Analyzing Social Structure using Multigraph Representations

Termeh Shafie
Department of Computational Social Science
GESIS - Leibniz Institute for the Social Sciences

2 Nov 2022

mrs.schochastics.net
@termehshafie
@termehs

the theoretical background

xkcd.com

✓ Shafie, T. (2015). A multigraph approach to social network analysis. *Journal of Social Structure*, 16, 1-21.

✓ Shafie, T. (2016). Analyzing local and global properties of multigraphs. *The Journal of Mathematical Sociology*, 40(4), 239-264.

✓ Frank, O., Shafie, T., (2018). Random Multigraphs and Aggregated Triads with Fixed Degrees. *Network Science*, 6(2), 232-250.

✓ Shafie, T., Schoch, D. (2021) Multiplexity analysis of networks using multigraph representations. *Statistical Methods & Applications* 30, 1425–1444.

✓ Shafie, T. (2022). Goodness of fit tests for random multigraph models, *Journal of Applied Statistics*. 1-26

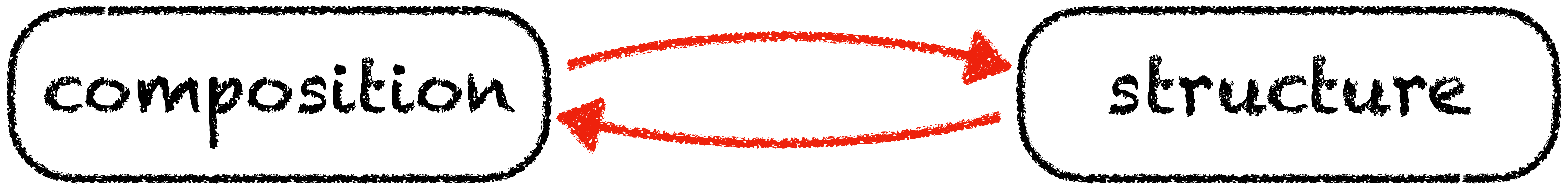
multigraphr

R package: https://cran.r-project.org/package=multigraphr

multivariate networks

multivariate networks comprise

- ☑ vertex set with at least one type of edge between pairs of nodes
- ☑ numerical and/or qualitative attributes on the vertices and edges



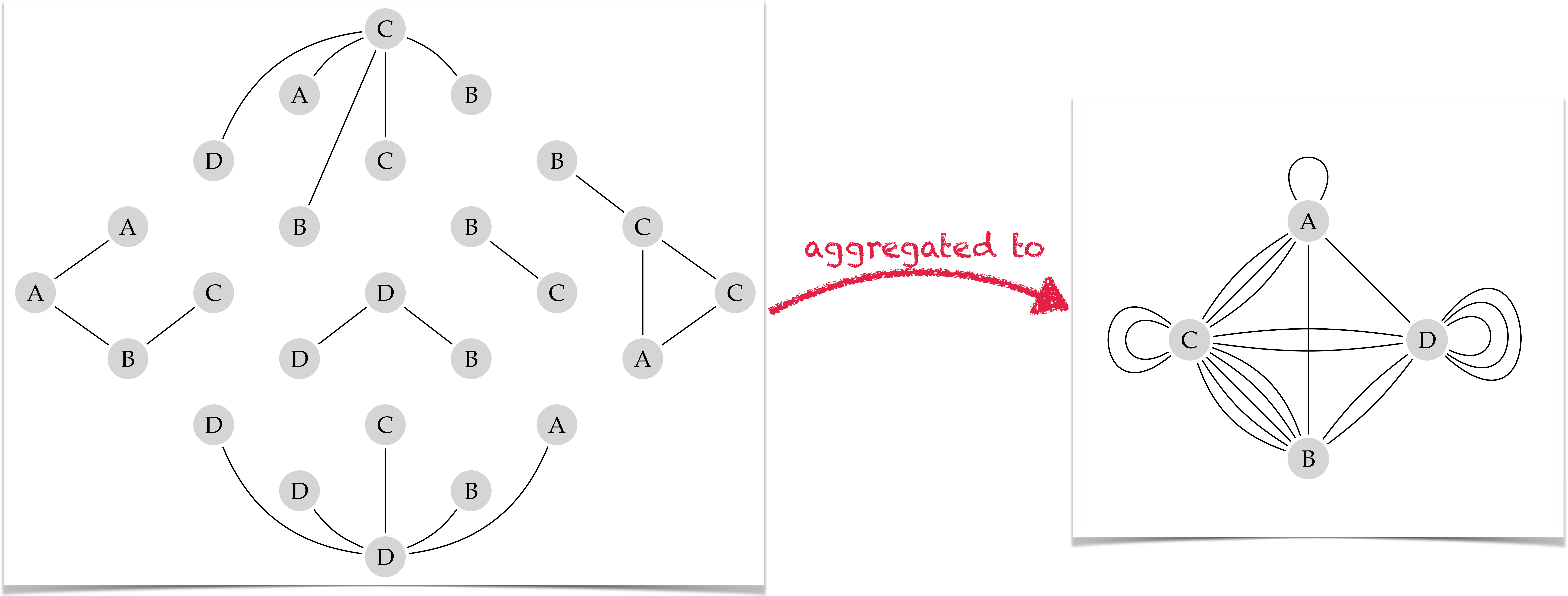
multivariate network data represented as **multigraphs**:

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

- ☑ can appear directly in applications (although scarce)
- ☑ can be constructed by different kinds of aggregations in graphs
  - ✓ node aggregation based on node attributes
  - ✓ tie aggregation based on tie attributes

aggregated multigraphs

example:



informative statistics in multigraphs

statistics for analyzing local and global social structural features

- ☑ 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

how do we quantify these statistics?

* *"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]

multigraph representation of network data

- multigraph represented by their edge multiplicity sequence

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

where R is the canonical site space for undirected edges $R = \{(i,j) : 1 \leq i \leq j \leq n\}$

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

- the number of vertex pair sites is given by

$$r = \binom{n+1}{2}$$

- edge multiplicities as entries in a matrix

$$\mathbf{M} = \begin{bmatrix} M_{11} & M_{12} & \ldots & M_{1n} \\ 0 & M_{22} & \ldots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_{nn} \end{bmatrix} \quad \mathbf{M} + \mathbf{M}' = \begin{bmatrix} 2M_{11} & M_{12} & \ldots & M_{1n} \\ M_{12} & 2M_{22} & \ldots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1n} & M_{2n} & \ldots & 2M_{nn} \end{bmatrix}$$

multigraph representation of network data

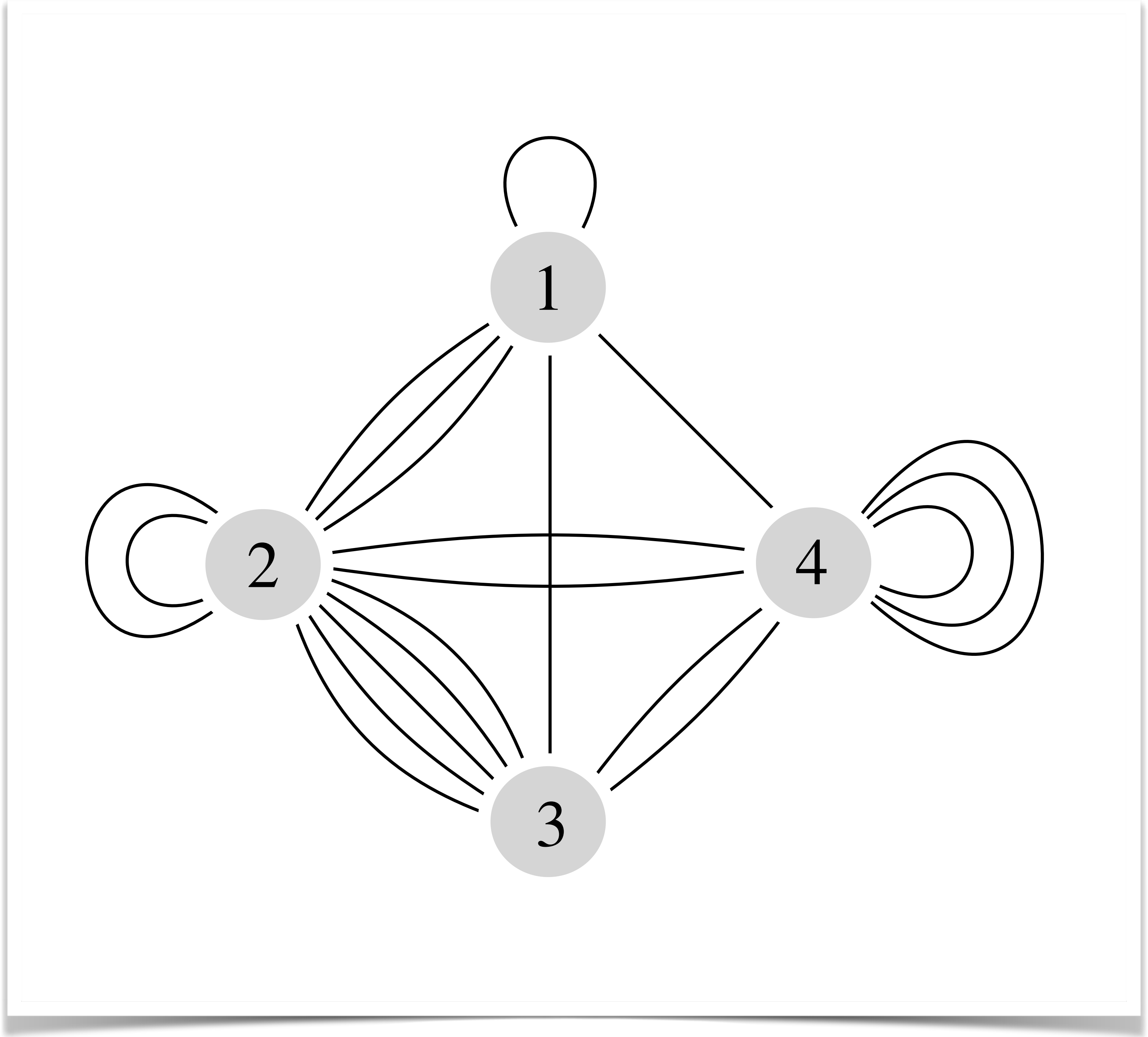
example:

- ☑ the number of vertex pair sites

$$r = \binom{n+1}{2} = \frac{5 \times 4}{2} = 10$$

- ☑ edge multiplicity sequence

$$\mathbf{M} = (M_{11}, M_{12}, M_{13}, M_{14}, M_{22}, M_{23}, M_{24}, M_{33}, M_{34}, M_{44})$$
$$= (1, \quad 3, \quad 1, \quad 1, \quad 2, \quad 5, \quad 2, \quad 0, \quad 2, \quad 3)$$



- ☑ edge multiplicities as entries in a matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 3 & 1 & 1 \\ 0 & 2 & 5 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix} \qquad \mathbf{M} + \mathbf{M}' = \begin{bmatrix} 2 & 3 & 1 & 1 \\ 3 & 4 & 5 & 2 \\ 1 & 5 & 0 & 2 \\ 1 & 2 & 2 & 6 \end{bmatrix}$$

statistics under random multigraph models

quantified defined using the distribution of edge multiplicities

- ☑ number of loops M_1 and number of non-loops M_2
- ☑ complexity sequence $\mathbf{R} = (R_0, R_1, \ldots, R_k)$ where

$$R_k = \sum \sum_{i \leq j} I(M_{ij} = k) \quad \text{for } k = 0,1,\ldots,m$$

is the frequencies of edge multiplicities

✓ M_1 and M_2

- tendency for within and between vertex category edges (homophily/heterophily)

✓ M_2 and R_2

- simplicity statistics
- single ties within vertex category (isolation)

✓ R_0 and R_1

- R_0: tendency for isolated vertices (network diffusion)
- R_1: simple occupancy of edges

✓ $R_0 + R_1$ compared to $R_3 + \cdots + R_k$

- tendency for strengthening ties (multiplexity)

✓ M_1 and R_1

- single ties within vertex category (isolation)

✓ interval estimates for R_k

- if overlapping for multiple edge types ⇒ multiplexity

random multigraph models

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

Random Stub Matching

given observed degree sequence $\mathbf{d}$

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

Independent Edge Assignments of Stubs

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

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

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

Independent Stub Assignments

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

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

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

Independent Edge Assignments

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

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

random multigraph models

random stub matching (RSM)

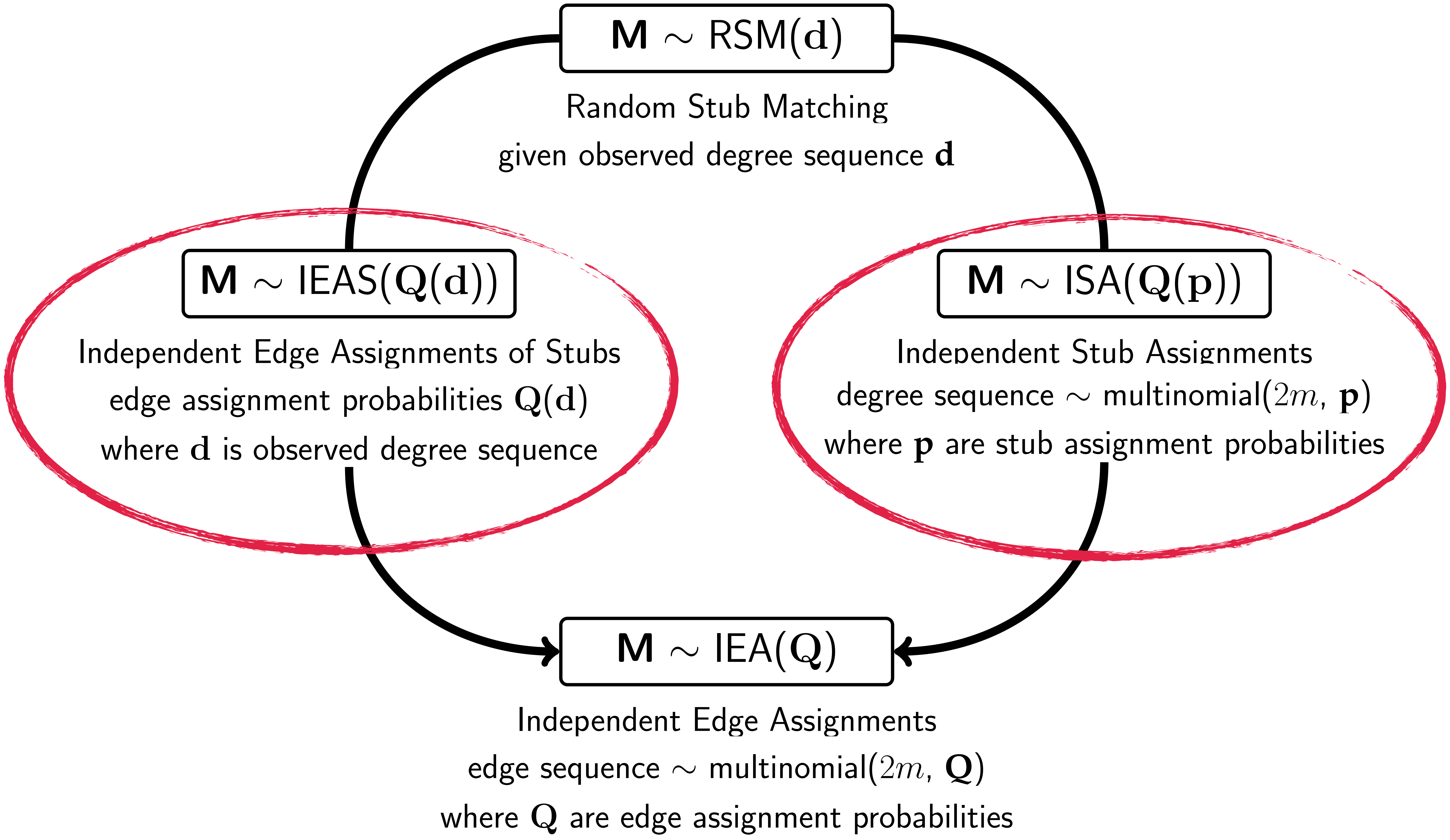
- ☑ edges are assigned to sites given fixed degree sequence $\mathbf{d} = (d_1, \ldots, d_n)$
- ☑ probability that an edge is assigned to site $(i,j) \in R$

$$Q_{ij} = \begin{cases} \binom{d_i}{2} \Big/ \binom{2m}{2} & \text{for } i = j \\ d_i d_j \Big/ \binom{2m}{2} & \text{for } i < j \end{cases}$$

