A Connected World

Data Analysis for Real World Network Data

Latent Variable Models

08.12.2022

A different angle to cope with dependencies

- So far, ERGM allowed us to explicitly account for (and measure) network dependencies
- Another way to capture network dependencies is by making use of latent variable models
- Models within this class assume that latent (unobserved) variables Z_i are associated with each node i, and that all dependencies between edges is due to these latent variables

Definition: Latent Variable Network Models

A latent variable network model is a statistical model that relates the set of observed edges $Y = (Y_{ij})$ to a set of latent variables $Z = (Z_i)$. The actor-specific latent variables Z_i can, in general, be of any dimension and be in the discrete or continuous domain. All dependence between edges Y_{ij} and Y_{kh} is assumed to be captured by the latent variables z_i, z_j, z_k, and *z_h.*

$$Y_{ij} | z_i, z_j \sim F(z_i, z_j)$$

Intuition

- Nodes possess some latent attributes (e.g. unobserved group membership, positioning in a social space) which influences tie behavior
- The idea is to estimate this latent structure, to gain an understanding of it and/or control for it while doing inference on covariates
- Let us start with the simplest (and most popular) application of latent variable models… community detection

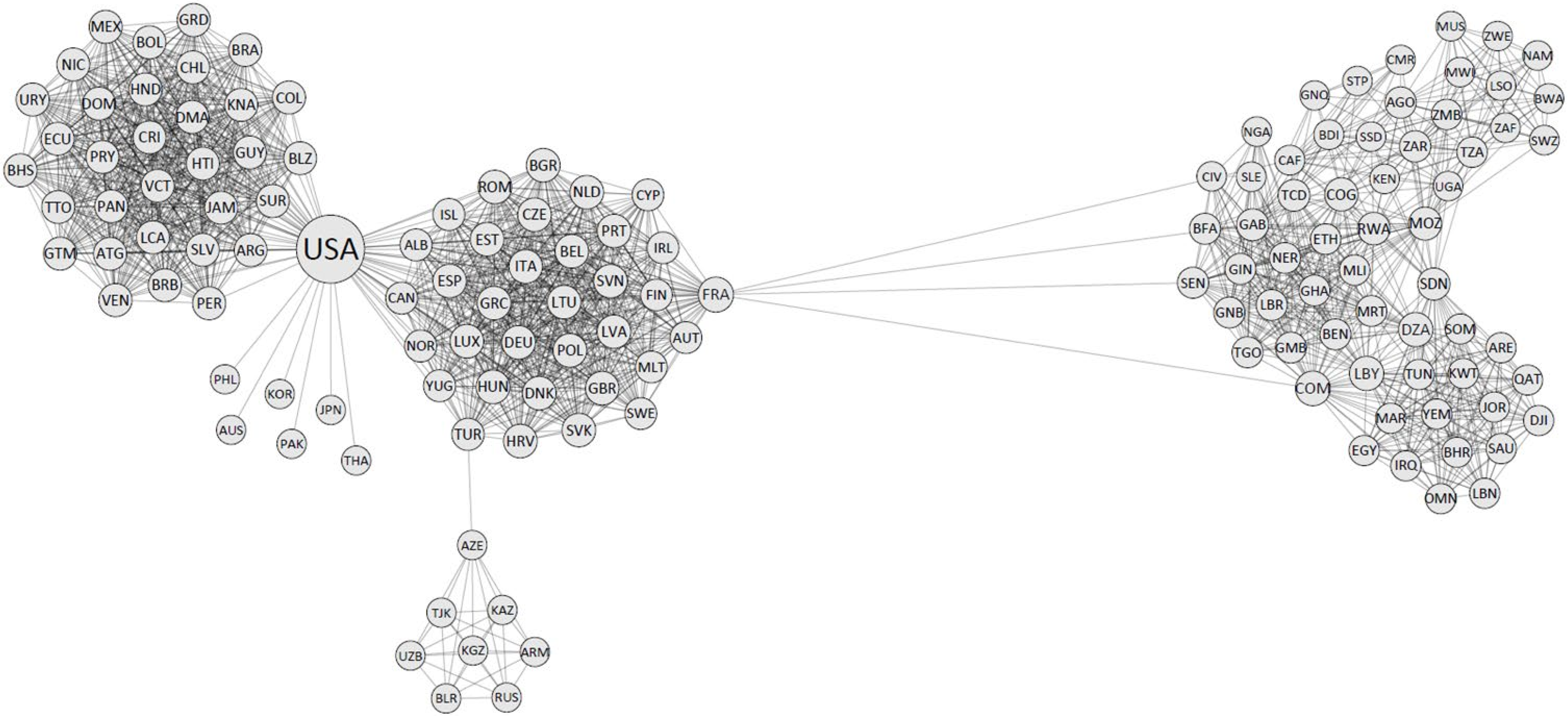
Community Detection in Networks

- Networks are often organized in smaller sub-groups
- Sometimes those subgroups are known and well defined (ex. political parties in a parliamentary network, classes in a school)
- More often that is not the case (ex. friendship circles on a facebook network, different cells in a network of terrorists)

