Statistical Analysis of Multivariate Egocentric Networks

Termeh Shafie
GESIS - Leibniz Institute for the Social Sciences

mrs.schochastics.net @termehshafie @termehs

multigraph representation of network data

multivariate network data represented as multigraphs:

“graphs where multiple edges and self-edges are permitted”

multigraph representation of network data

multivariate network data represented as multigraphs:
"graphs where multiple edges and self-edges are permitted"

methodological developments and software:

- [x] two random multigraph models
- [x] definition of several statistics to analyse structural features
- [x] formal goodness of fit tests
- [x] R package:
  https://cran.r-project.org/package=multigraphr

multigraph representation of network data

multivariate network data represented as multigraphs:
“graphs where multiple edges and self-edges are permitted”

methodological developments and software:

- ☑ two random multigraph models
- ☑ definition of several statistics to analyse structural features
- ☑ formal goodness of fit tests
- ☑ R package:
  https://cran.r-project.org/package=multigraphr

limitation: currently only implemented for exact distributions of multigraphs

possible solutions:

- ☑ MC methods for approximating probability distributions
- ☑ focus on small networks such as ego nets

multigraph representation of network data

multivariate network data represented as multigraphs:
“graphs where multiple edges and self-edges are permitted”

methodological developments and software:

- ☑ two random multigraph models
- ☑ definition of several statistics to analyse structural features
- ☑ formal goodness of fit tests
- ☑ R package:
  https://cran.r-project.org/package=multigraphr

limitation: currently only implemented for exact distributions of multigraphs

possible solutions:

- ☑ MC methods for approximating probability distributions
- ☑ focus on small networks such as ego nets

running example

Krackhardt's High-tech Managers Networks (1987)

cognitive social structure data from 21 management personnel in a high-tech firm

relations:	actor attributes:
- undirected friendship - directed advice	- department - level - age - tenure

(also includes the relations each ego perceived among all other managers)

running example

Krackhardt's High-tech Managers Networks (1987)

cognitive social structure data from 21 management personnel in a high-tech firm

relations:	actor attributes:
- undirected friendship - directed advice	- department - level - age - tenure

(also includes the relations each ego perceived among all other managers)

running example

Krackhardt's High-tech Managers Networks (1987)

cognitive social structure data from 21 management personnel in a high-tech firm

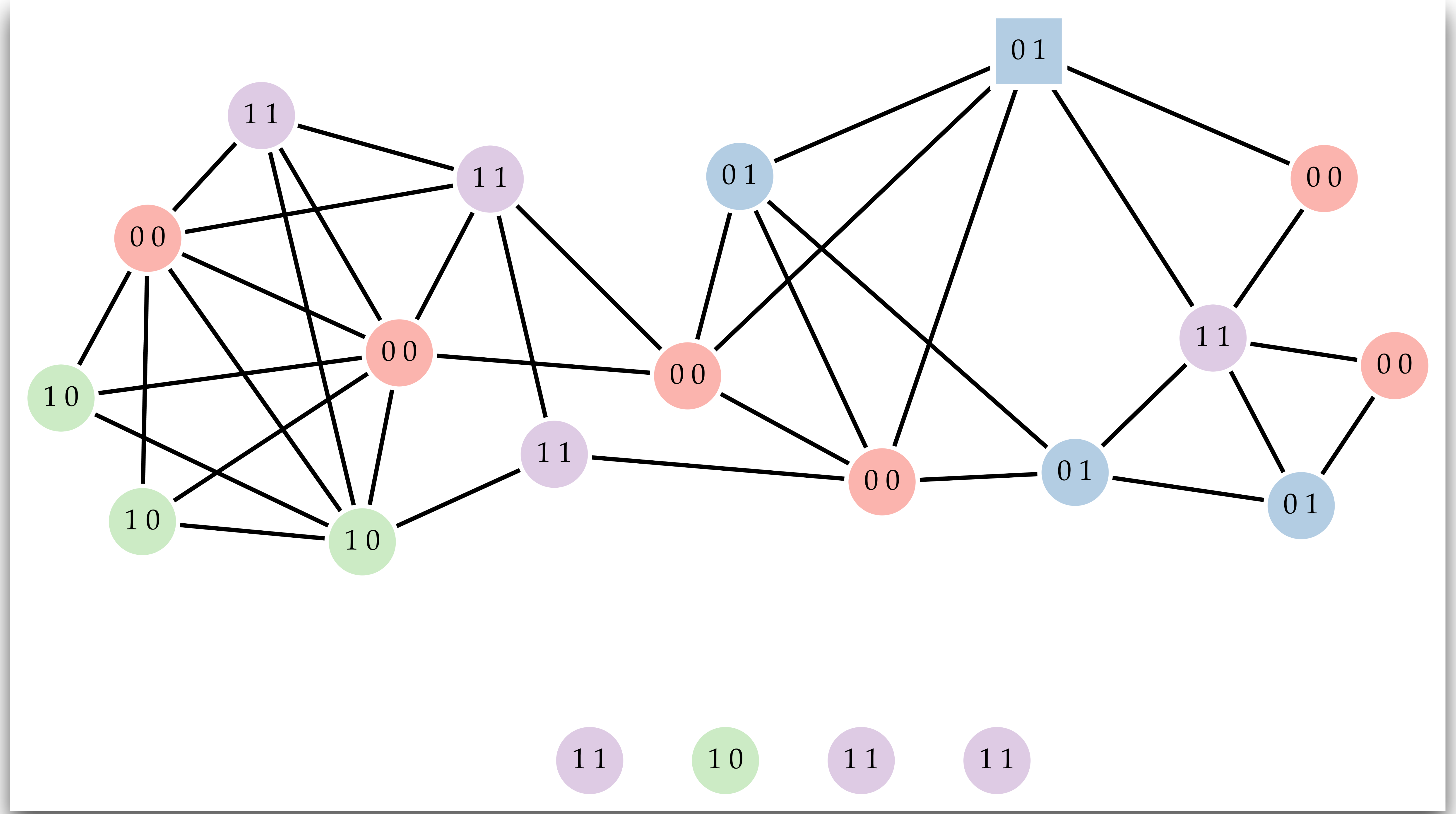
relations:	actor attributes:
- undirected friendship - directed advice	- department - level - age - tenure