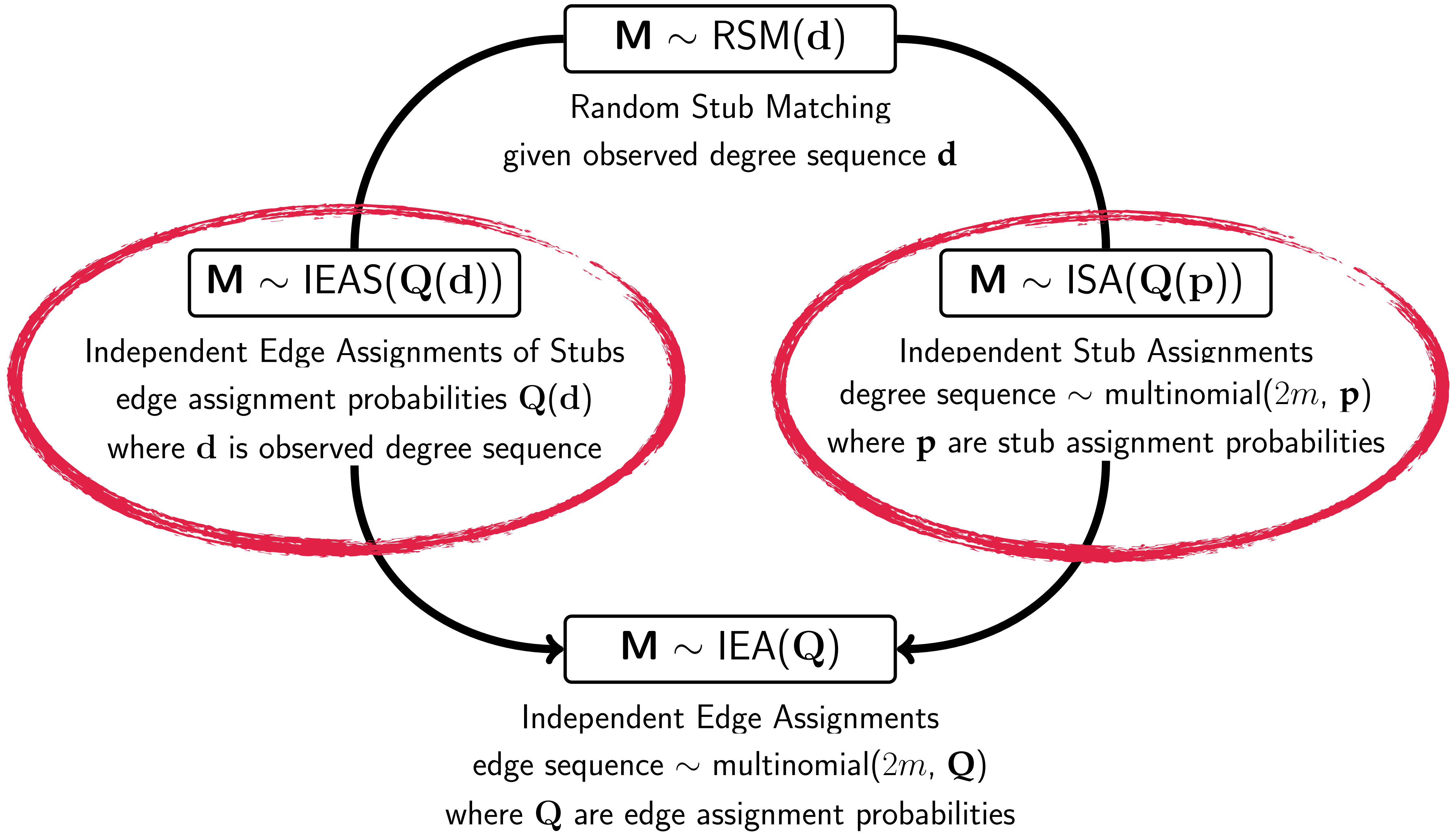
independent edge assignments (IEA)

- ☑ edges are independently assigned to vertex pairs in site space R
- ☑ edge assignment probabilities $\mathbf{Q} = (Q_{ij} : (i,j) \in R)$
- ☑ $\mathbf{M}$ is multinomial distributed with parameters m and $\mathbf{Q}$
- ☑ moments of statistics for analysing local and global structure are easily derived
- ☑ can be used as an approximation to the RSM model

random multigraph models



approximate IEA models



independent edge assignment of stubs (IEAS)

- ☑ edges assignment probabilities defined by observed degree sequence $\mathbf{Q} = \mathbf{Q}(\mathbf{d})$

independent stub assignment (ISA)

- ☑ Bayesian model for stub frequencies
- ☑ degree sequence $\mathbf{D} \sim$ multinomial$(2m, \mathbf{p})$ where $\mathbf{p}$ are stub assignment probabilities

statistics under random multigraph models

✓ M_1 and M_2

- tendency for within and between vertex category edges (homophily/heterophily)

✓ M_2 and R_2

- simplicity statistics
- single ties within vertex category (isolation)

✓ R_0 and R_1

- R_0: tendency for isolated vertices (network diffusion)
- R_1: simple occupancy of edges

✓ $R_0 + R_1$ compared to $R_3 + \cdots + R_k$

- tendency for strengthening ties (multiplexity)

✓ M_1 and R_1

- single ties within vertex category (isolation)

✓ interval estimates for R_k

- if overlapping for multiple edge types ⇒ multiplexity

moments of these statistics can be derived under IEA but not under RSM

$\Longrightarrow$ to avoid computational difficulties we can to use the IEA approximations

approx 95% intervals

$$\hat{E} \pm 2\sqrt{\hat{V}}$$

goodness of fit tests

gof measures between observed and expected edge multiplicity sequence

test statistics:

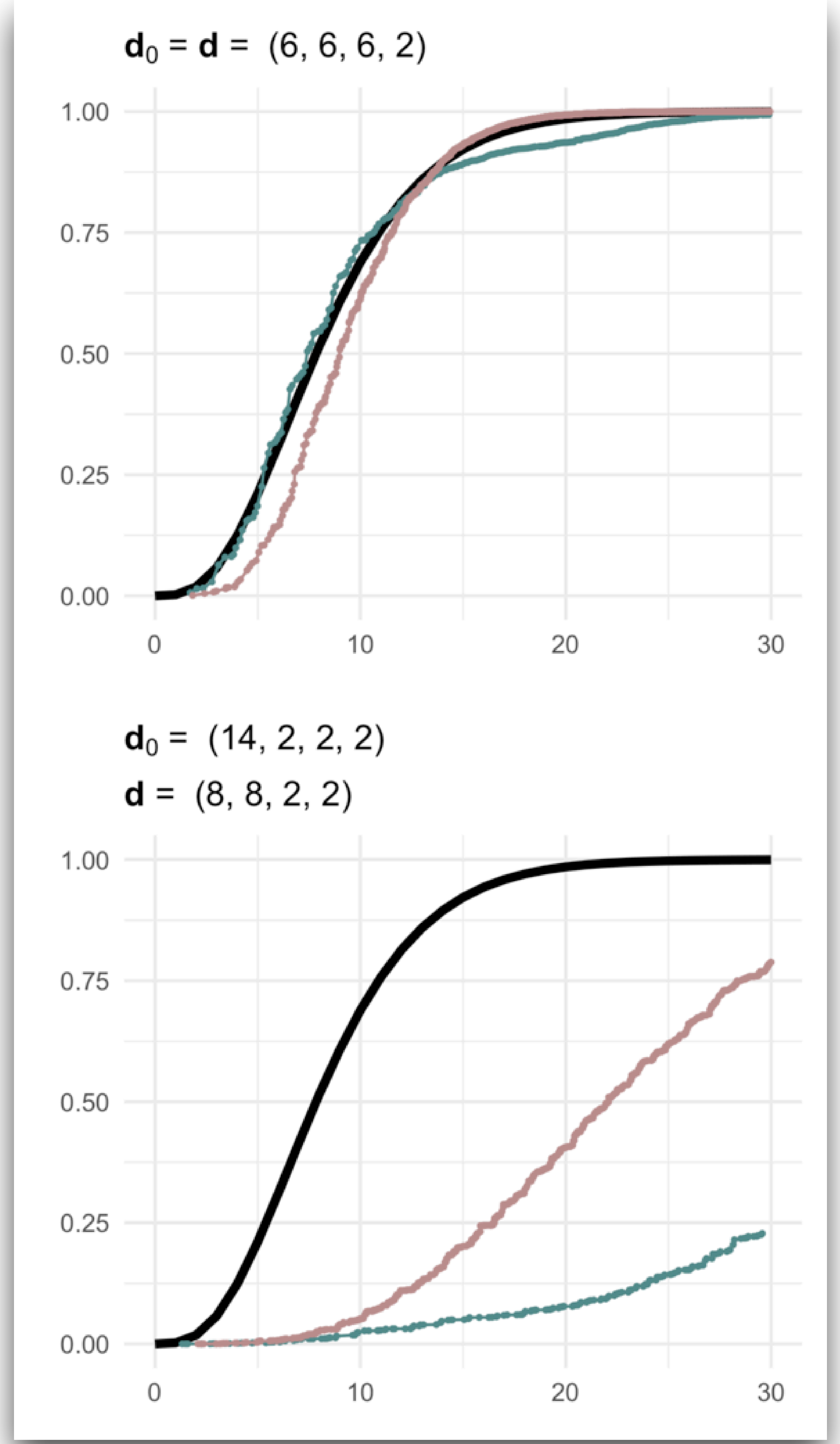
☑ S of Pearson type

☑ A of information divergence type

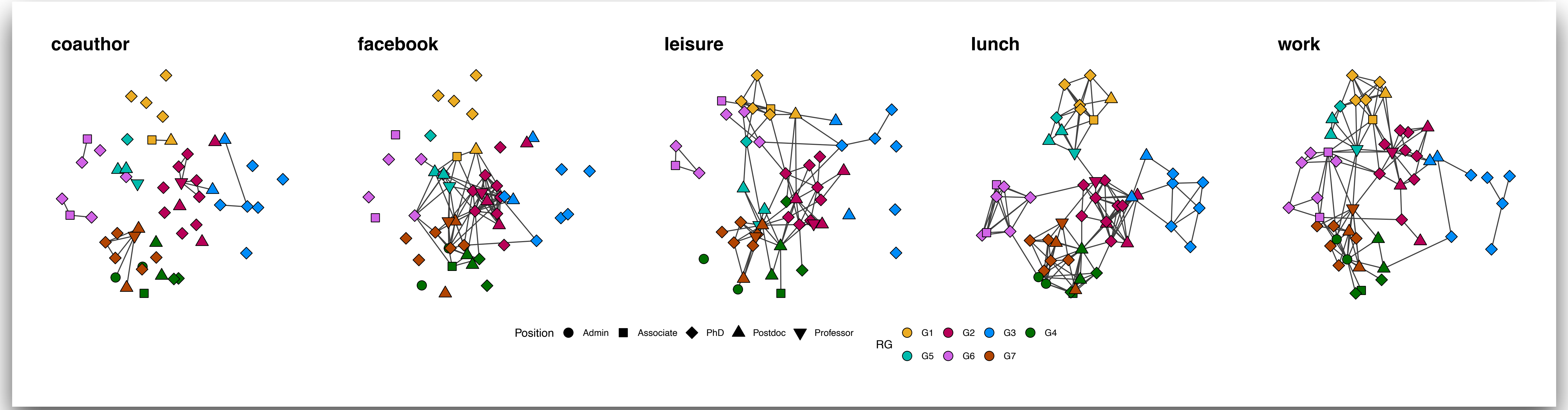
some results:

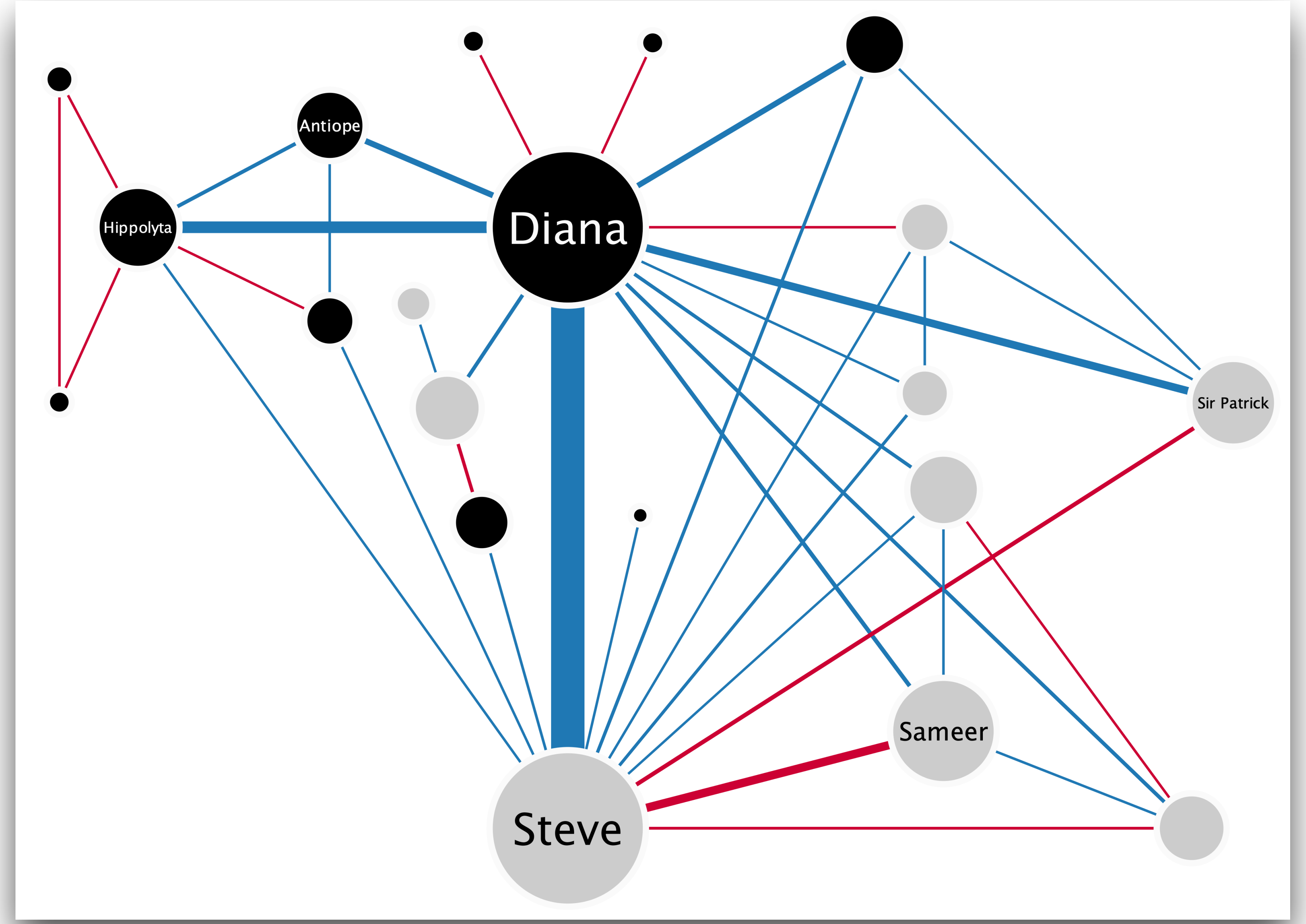
☑ even for very small *m*, null distributions of test statistics under IEA model are well approximated by asymptotic distributions