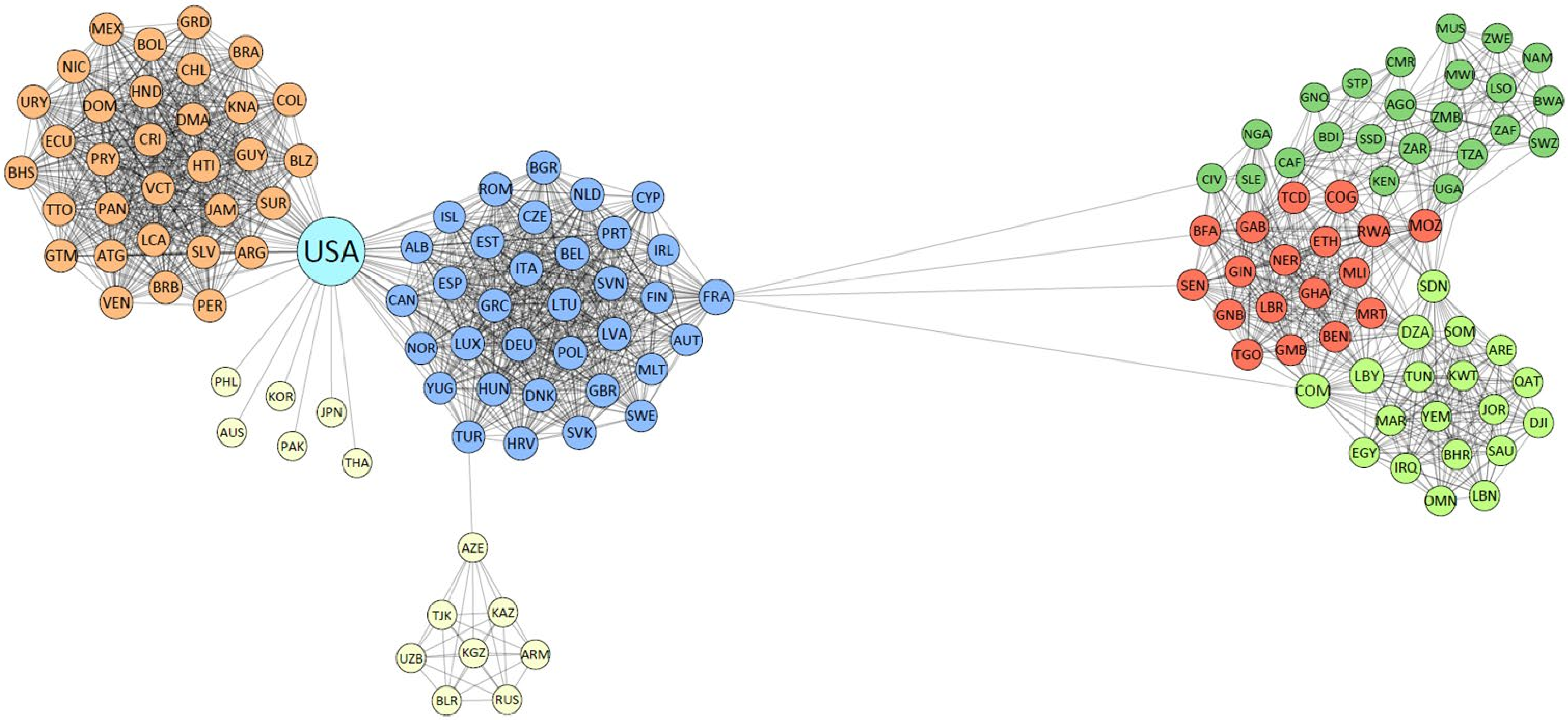
Community Detection in Networks

- We can treat the community membership as a latent (unobserved) variable and try to estimate it
- In models with built-in community structure, the probability of forming a tie within a group is typically higher than forming one between groups
- Other types of structures, such as core-periphery, are possible

We want to go from here…



…to here!



USA
MEX
GRD
BOL
BRA
NIC
CHL
HND
URY
DOM
KNA
COL
DMA
ECU
CRI
GUY
BLZ
PRY
HTI
BHS
VCT
SUR
TTO
PAN
JAM
LCA
GTM
ATG
SLV
ARG
BRB
VEN
PER
PHL
KOR
JPN
AUS
PAK
THA
BGR
ROM
NLD
CYP
ISL
CZE
PRT
EST
IRL
ALB
BEL
ITA
ESP
SVN
CAN
GRC
LTU
FIN
FRA
LVA
AUT
NOR
LUX
DEU
POL
MLT
YUG
HUN
DNK
GBR
SWE
TUR
HRV
SVK
AZE
TJK
KAZ
UZB
KGZ
ARM
BLR
RUS
MUS
ZWE
NAM
CMR
MWI
STP
LSO
GNQ
AGO
BWA
ZMB
BDI
SSD
ZAF
NGA
ZAR
SWZ
CAF
TZA
CIV
SLE
KEN
UGA
TCD
COG
BFA
GAB
RWA
MOZ
ETH
NER
SEN
GIN
MLI
SDN
GHA
GNB
LBR
MRT
DZA
SOM
BEN
ARE
TGO
GMB
LBY
TUN
KWT
QAT
COM
JOR
MAR
YEM
DJI
EGY
BHR
SAU
IRQ
OMN
LBN

How to perform community detection?

- Many heuristic methods available (see Fortunato & Hric, 2016)
- Most popular is modularity maximization: assign node to groups in a way that maximises some target function (“modularity”)
- Other methods based on matrix factorization (i.e. spectral decomposition)

