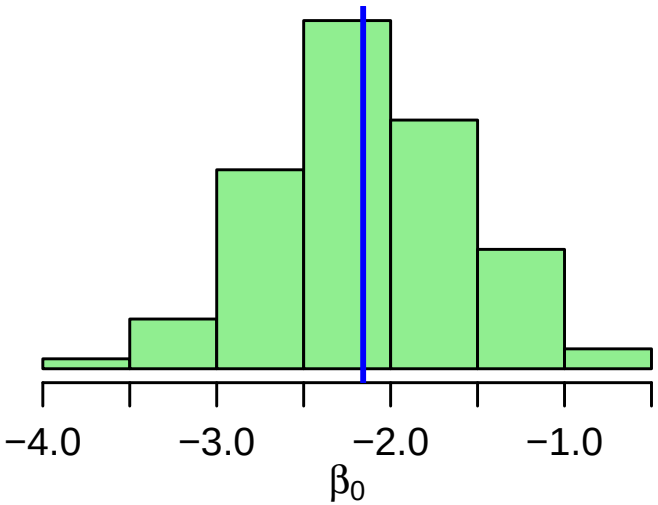
`fit.pop` = fitted model based on complete network data

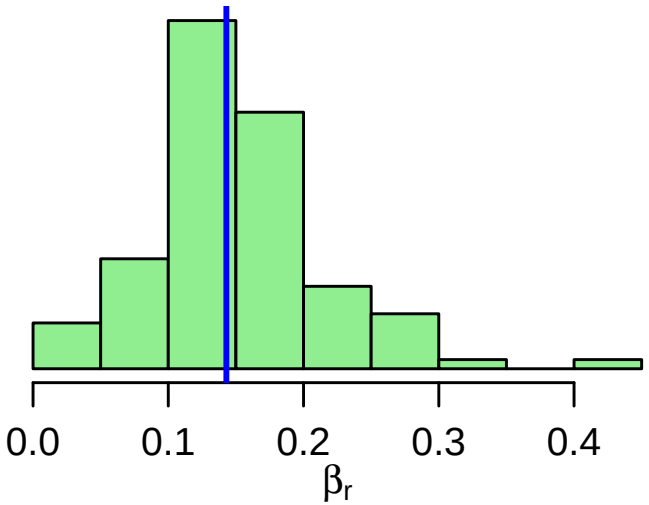
`fit.samp` = fitted model based on sampled network data

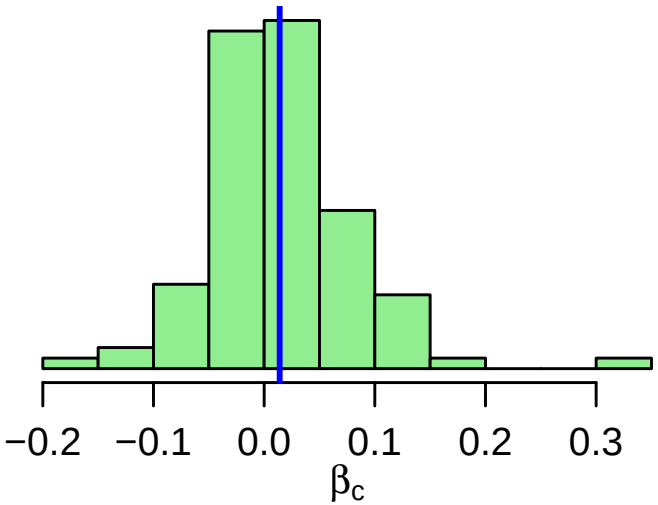
How do the parameter estimates of `fit.samp` compare to those of `fit.pop`?

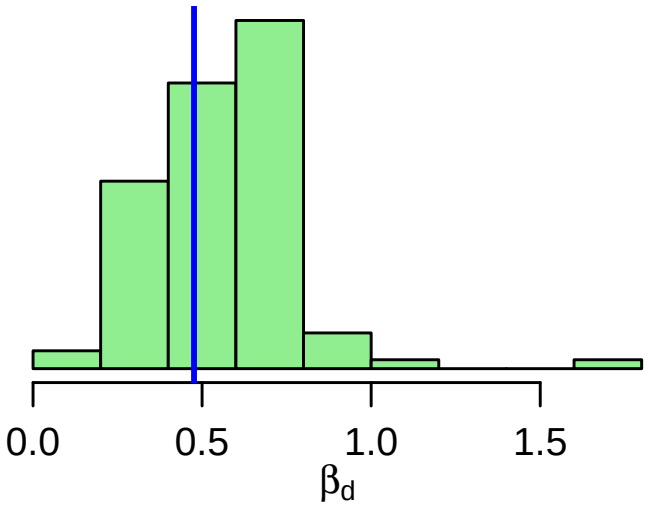
Node-induced subgraph sample

$n_p = 32,\ n_s = 10$









Egocentric sample

$n_p = 32,\ n_{s_1} = 4$