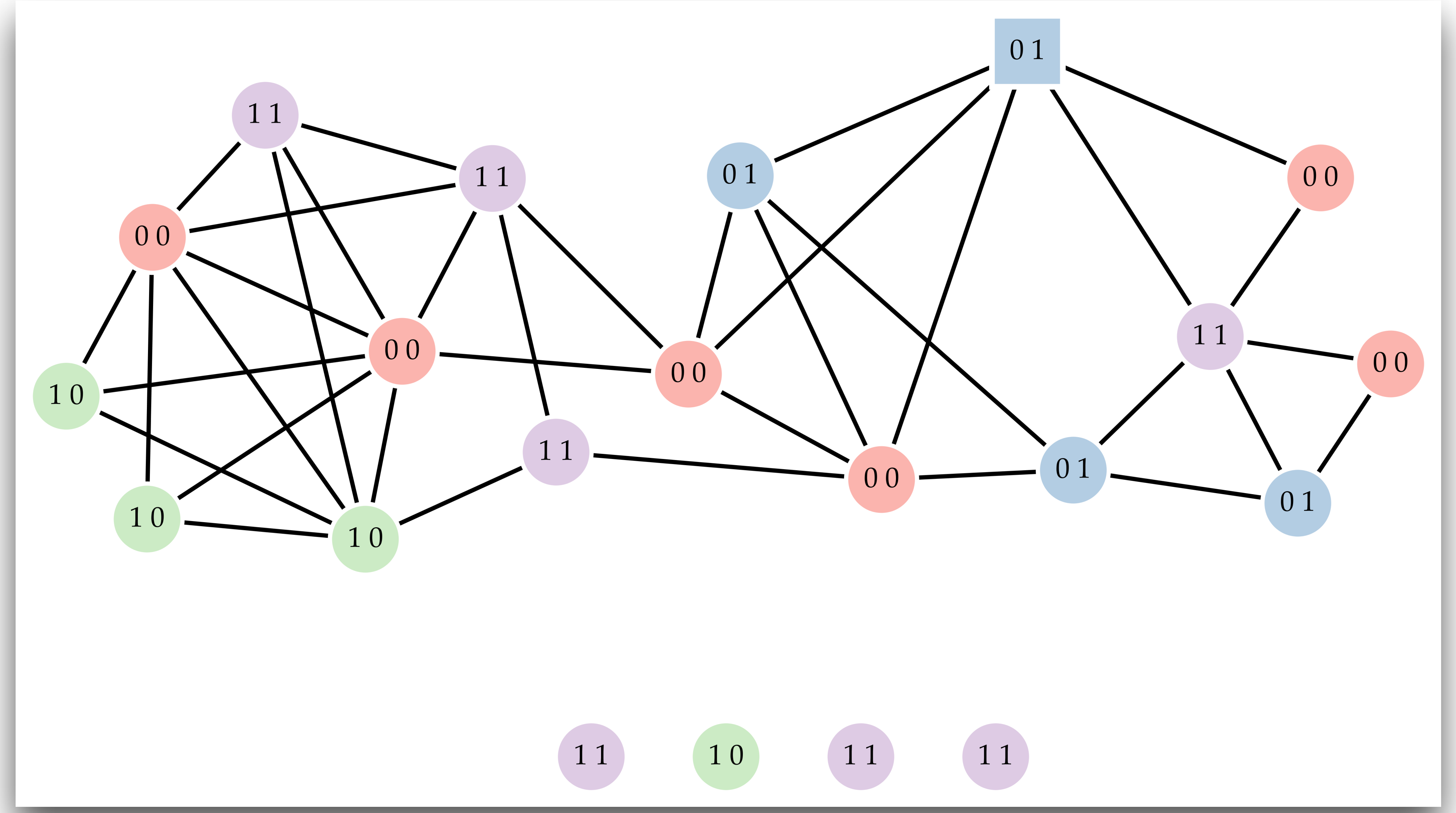
☑ the convergence of the cdf's of test statistics are rapid and depend on parameters in models



empirical examples





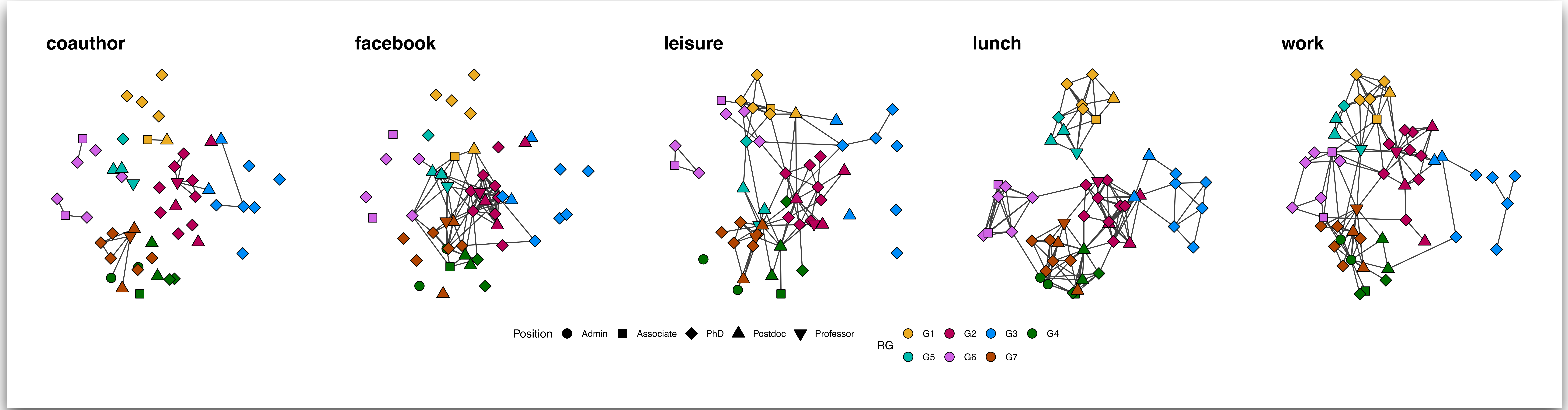


multivariate social networks

the AUCS dataset: relations between faculty and staff members at a university

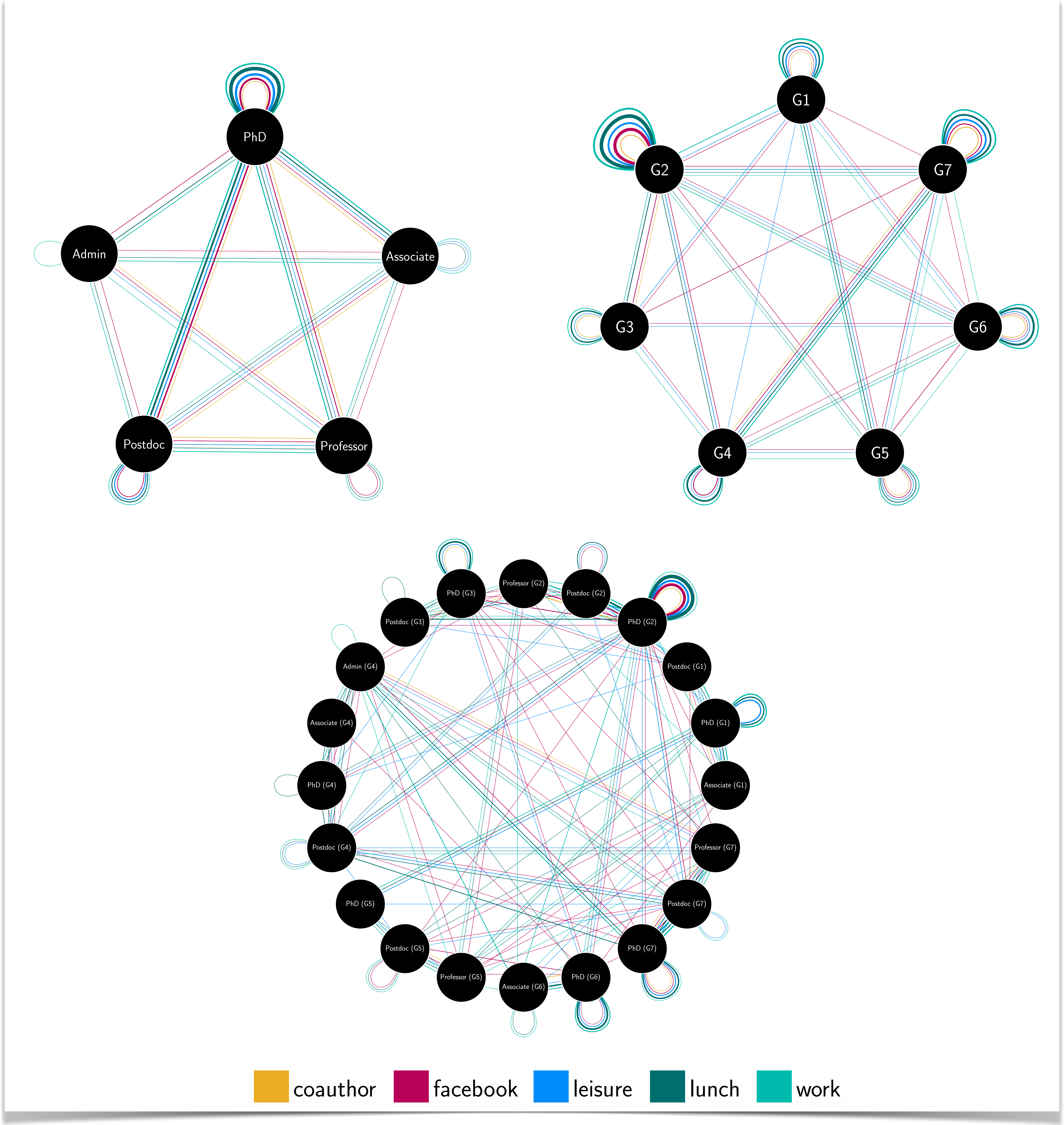
a multivariate network with multiple types of ties and vertex attributes

- ☑ five types of relations of the considered network dataset
- ☑ vertex attributes are research group (RG) and academic position