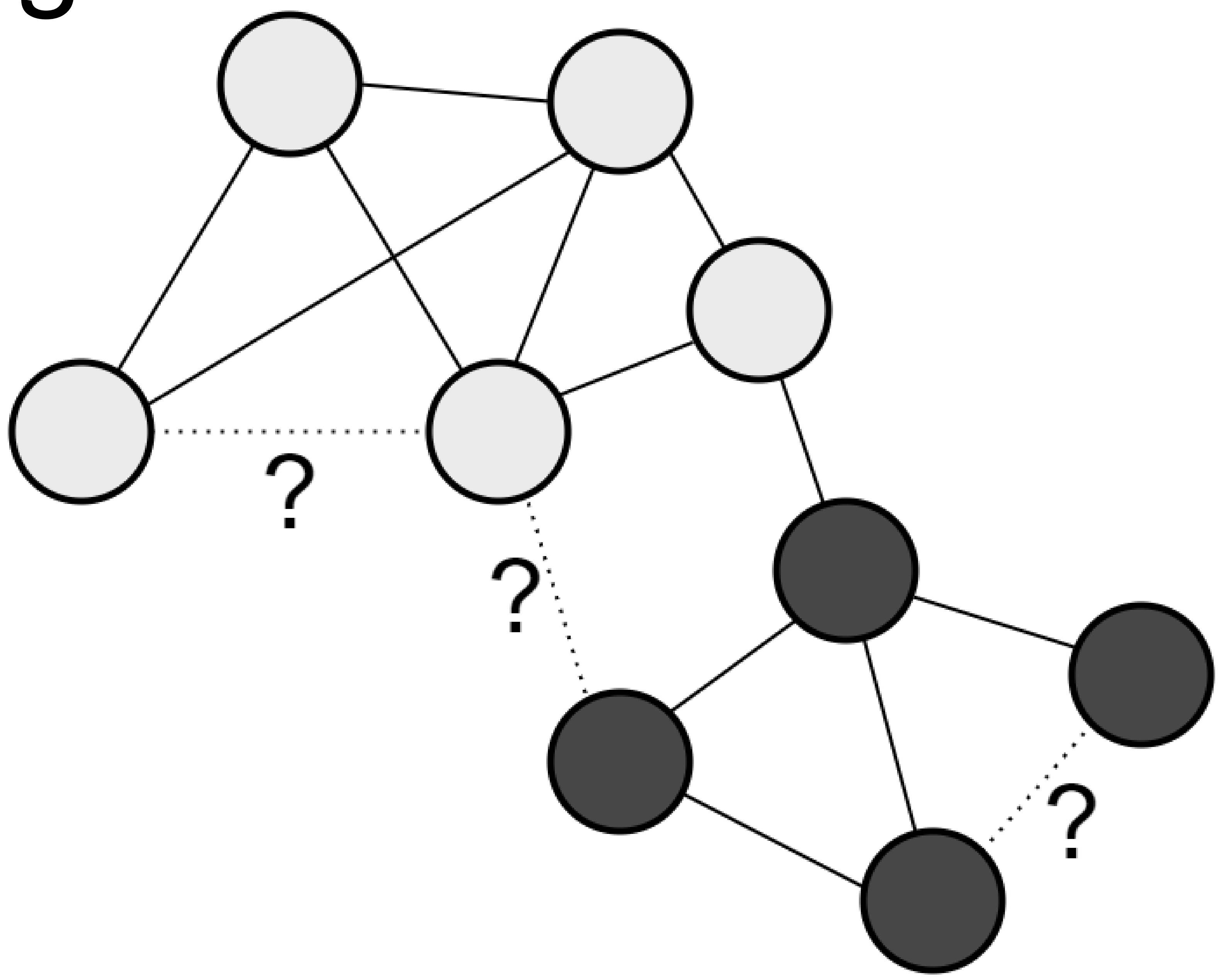
Heuristics - Pros and Cons

- **Pros:**
  - Fast
  - Good at “pure” community detection
- **Cons:**
  - Not a statistical model: no uncertainty estimations, no theoretical guarantees
  - Usually not possible/straightforward to include covariates
- Simplest statistical approach: **The Stochastic Blockmodel**

Stochastic Blockmodels

Stochastic Blockmodels - Main ideas

- A probabilistic model for networks (the edges are random)
- Each node belongs to one (unobserved) class or “block”
- The probability of any two nodes to connect depends solely on the blocks to which the two nodes belong