(also includes the relations each ego perceived among all other managers)

- ☑ age and tenure binarized to indicate low/high (0/1)
- ☑ each node thus has 4 possible cross-classified attribute outcomes: (0,0), (0,1), (1,0), (1,1)
- ☑ multigraphs aggregated based on these four possible outcomes represented as nodes

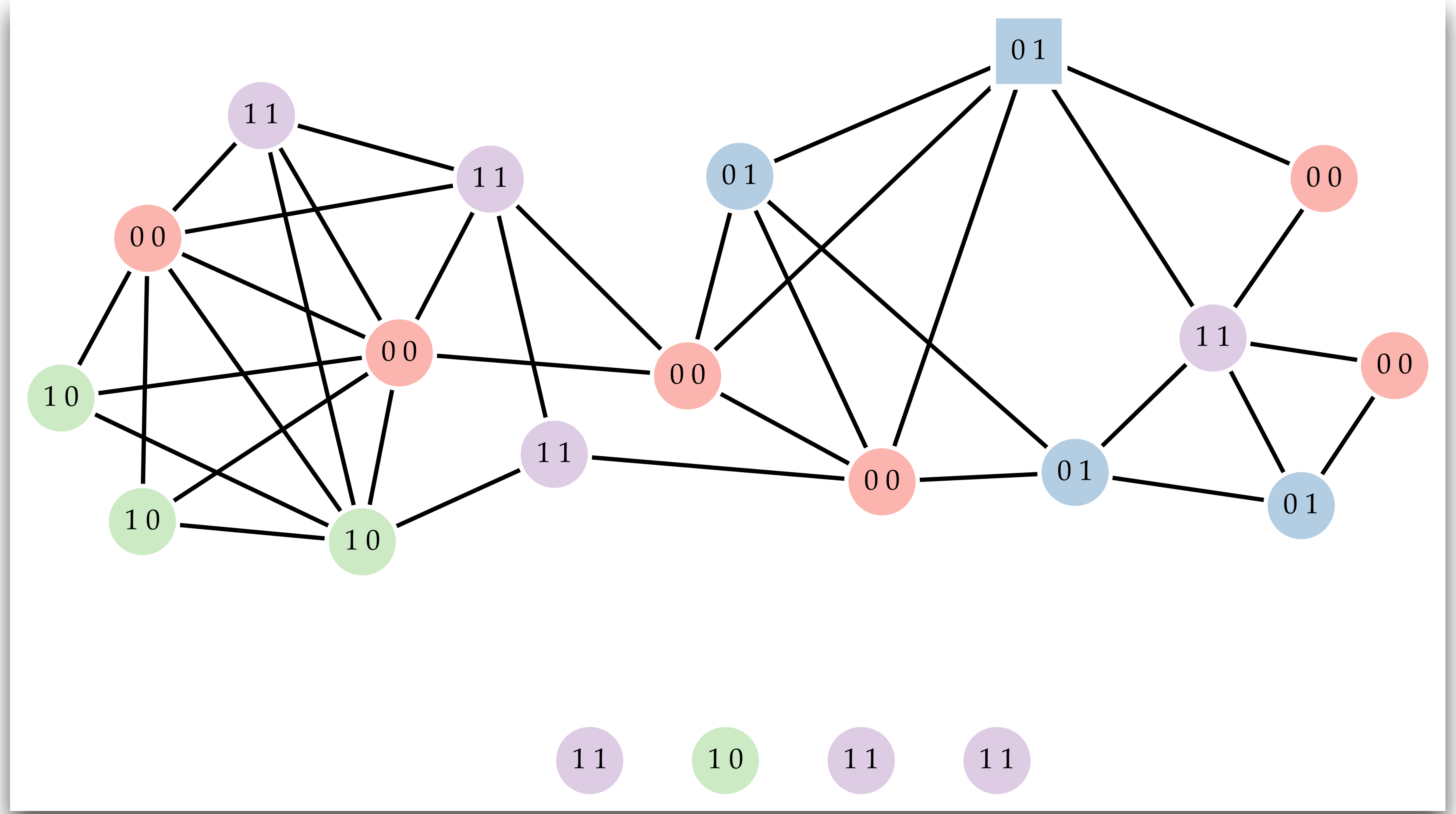
example: aggregated multigraphs

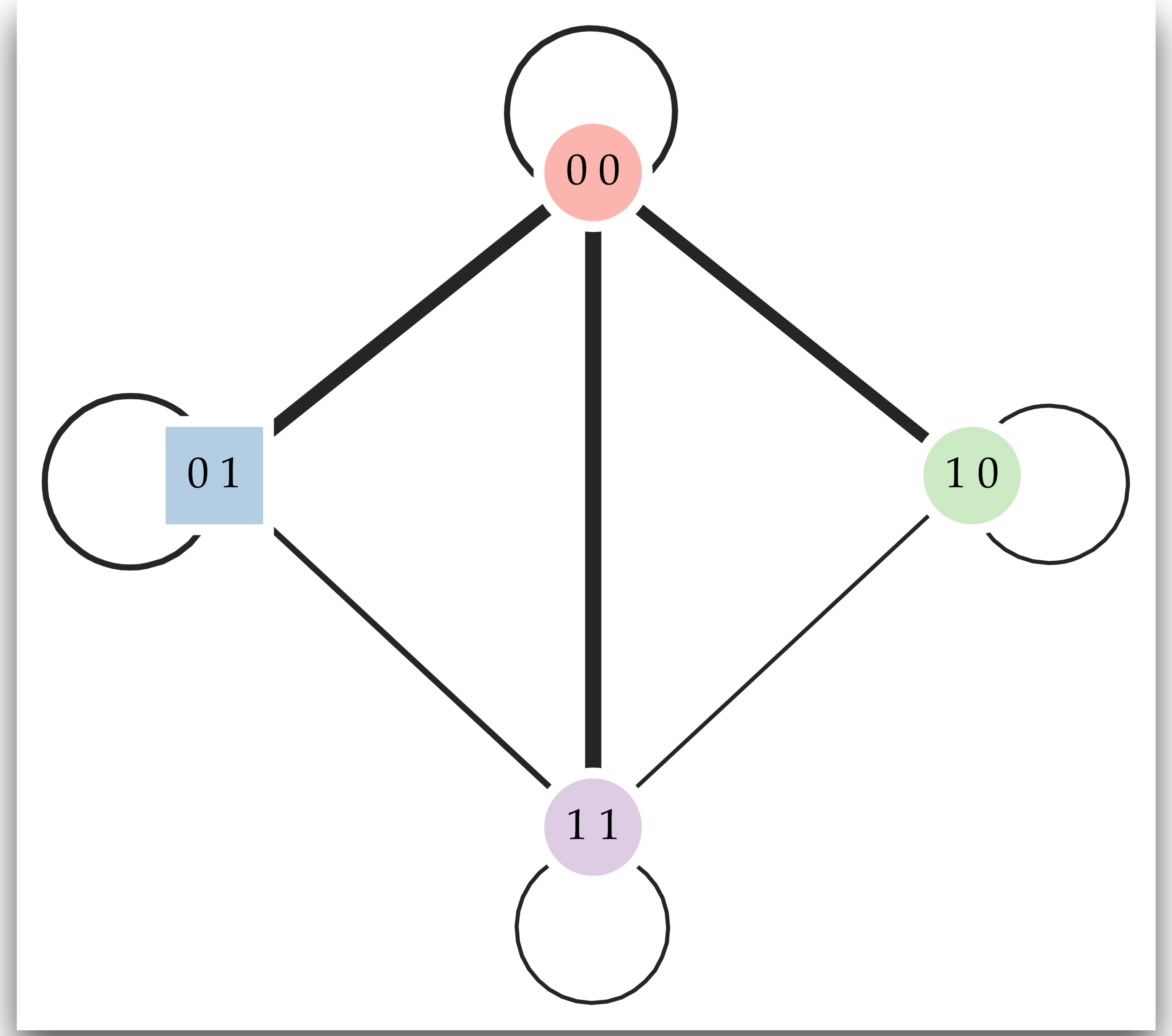
ego 1's original network and aggregated multigraph



example: aggregated multigraphs

ego 1's original network and aggregated multigraph





aggregated to

example: aggregated multigraphs

random multigraph models

- ☑ multigraphs represented by their edge multiplicity sequence

$$\mathbf{M} = (M_{ij} : (i,j) \in R)$$

where R is the canonical site space for undirected edges

$$R = \{(i,j) : 1 \leq i \leq j \leq n\}$$

that is

$$(1,1) < (1,2) < \cdots < (1,n) < (2,2) < (2,3) \cdots < (2,n) < \cdots < (n,n)$$

and n is number of nodes

random multigraph models

- multigraphs represented by their edge multiplicity sequence

$$\mathbf{M} = (M_{ij} : (i,j) \in R)$$

where R is the canonical site space for undirected edges

$$R = \{(i,j) : 1 \leq i \leq j \leq n\}$$

that is

$$(1,1) < (1,2) < \cdots < (1,n) < (2,2) < (2,3)\cdots < (2,n) < \cdots < (n,n)$$

and n is number of nodes

- so for our examples with multigraphs on 4 nodes the number of edge sites is 10:

$$(1,1), (1,2), (1,3), (1,4), (2,2), (2,3), (2,4), (3,3), (3,4), (4,4)$$

random multigraph models

$\mathbf{M} \sim \mathsf{RSM}(\mathbf{d})$

Random Stub Matching

given observed degree sequence $\mathbf{d}$

$\mathbf{M} \sim \mathsf{IEAS}(\mathbf{Q}(\mathbf{d}))$

Independent Edge Assignments of Stubs

edge assignment probabilities $\mathbf{Q}(\mathbf{d})$

where $\mathbf{d}$ is observed degree sequence

$\mathbf{M} \sim \mathsf{ISA}(\mathbf{Q}(\mathbf{p}))$

Independent Stub Assignments

degree sequence $\sim$ multinomial($2m$, $\mathbf{p}$)

where $\mathbf{p}$ are stub assignment probabilities

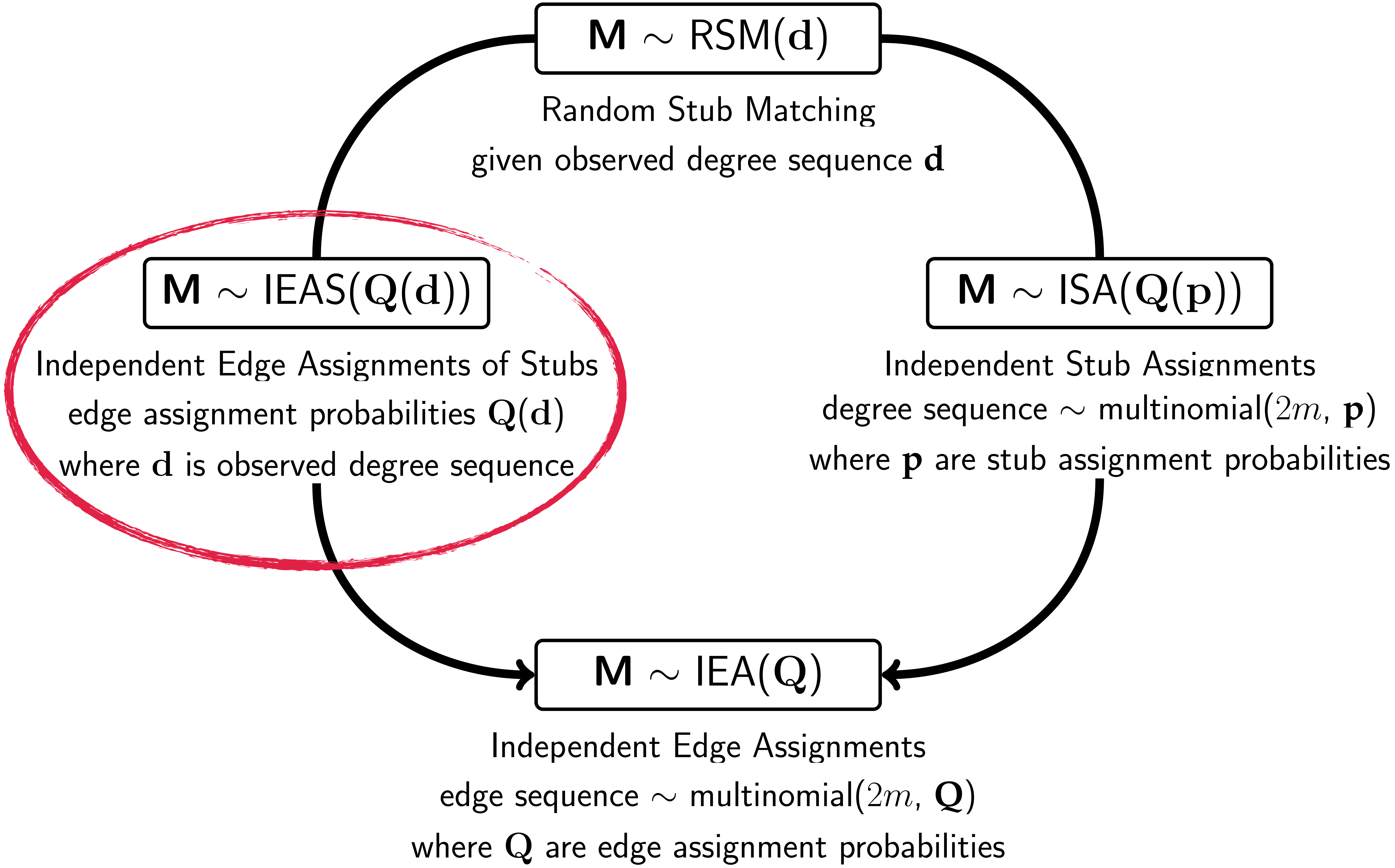
$\mathbf{M} \sim \mathsf{IEA}(\mathbf{Q})$

Independent Edge Assignments

edge sequence $\sim$ multinomial($2m$, $\mathbf{Q}$)

where $\mathbf{Q}$ are edge assignment probabilities

random multigraph models



random multigraph models: statistics

statistics for analysing structural features under multigraph models

measures defined using the distribution of edge multiplicities:

- ☑ number of loops and non-loops: tendency for within and between vertex category edges
  ➞ homophily/heterophily
- ☑ tendency for isolated vertices ➞ network diffusion
- ☑ simple occupancy of edges ➞ simple/complex network*
- ☑ single ties within vertex category ➞ isolation
- ☑ tendency for strengthening ties and if overlapping for multiple edge types ➞ multiplexity

* *"if a graph contains loops and/or any pairs of nodes is adjacent via more than one line a graph is complex"* [Wasserman and Faust, 1994]

random multigraph models: statistics

statistics for analysing structural features under multigraph models

measures defined using the distribution of edge multiplicities:

- ☑ number of loops and non-loops: tendency for within and between vertex category edges
  ⟶ homophily/heterophily
- ☑ tendency for isolated vertices ⟶ network diffusion
- ☑ simple occupancy of edges ⟶ simple/complex network*
- ☑ single ties within vertex category ⟶ isolation
- ☑ tendency for strengthening ties and if overlapping for multiple edge types ⟶ multiplexity

* *"if a graph contains loops and/or any pairs of nodes is adjacent via more than one line a graph is complex"* [Wasserman and Faust, 1994]

random multigraph models: statistics

statistics for analysing structural features under multigraph models

measures defined using the distribution of edge multiplicities:

☑ number of loops and non-loops: tendency for within and between vertex category edges
→ homophily/heterophily

approx 95% intervals
$\hat{E} \pm 2\sqrt{\hat{V}}$

☑ tendency for isolated vertices → network diffusion

☑ simple occupancy of edges → simple/complex network*

☑ single ties within vertex category → isolation

☑ tendency for strengthening ties and if overlapping for multiple edge types → multiplexity

* *"if a graph contains loops and/or any pairs of nodes is adjacent via more than one line a graph is complex"* [Wasserman and Faust, 1994]

random multigraph models: goodness of fit

gof measures between observed and expected edge multiplicity sequence
under simple or composite hypothesis

test statistics:

- ☑ S of Pearson type
- ☑ A of information divergence type

summary:

- ☑ even for very small *m*, the null distributions of the test statistics under the IEA model are well approximated by their asymptotic distributions
- ☑ the convergence of the cdf's of test statistics are rapid and depend on parameters in models
- ☑ approximations can be obtained using adjustments of χ^2-distributions yielding better power
- ☑ influence of RSM on both test statistics is substantial for small *m*: a shift of their distributions towards smaller values compared to what holds true for null distributions under IEA

random multigraph models: goodness of fit

gof measures between observed and expected edge multiplicity sequence
under simple or composite hypothesis

test statistics:

- ☑ S of Pearson type
- ☑ A of information divergence type

summary:

- ☑ even for very small *m*, the null distributions of the test statistics under the IEA model are well approximated by their asymptotic distributions
- ☑ the convergence of the cdf's of test statistics are rapid and depend on parameters in models
- ☑ approximations can be obtained using adjustments of χ^2-distributions yielding better power
- ☑ influence of RSM on both test statistics is substantial for small *m*: a shift of their distributions towards smaller values compared to what holds true for null distributions under IEA

random multigraph models: goodness of fit

gof measures between observed and expected edge multiplicity sequence
under simple or composite hypothesis

test statistics:

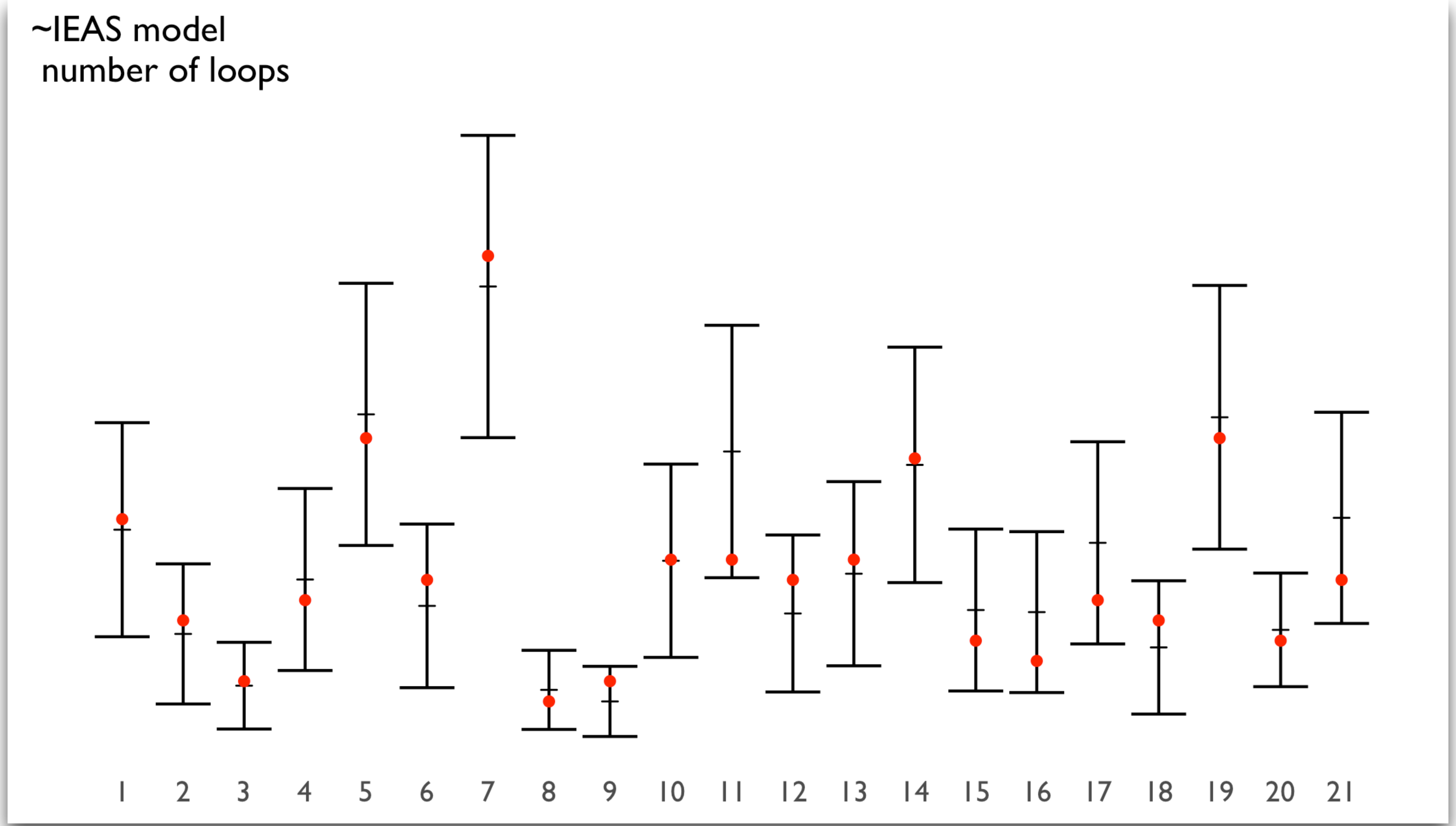
- ☑ S of Pearson type
- ☑ A of information divergence type

significance level
$\alpha = 0.05$

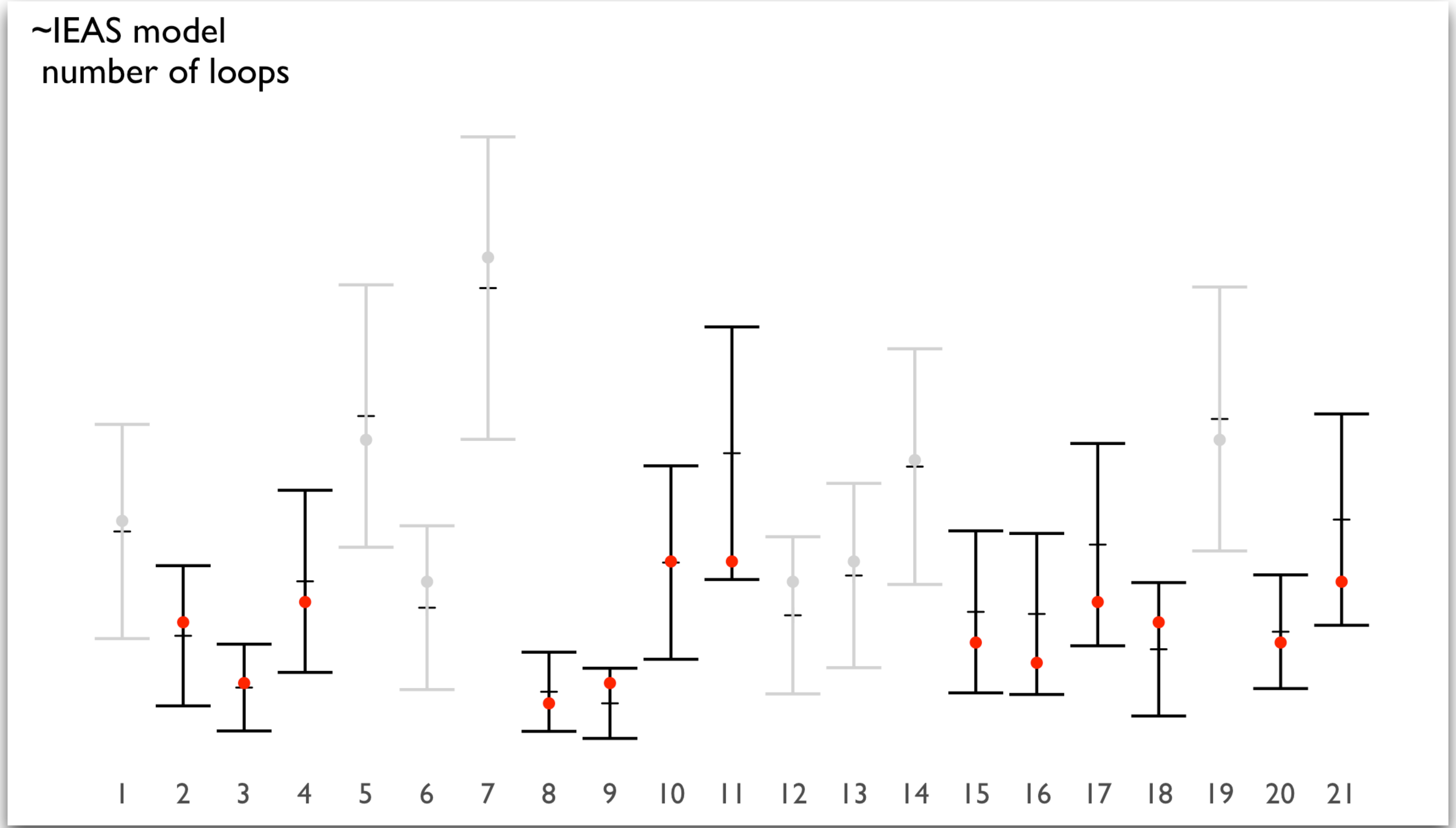
summary:

- ☑ even for very small *m*, the null distributions of the test statistics under the IEA model are well approximated by their asymptotic distributions
- ☑ the convergence of the cdf's of test statistics are rapid and depend on parameters in models
- ☑ approximations can be obtained using adjustments of χ^2-distributions yielding better power
- ☑ influence of RSM on both test statistics is substantial for small *m*: a shift of their distributions towards smaller values compared to what holds true for null distributions under IEA

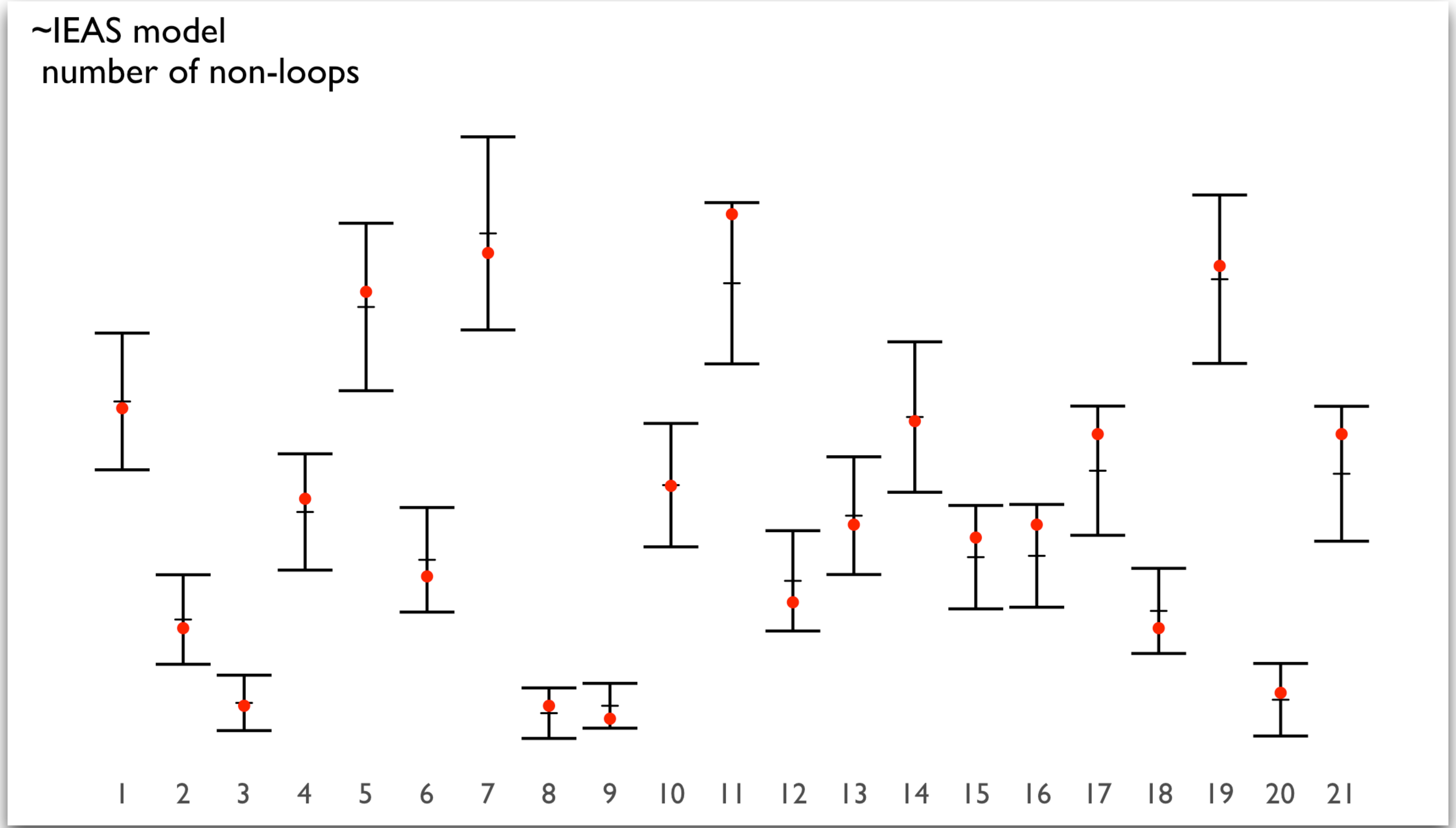
example: number of loops



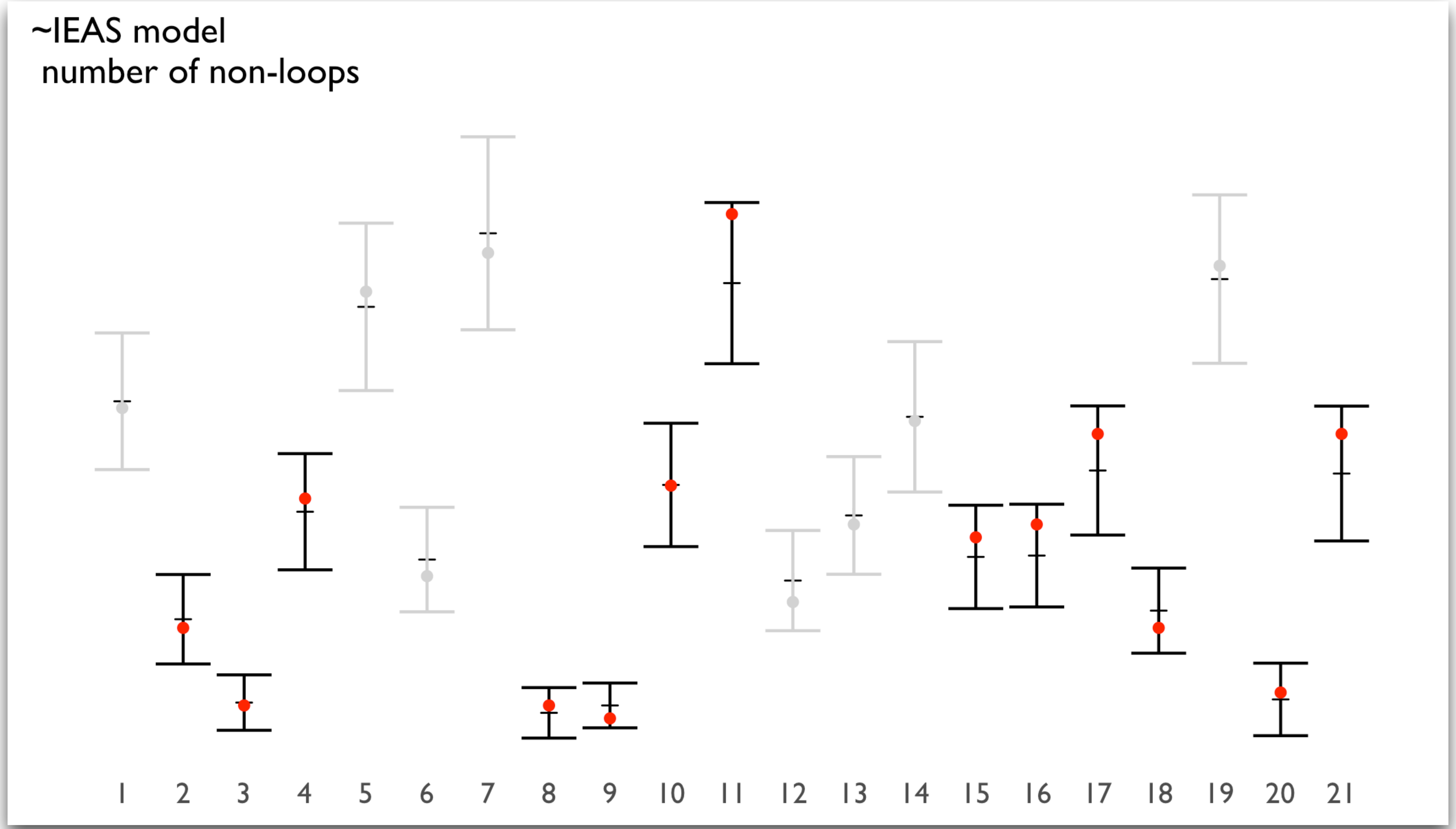
example: goodness of fit



example: number of non-loops



example: goodness of fit



continuation of work

- ☑ analyze the co-occurrences of multiple relation types (undirected/directed/signed)
- ☑ define statistics for quantifying structural holes in multigraphs
- ☑ scale it up in terms of number of ego-nets:
    - ✓ using digital trace data e.g. social media accounts
    - ✓ how to include perceived alter-alter ties?
- ☑ community detection based on egos’ structural features

references

- Shafie, T. (2015). A multigraph approach to social network analysis. *Journal of Social Structure, 16.*
- Shafie, T. (2016). Analyzing local and global properties of multigraphs. *The Journal of Mathematical Sociology*, 40(4), 239–264.
- Frank, O. and Shafie, T. (2018). Random multigraphs and aggregated triads with fixed degrees. *Network Science*, 6(2), 232–250.
- Shafie, T. and Schoch, D., (2021). Multiplexity analysis of networks using multigraph representations. *Statistical Methods & Applications,* 30(5), 1425–1444.
- Shafie, T. (2022). Goodness of fit tests for random multigraph models. *Journal of Applied Statistics.* https://doi.org/10.1080/02664763.2022.2099816
- Shafie, T. (2021). *multigraphr: Probability models and statistical analysis of random multigraphs.* Available at https://cran.r-project.org/package=multigraphr, R package version 0.1.0.