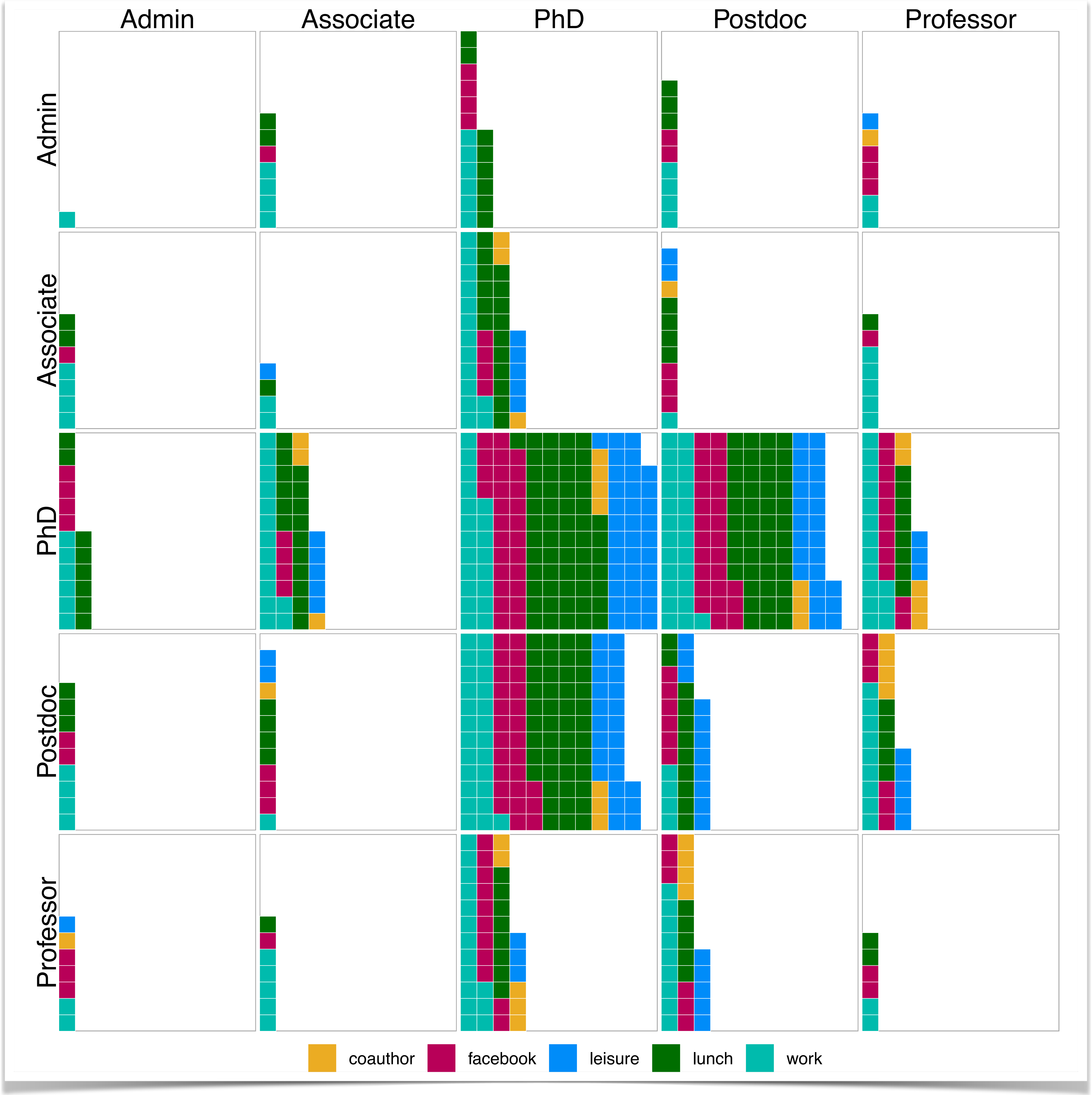


aggregation based on single or combined vertex attributes ⇒ three multigraphs

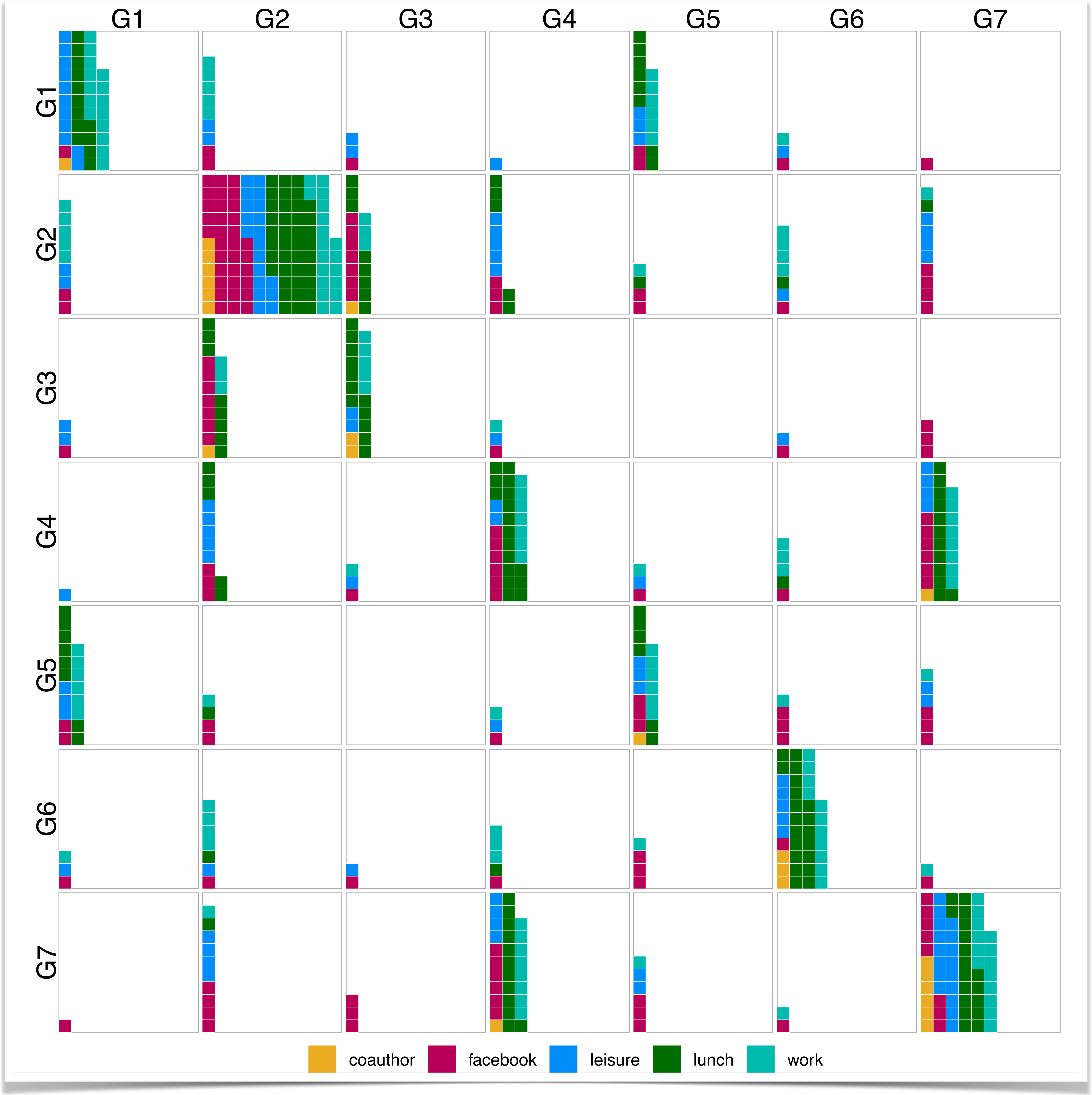
aggregated multigraphs



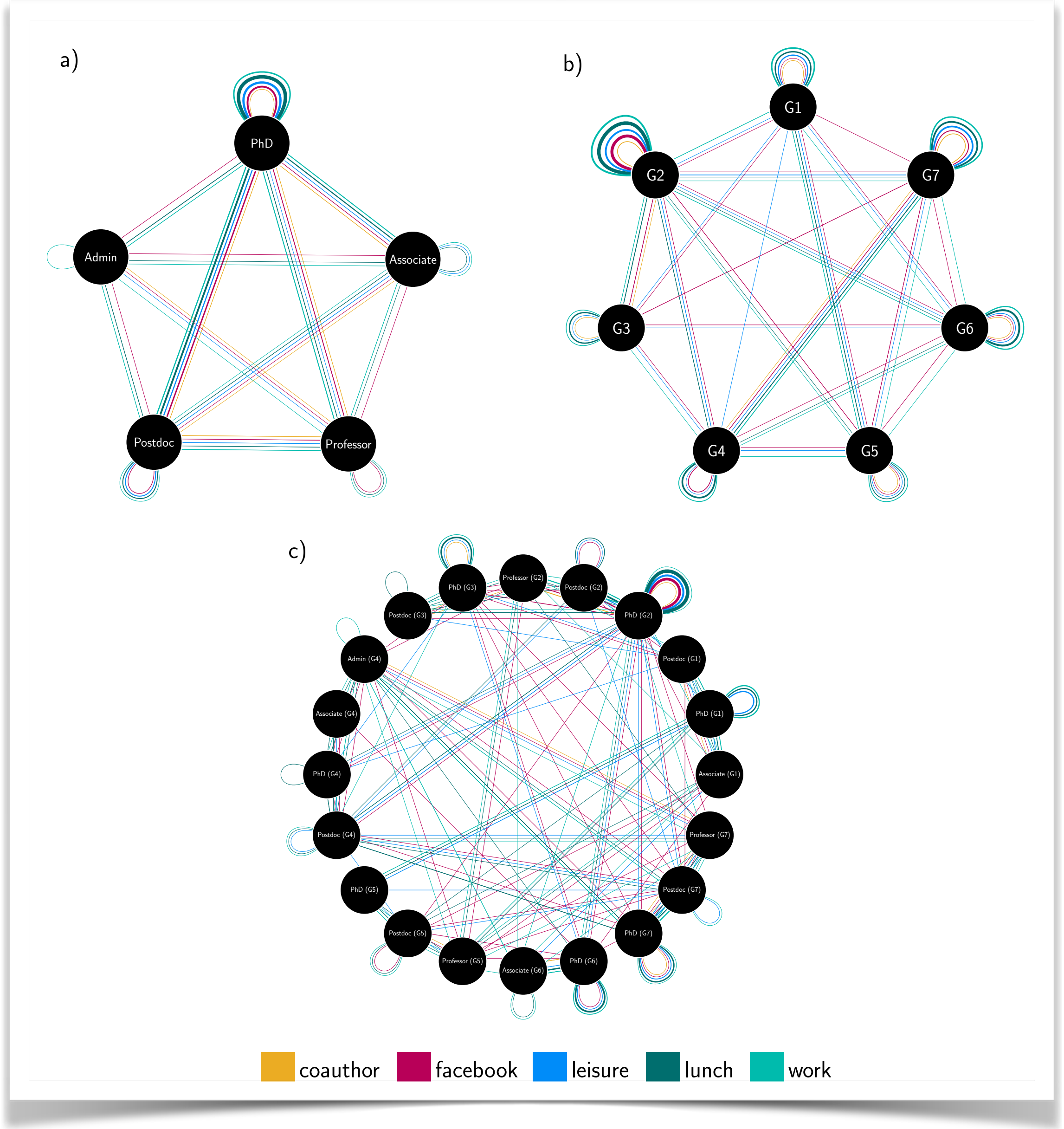
aggregated multigraphs: waffle matrices

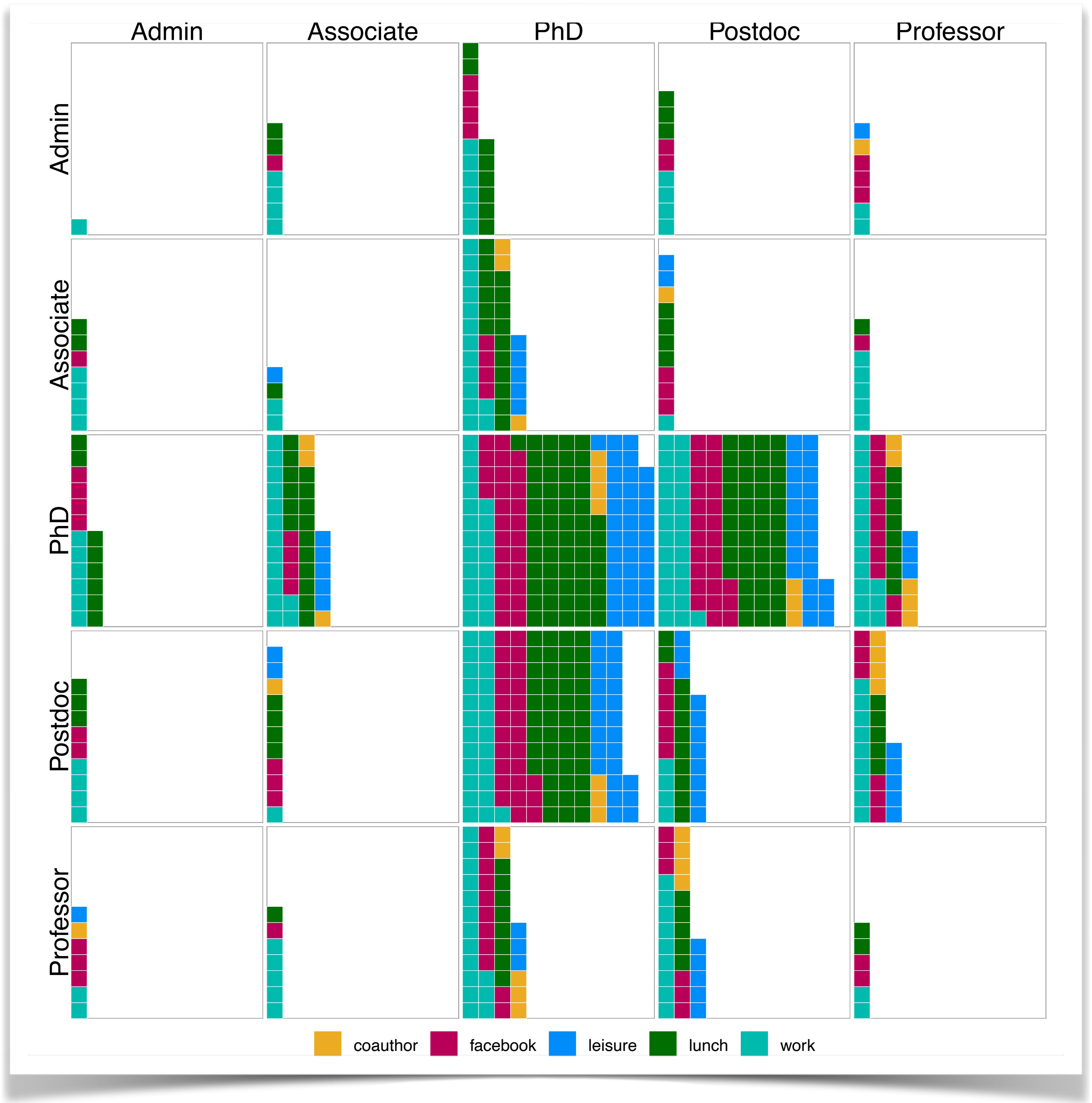


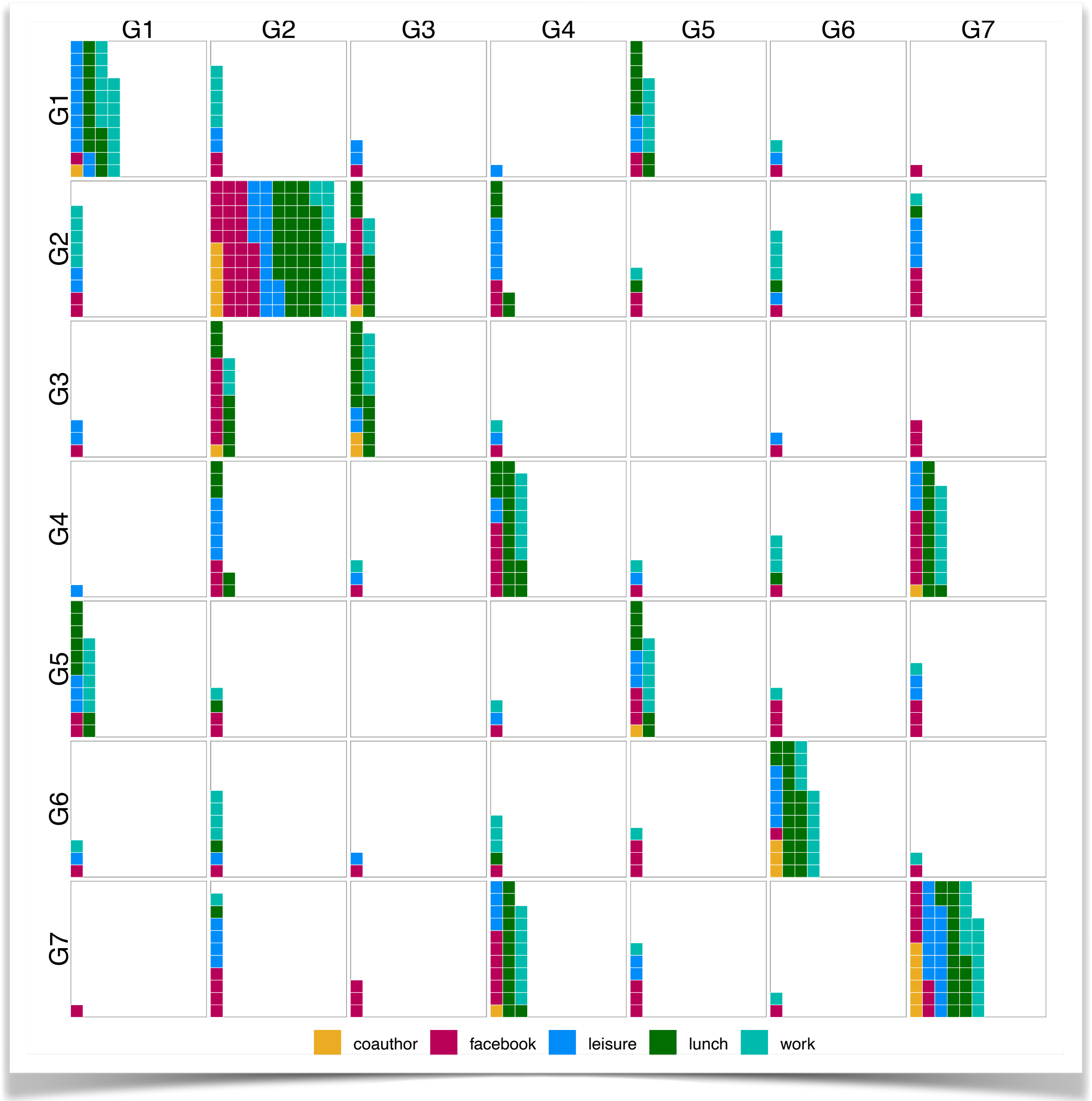
aggregated multigraphs: waffle matrices



aggregated multigraphs: waffle matrices







✓ M_1 and M_2

- tendency for within and between vertex category edges (homophily/heterophily)