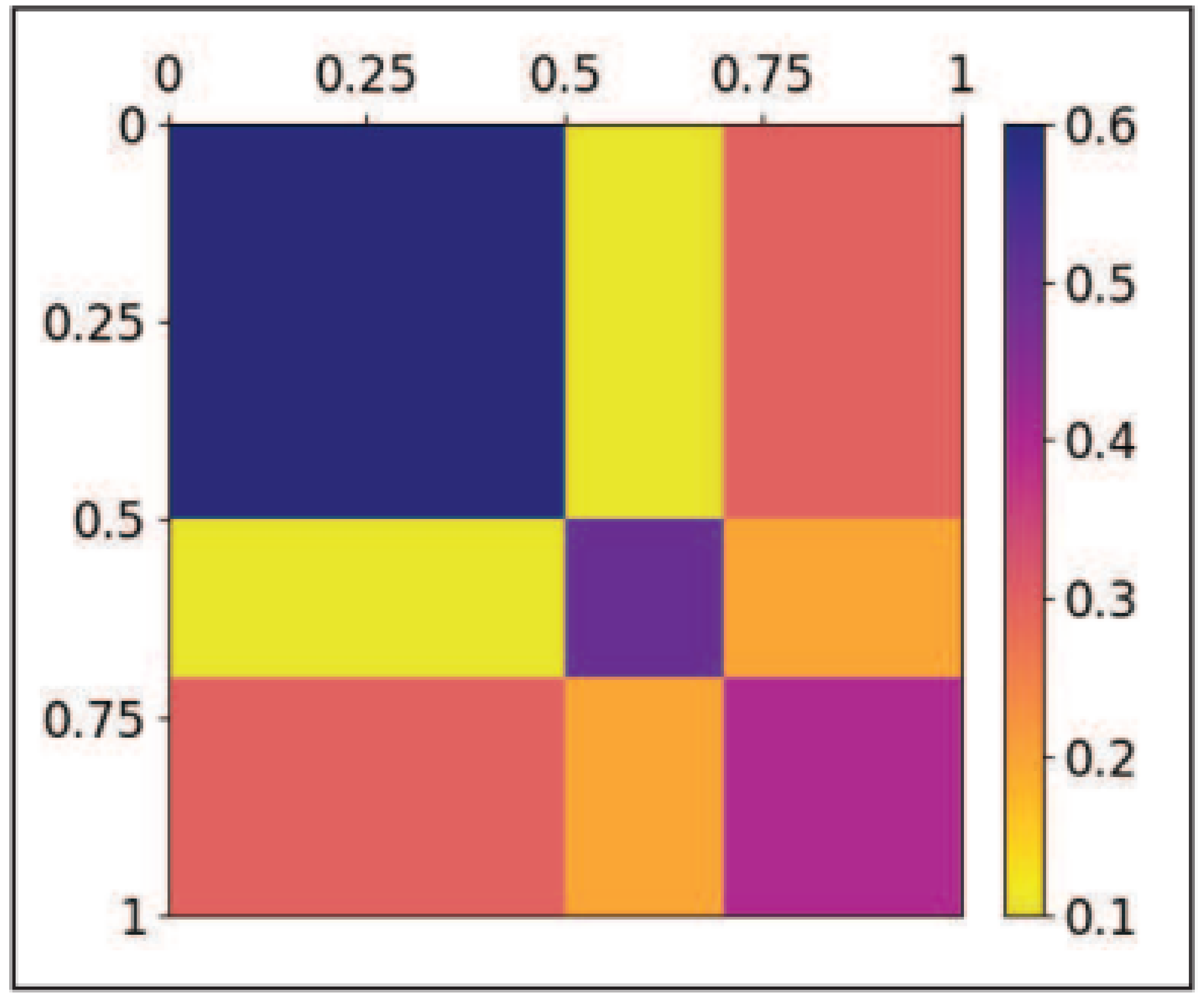
The Stochastic Blockmodel

We assume the conditional probability of a tie Y_{ij} to follow:

$$Y_{ij}|z_i, z_j \sim Bernoulli(p(z_i, z_j))$$

With $p(z_i,z_j)$ governed by a block-probability matrix $\boldsymbol{P}$.

$$\left.\begin{array}{l} \boldsymbol{\pi} = (0.5, 0.2, 0.3) \\ \boldsymbol{P} = \begin{pmatrix} 0.6 & 0.1 & 0.3 \\ 0.1 & 0.5 & 0.2 \\ 0.3 & 0.2 & 0.4 \end{pmatrix} \end{array}\right\} \iff$$