✓ M_2 and R_2

- simplicity statistics
- single ties within vertex category (isolation)

✓ R_0 and R_1

- R_0: tendency for isolated vertices (network diffusion)
- R_1: simple occupancy of edges

✓ $R_0 + R_1$ compared to $R_3 + \cdots + R_k$

- tendency for strengthening ties (multiplexity)

✓ M_1 and R_1

- single ties within vertex category (isolation)

✓ interval estimates for R_k

- if overlapping for multiple edge types ⇒ multiplexity

observed edge multiplicities

- complexity sequence $\mathbf{R} = (R_0, R_1, \ldots, R_k)$ where

$$R_k = \sum_{i \leq j} \sum I(M_{ij} = k) \quad \text{for } k = 0,1,\ldots,m$$

is the frequencies of edge multiplicities

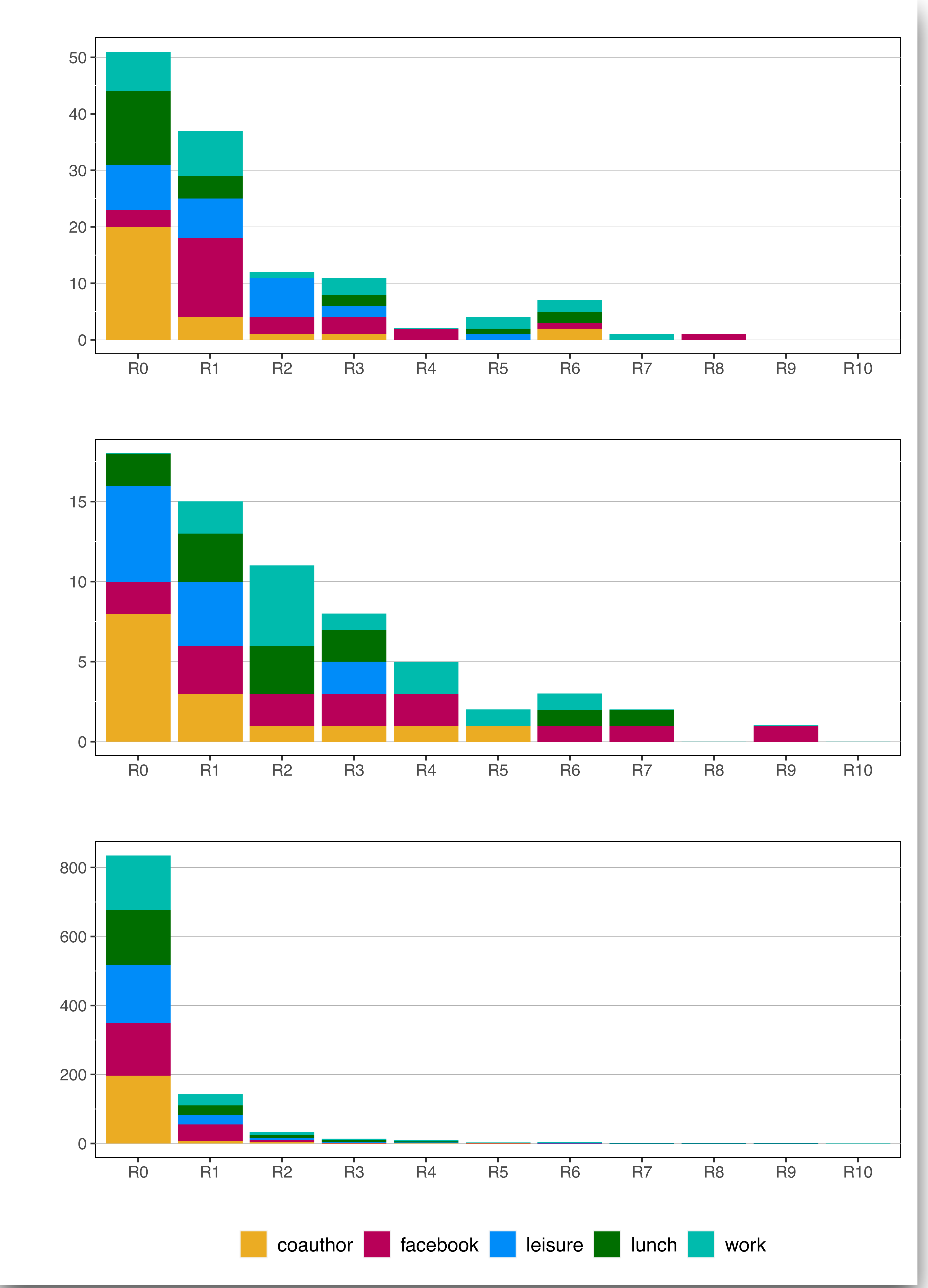
✓ R_0 number of vertex pair sites with no edge occupancy

✓ R_1 number of vertex pair sites with single edge occupancy

✓ R_2 number of vertex pair sites with double edge occupancy

$\vdots$

compare to expected values from random multigraph models



expected edge multiplicities

expected values and variance of R_k are derived and estimated under models

- ☑ $\sim$ IEA($\mathbf{Q}$)

  MLE of the edge assignment probabilities given by the empirical fraction of each edge type

- ☑ $\sim$ IEAS($\mathbf{Q}(\mathbf{d})$)

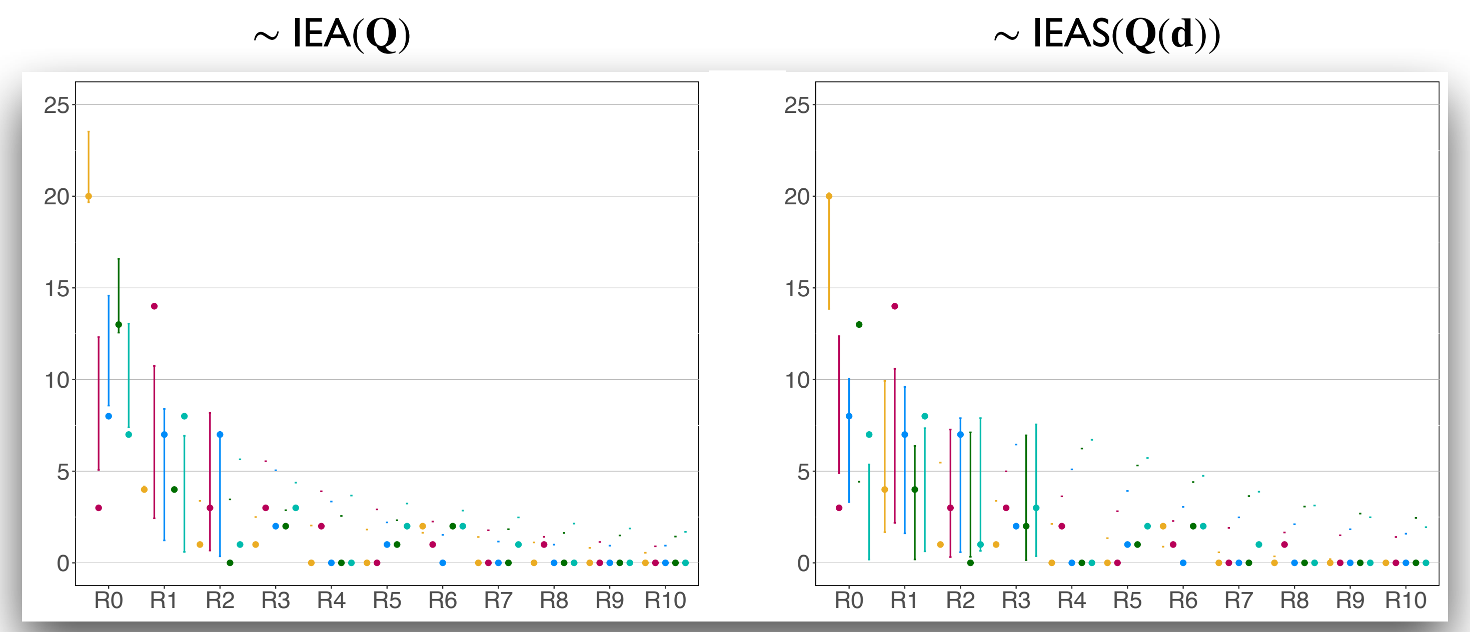
  (IEA approximation of RSM)

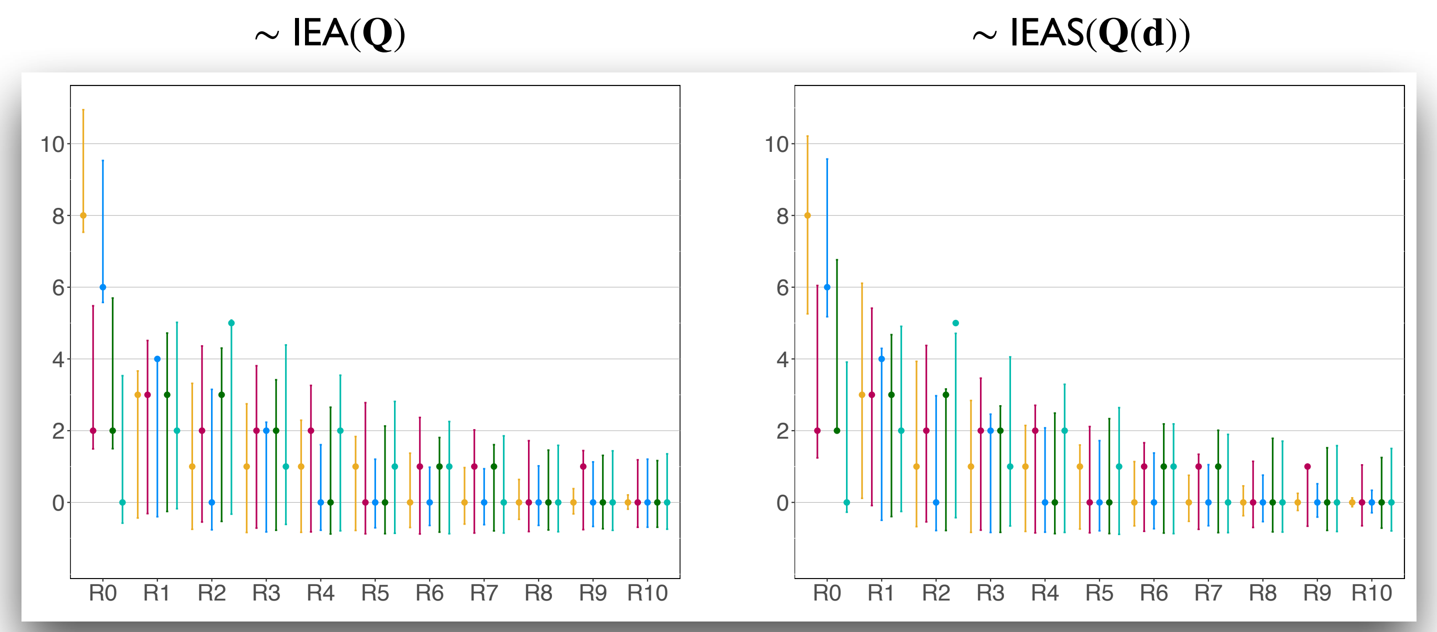
  edge assignment probabilities given by the observed degree sequence of each edge type

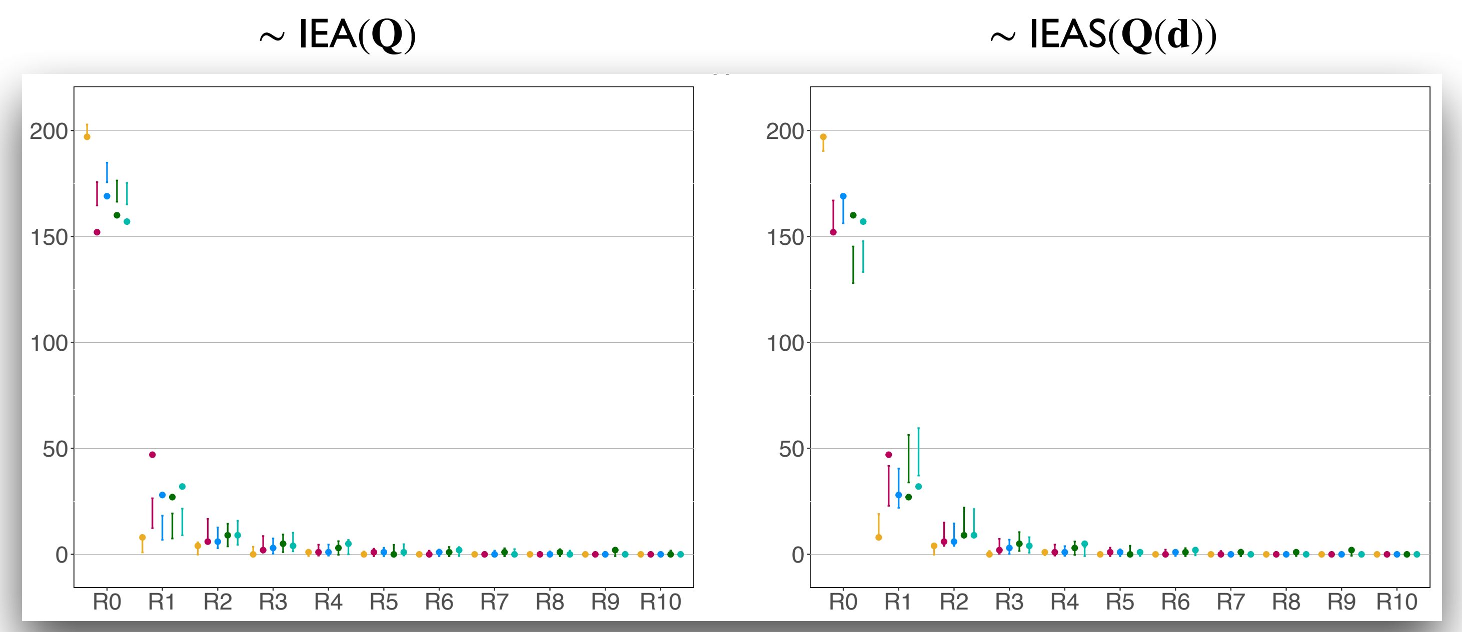
approx 95% intervals illustrated

$$\hat{E} \pm 2\sqrt{\hat{V}}$$

multiplexity analysis







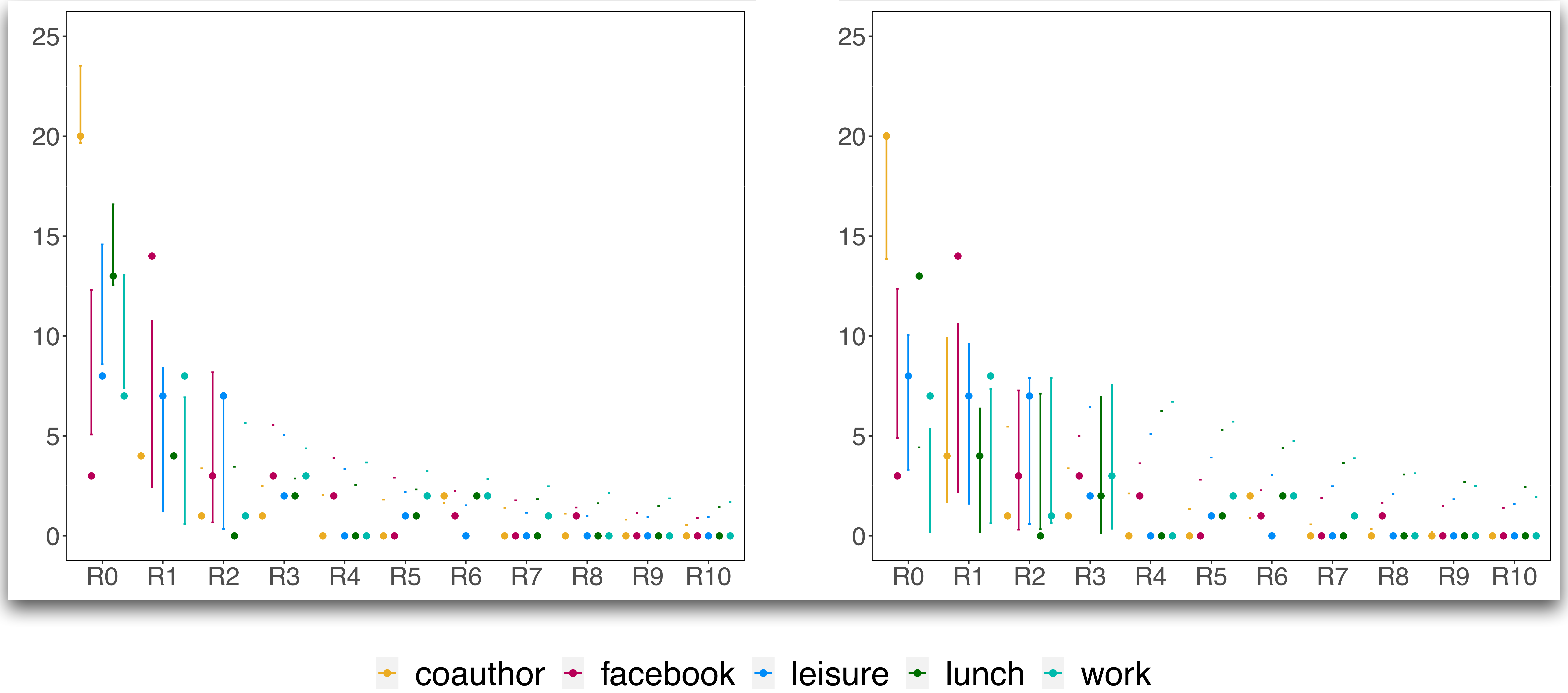
approx 95% intervals illustrated

$$\hat{E} \pm 2\sqrt{\hat{V}}$$

multiplexity analysis

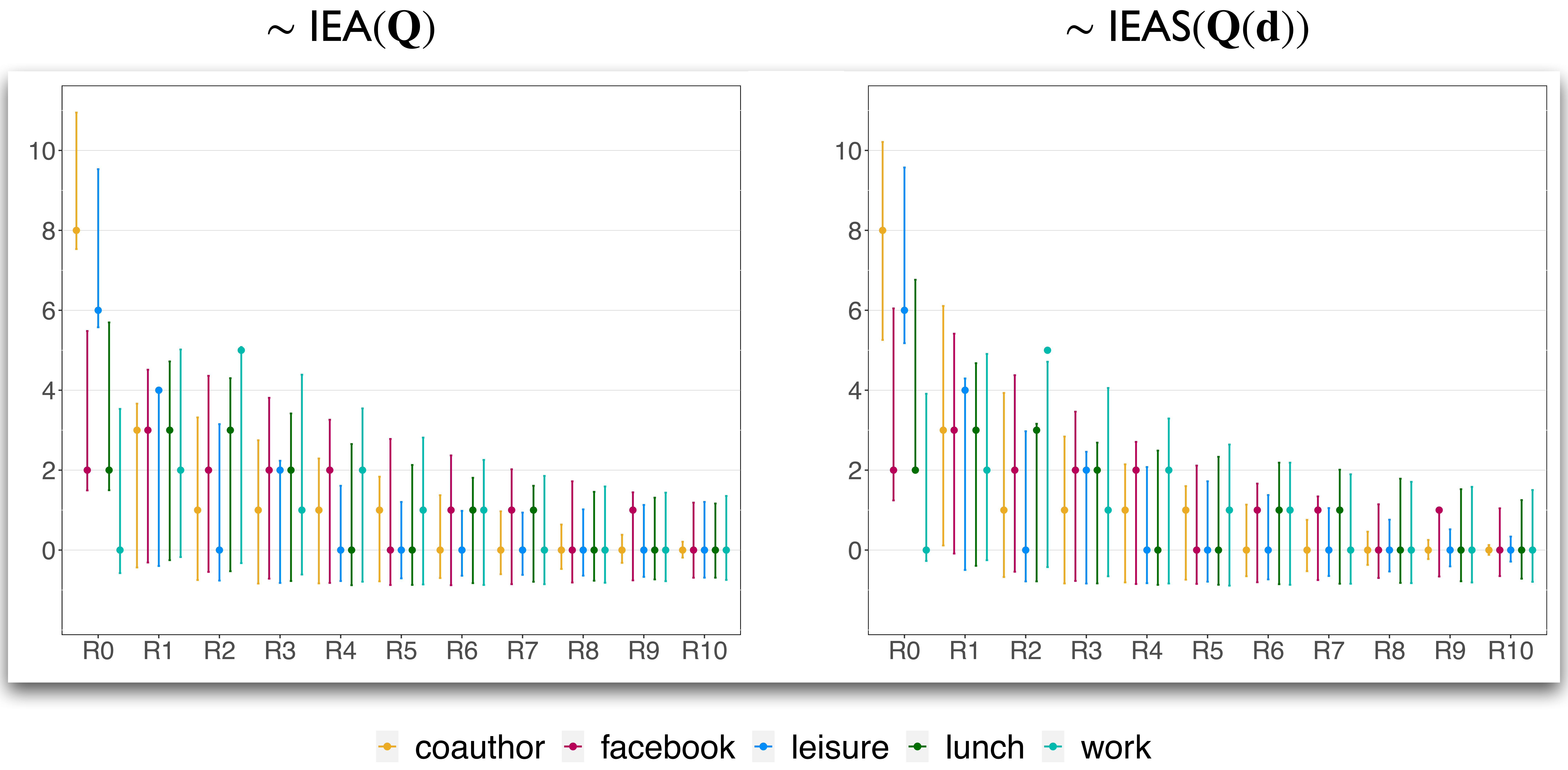
multigraph based on position

$\sim$ IEA($\mathbf{Q}$) $\sim$ IEAS($\mathbf{Q}(\mathbf{d})$)



multiplexity analysis

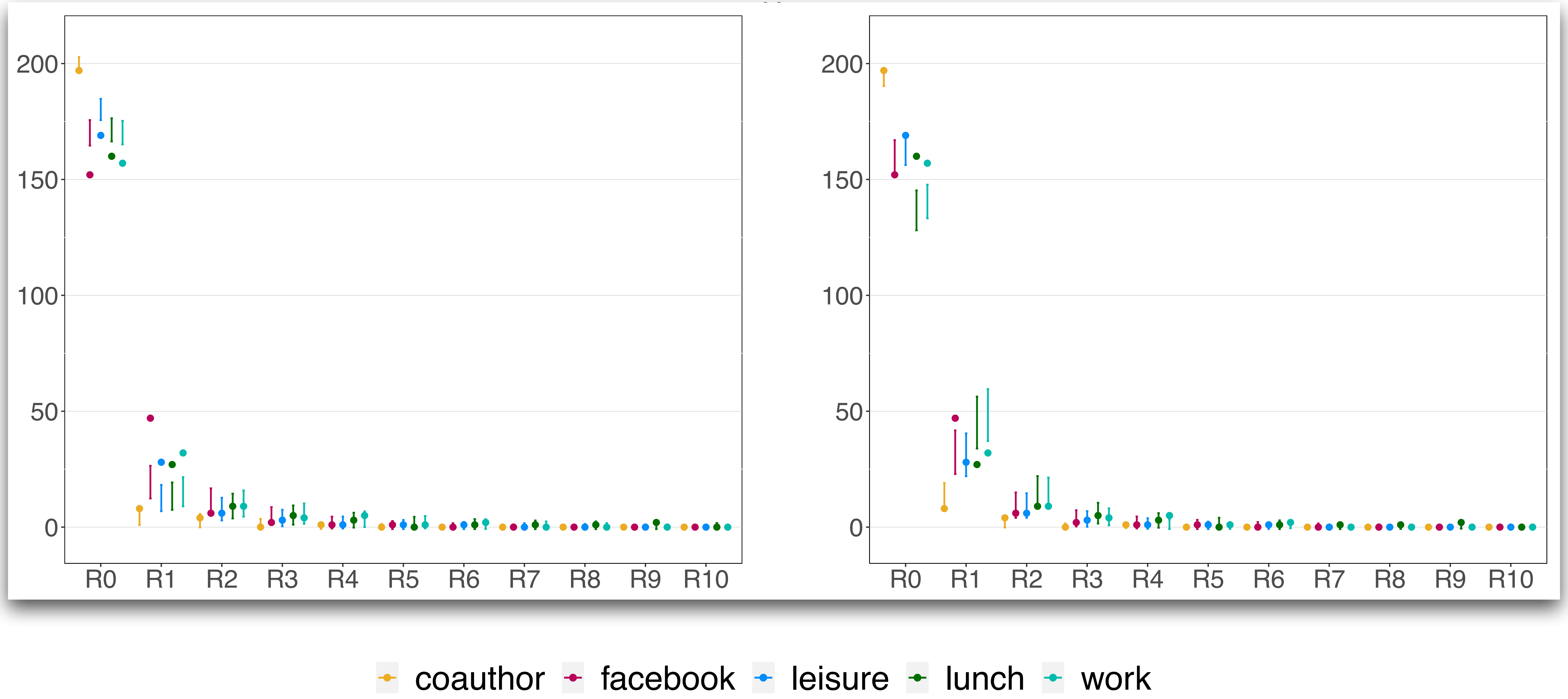
multigraph based on research group



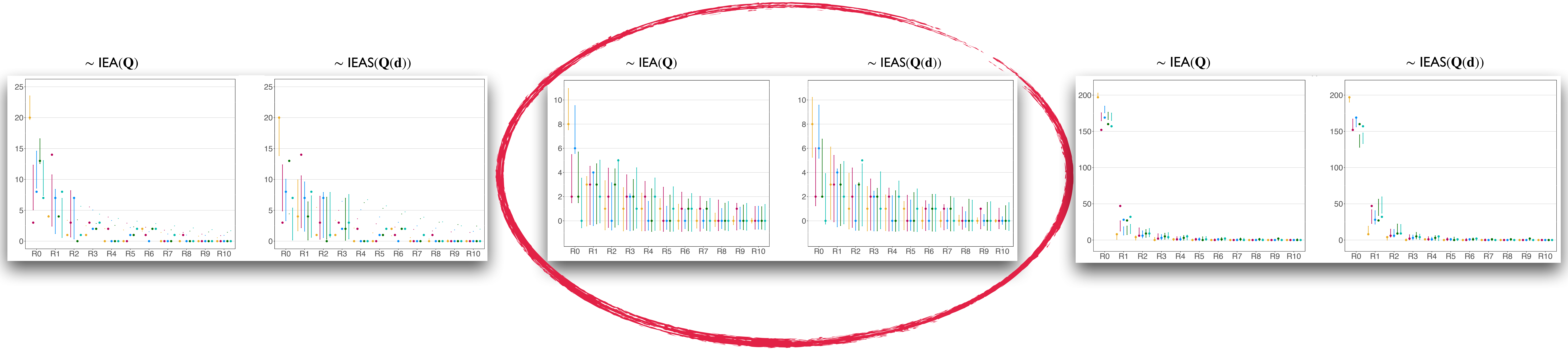
multiplexity analysis

multigraph based on position and research group

$\sim$ IEA($\mathbf{Q}$) $\sim$ IEAS($\mathbf{Q}(\mathbf{d})$)



multiplexity analysis



- ☑ both models provide good fits for multigraphs based on research groups
- ☑ intervals overlapping implies
  - ✓ indicating that tie occurrences are not significantly different
  - ✓ tie occurrences are not independent implying
  - ✓ some form of edge dependency is needed in the model specification

analysing ego networks

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)

analysing ego networks

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)

analysing ego networks

Krackhardt's High-tech Managers Networks (1987)

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

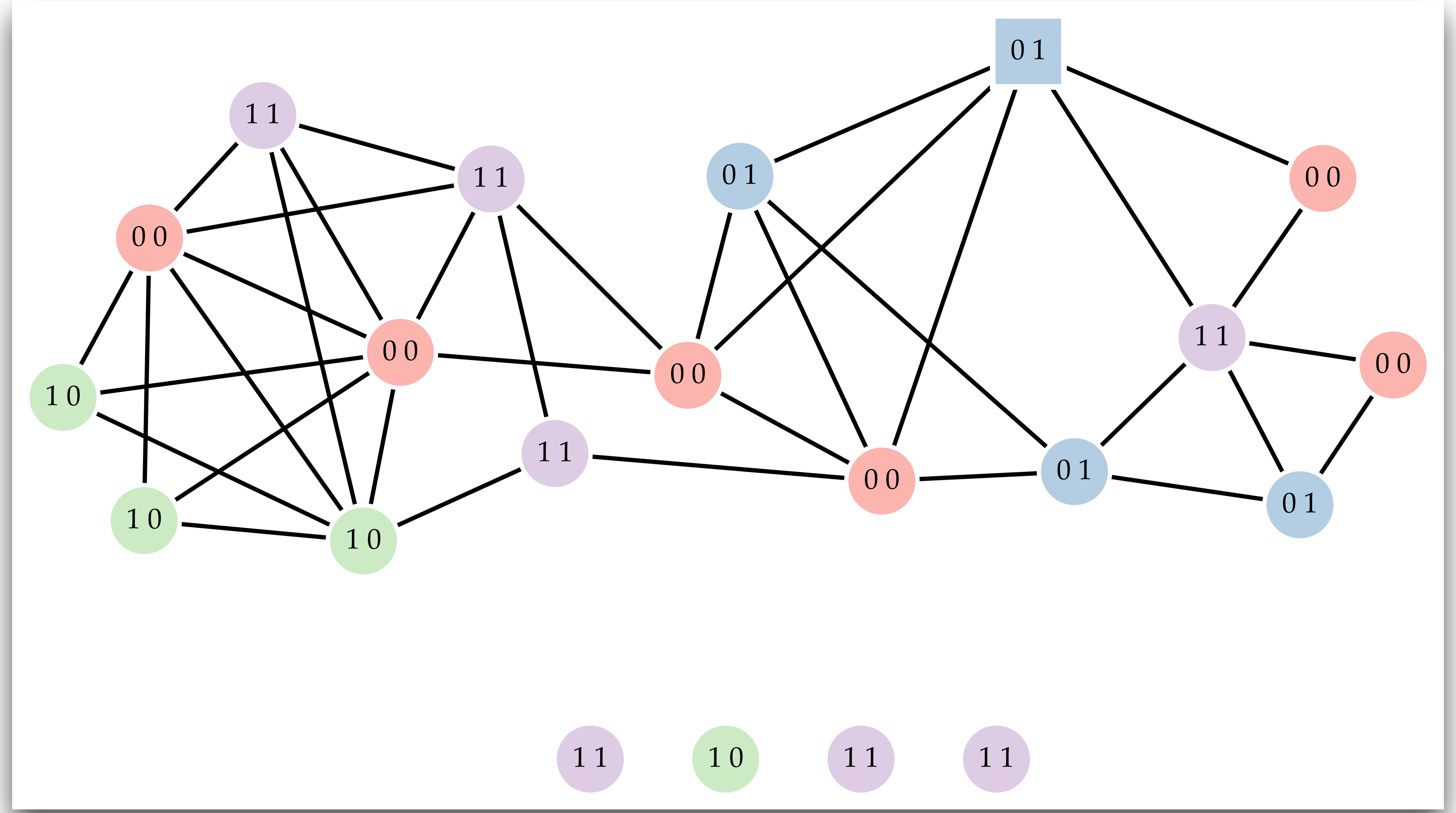
relations:	actor attributes:
- undirected friendship - directed advice	- department - level - age - tenure