From De Nicola et al., 2022

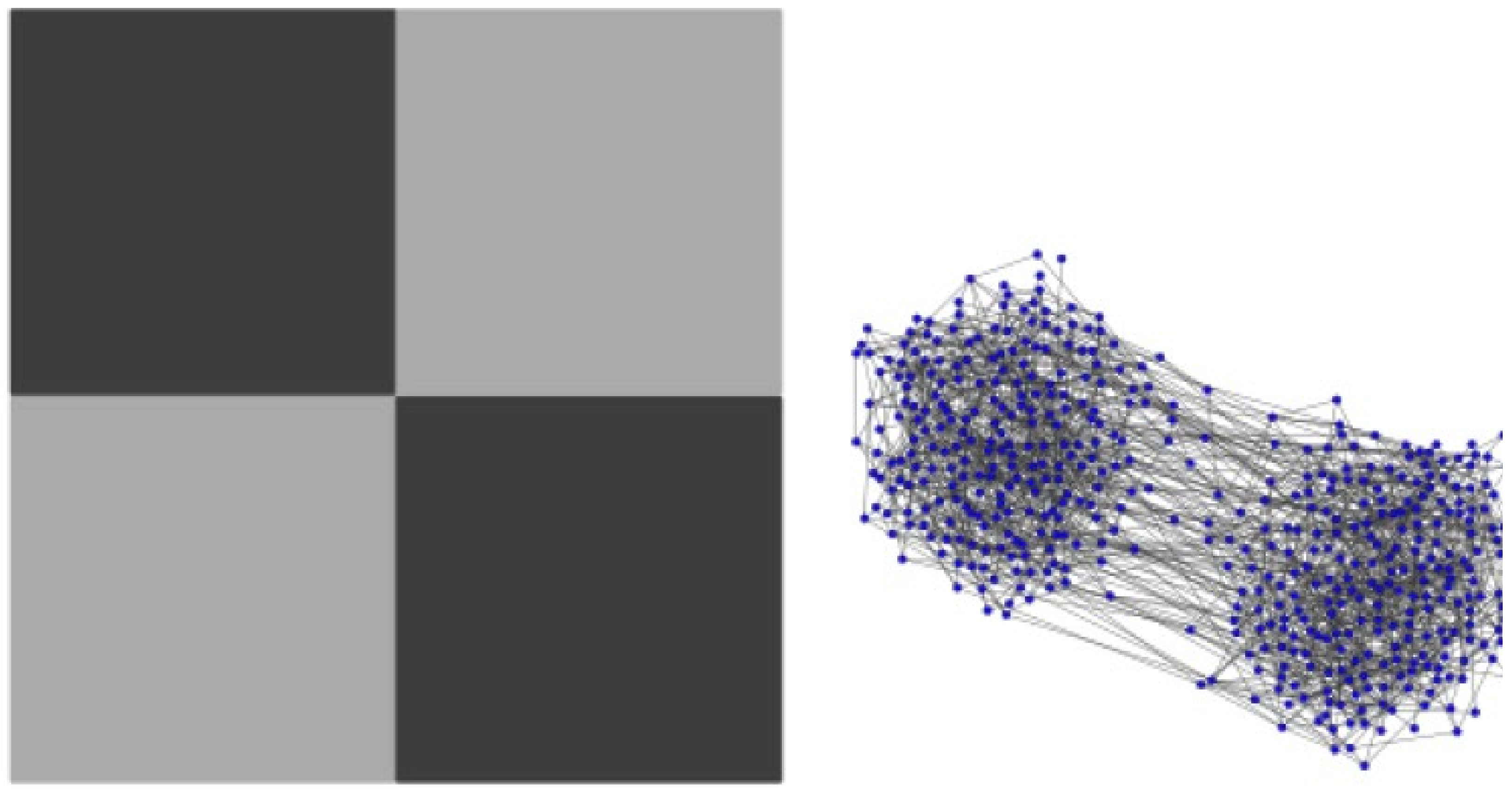
Stochastic Blockmodel - Estimation

- Everything looks very simple, but…
- Block-memberships are unknown, and need to be estimated!
- The complete data likelihood is untreatable