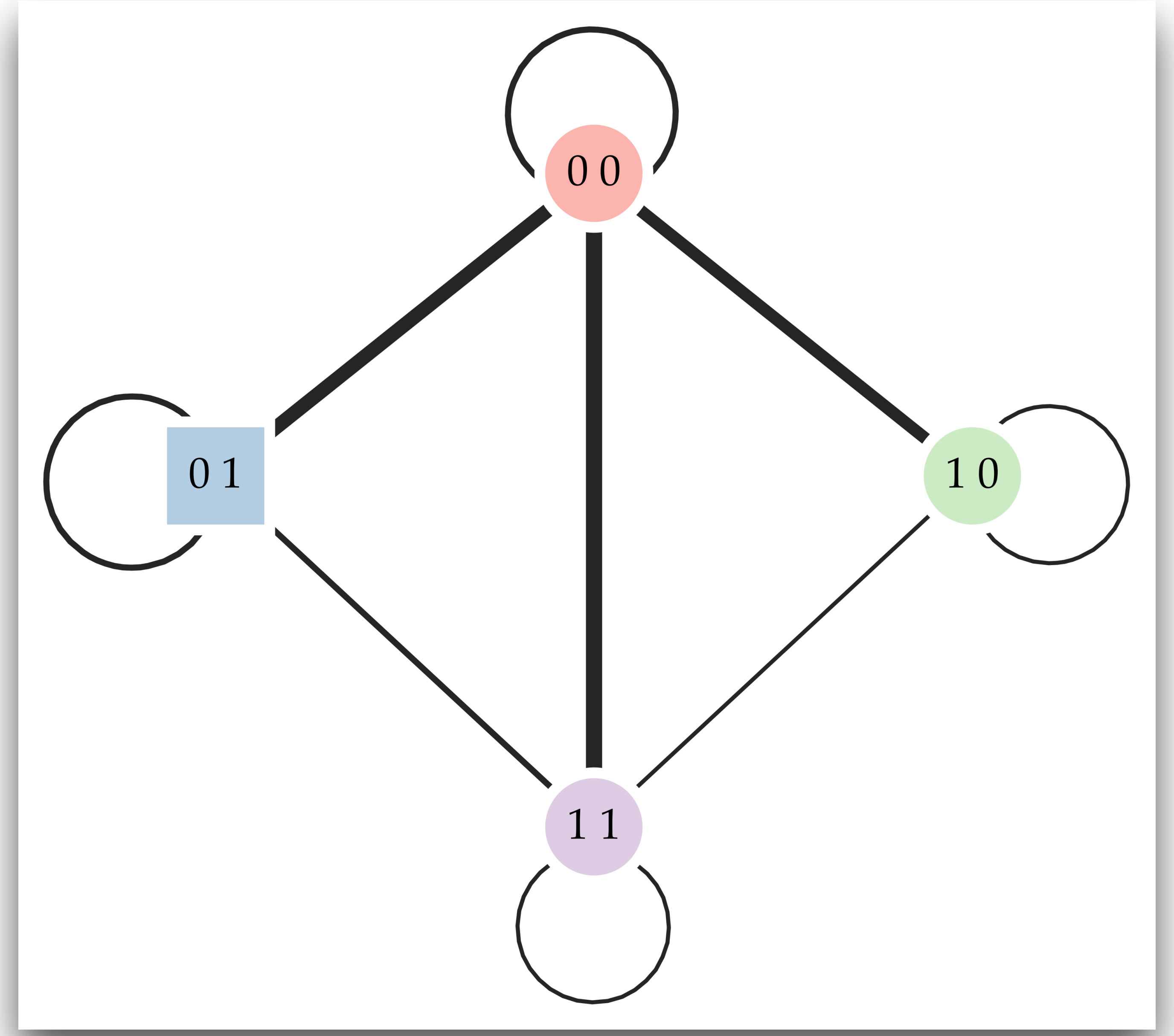
(also includes the relations each ego perceived among all other managers)

- ☑ 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

aggregated multigraphs

ego 1's original network and aggregated multigraph

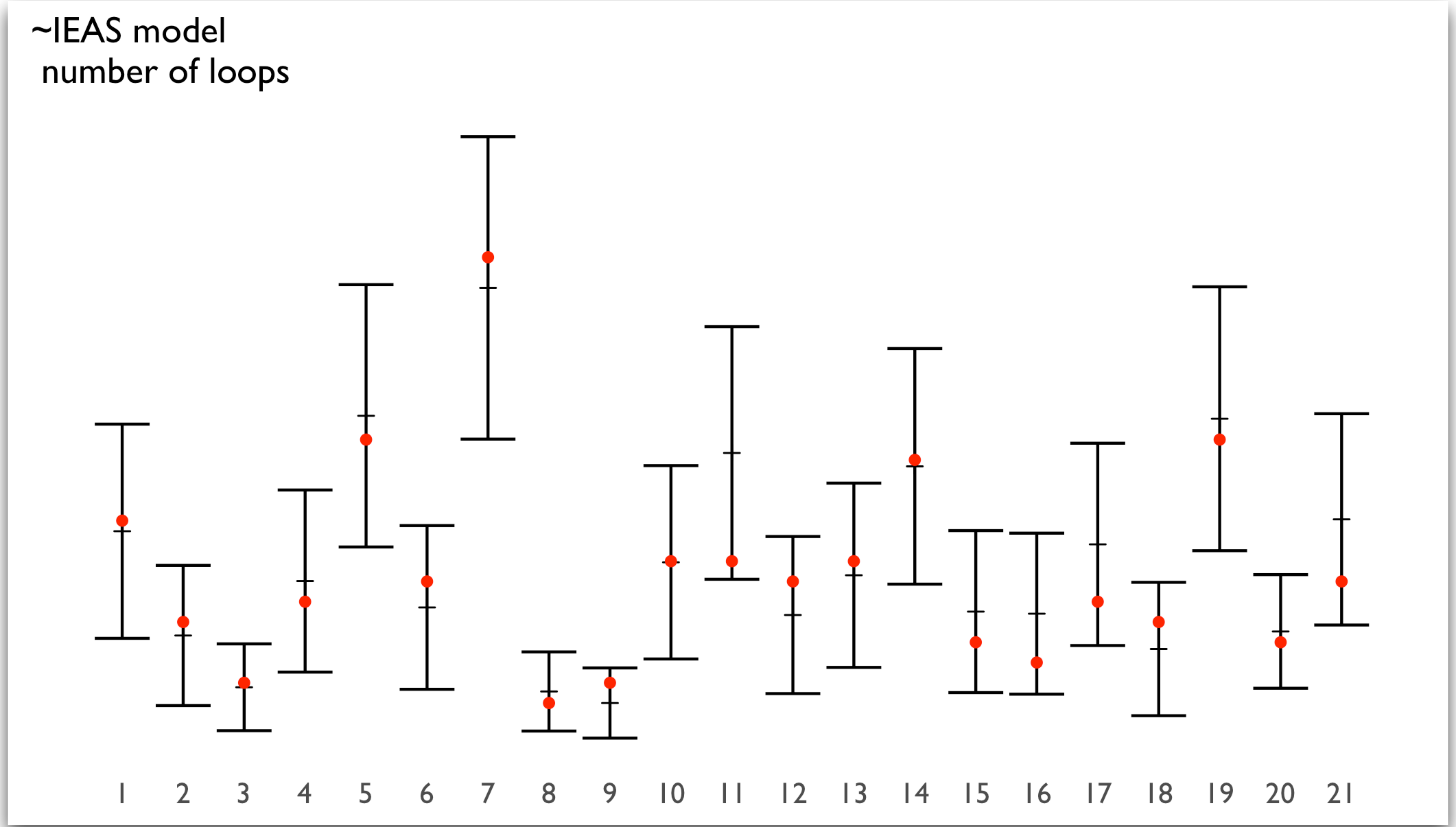




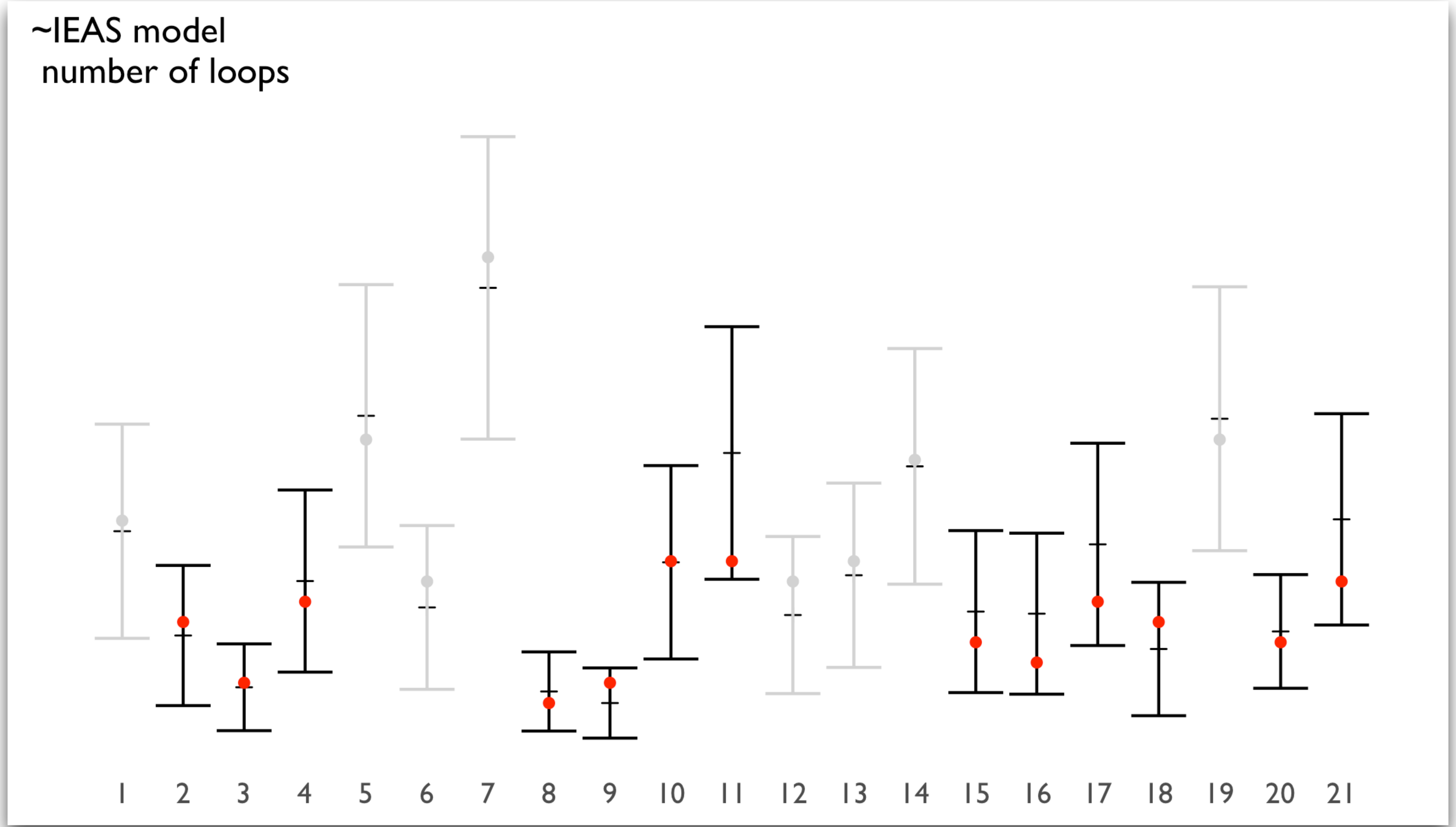
aggregated to

aggregated multigraphs

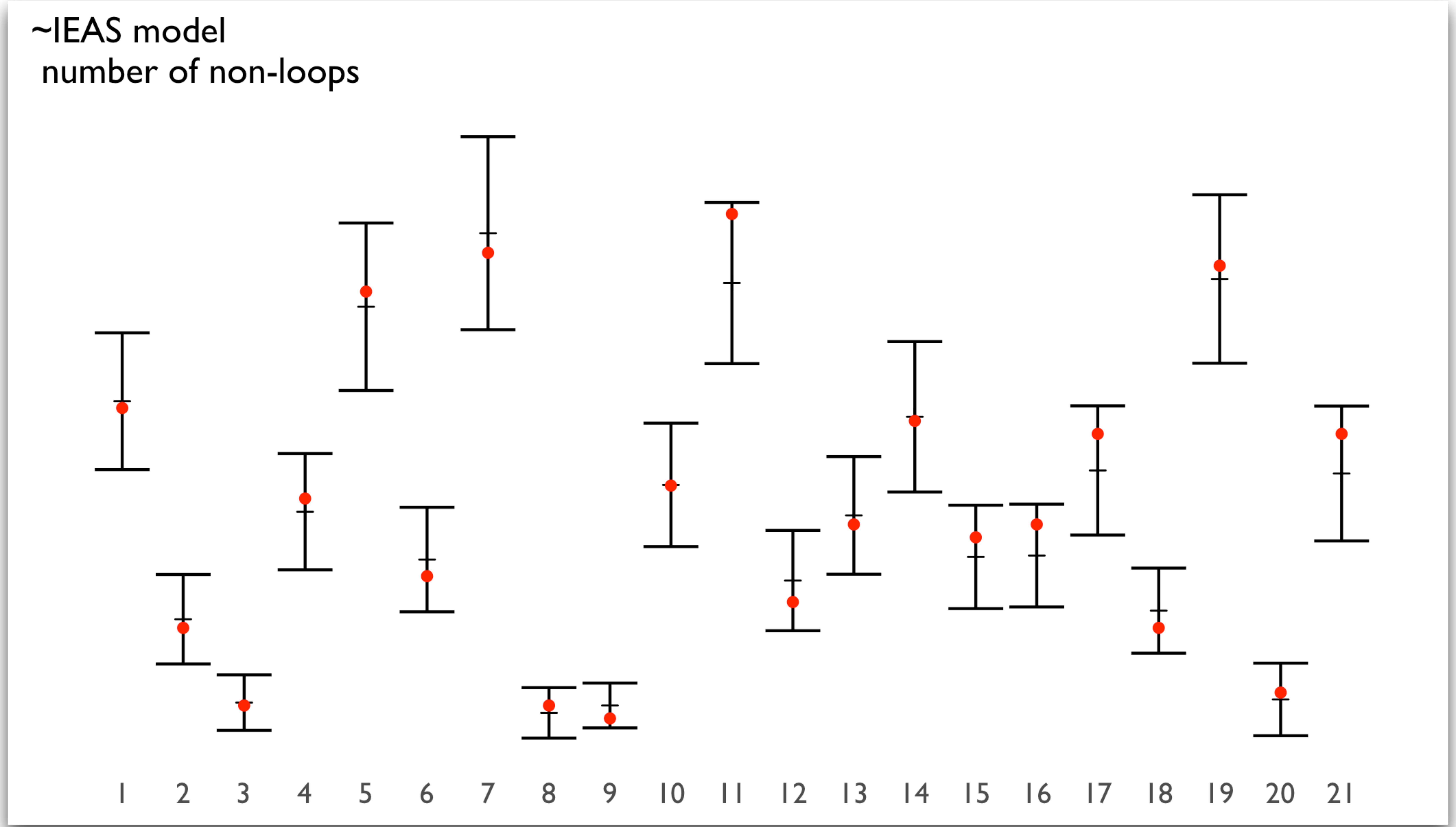
example: number of loops



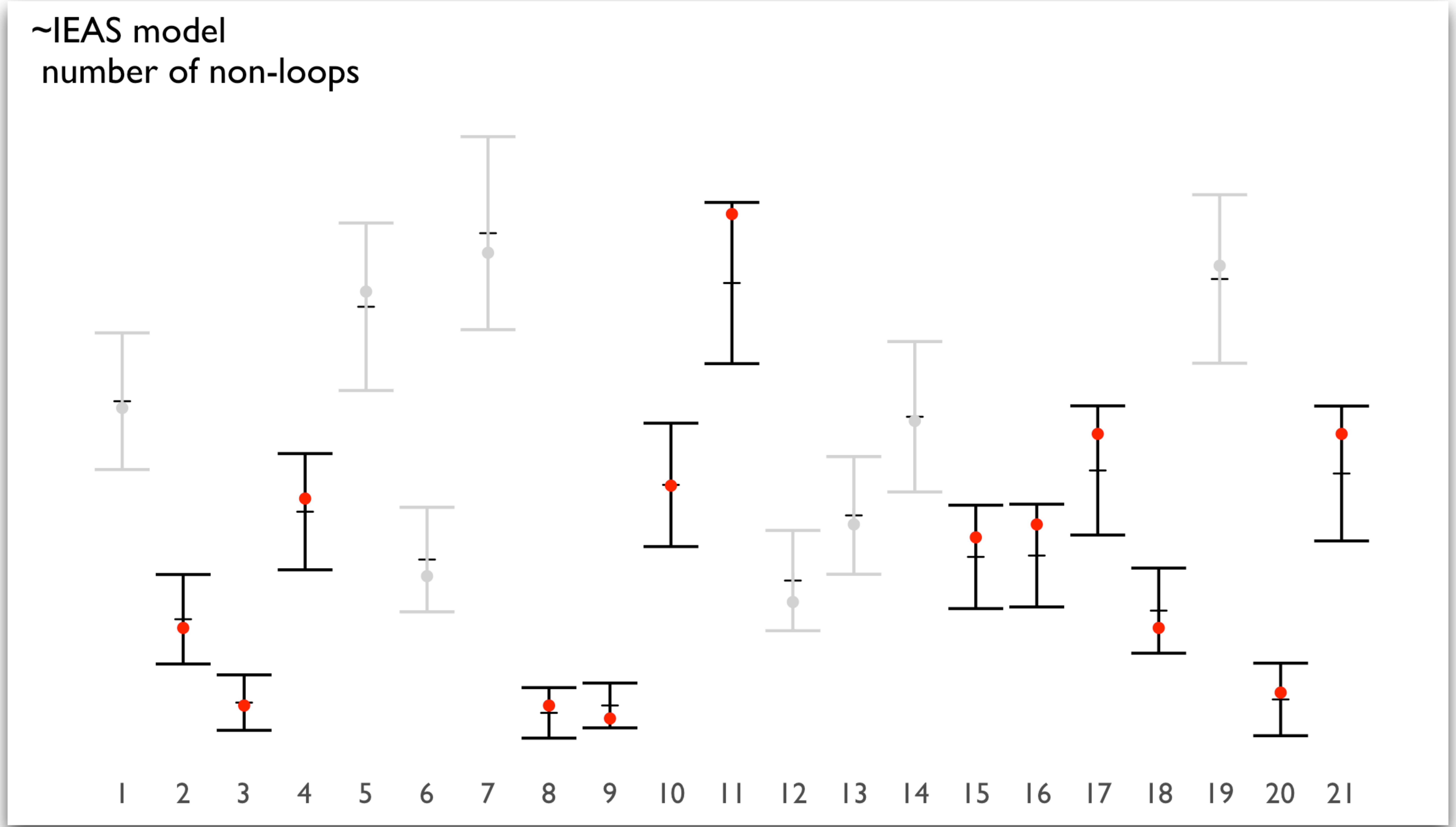
example: goodness of fit



example: number of non-loops



example: goodness of fit



character networks

the under-/misrepresentation of female characters in movies

- ☑ male vs. female frequency of appearances
- ☑ gender role and content stereotyping
- ☑ structure and dynamics of narrative texts

Alison Bechdel's
"Dykes to Watch Out For" (1985)

data (~ 10 000 movies):

☑ character networks

(e.g. Cornell Movie-Dialogues Corpus)

- ✓ type, frequency and direction of interactions
- ✓ topic of dialogues
- ✓ number of lines

☑ meta data

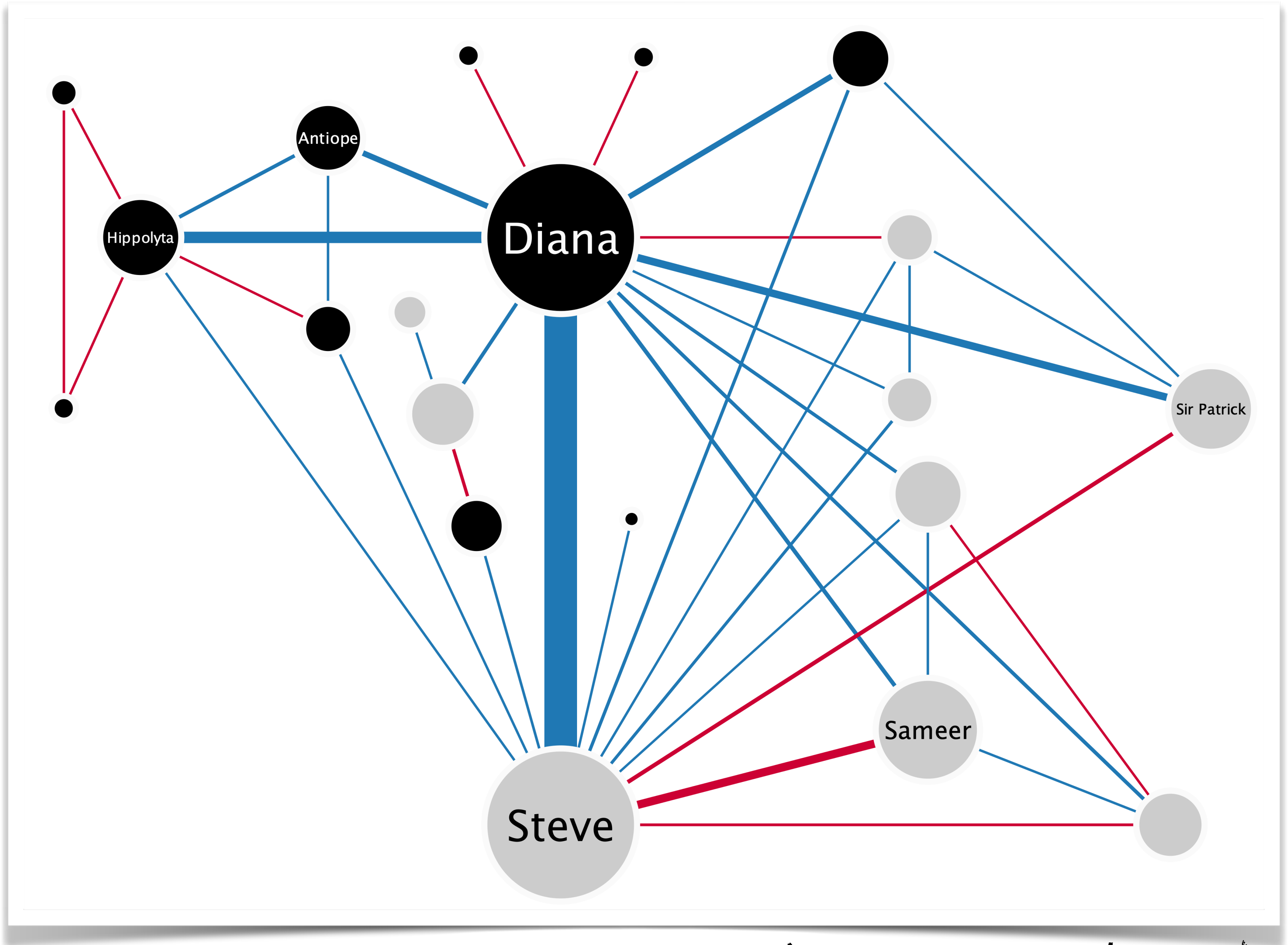
(from e.g. IMDb.com, bechdeltest.com)

- ✓ gender of writer(s), director(s), lead actor(s)
- ✓ year
- ✓ rating
- ✓ country of production
- ✓ box office revenue

character networks

the under-/misrepresentation of female characters in movies

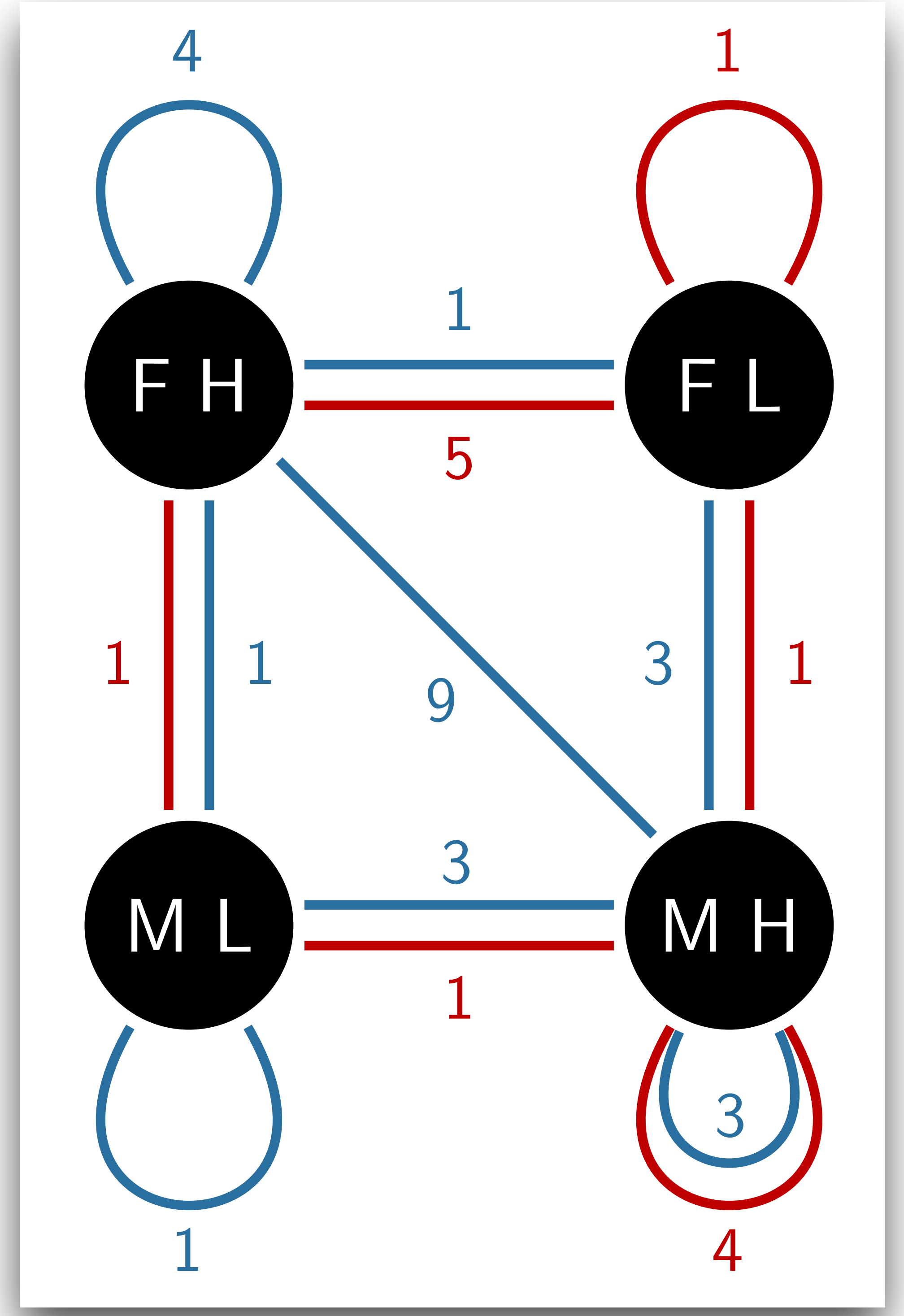
- ☑ male vs. female frequency of appearances
- ☑ gender role and content stereotyping
- ☑ structure and dynamics of narrative texts



Wonder Woman (2017)

multigraph aggregations based on

- ☑ gender (female/male)
- ☑ number of lines (low/high)
- ☑ topic (pass or fail bechdel test)



models used to study
e.g. homophily/heterophily

final words on presented framework

- ☑ let research question and social theories guide data transformations
- ☑ attention to density of various edges and vertex variable distributions
- ☑ only applicable to undirected networks
- ☑ visual inspections of waffle matrices are only feasible for small multigraphs
- ☑ direction of associations between different edge types not revealed

R package: https://cran.r-project.org/package=multigraphr

```
install.packages("multigraphr")

# development version
devtools::install_github("termehs/multigraphr")
```

more guides available on my website, package vignette, and GitHub

multigraphr