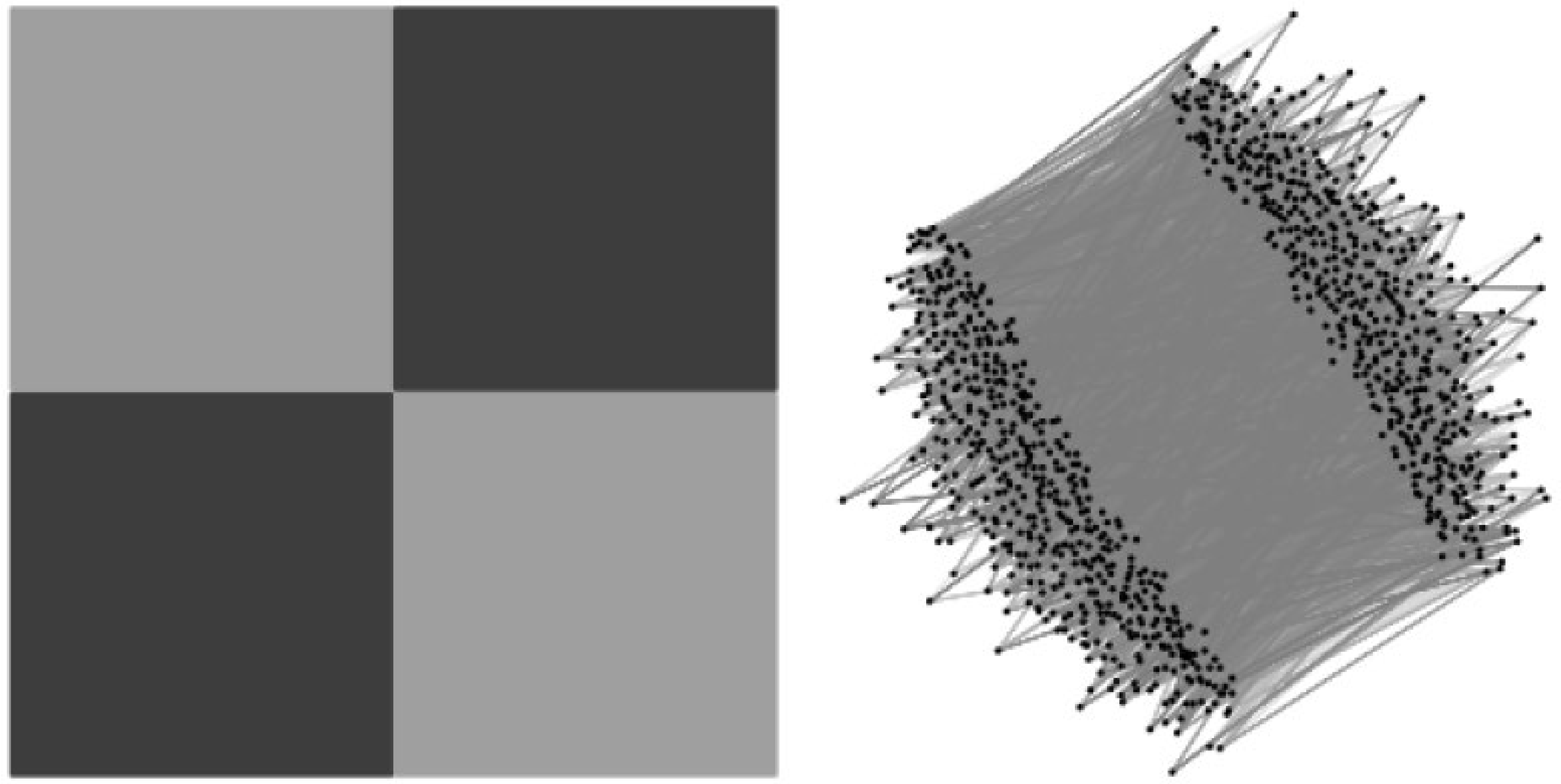
- Need to solve a complex estimation problem. Some routes:
  - Variational inference
  - Vertex-switching algorithms
  - MCEM algorithms
  - …

The great thing about the SBM

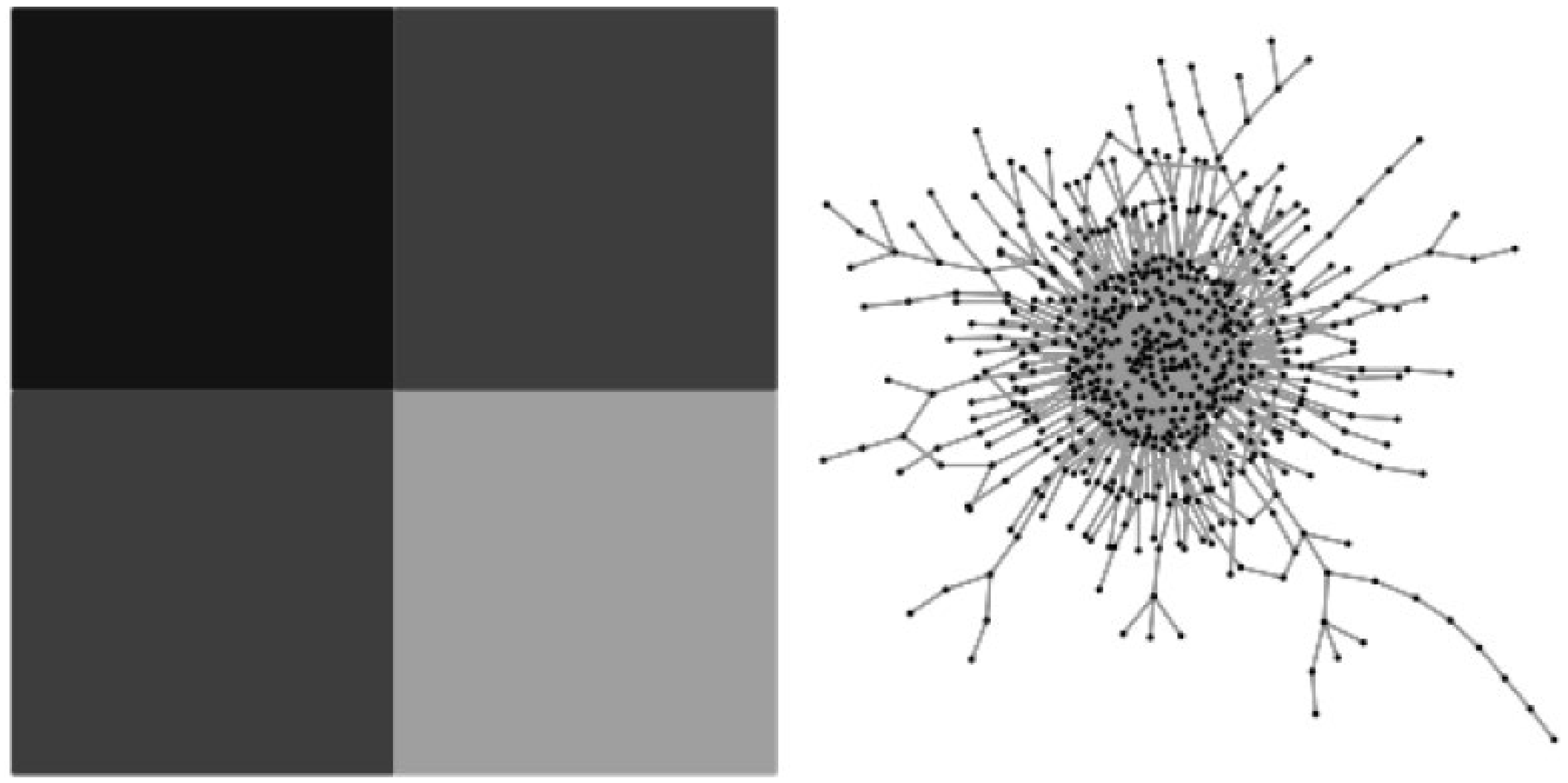
- Unlike pure “community detection” algorithms, able to find any type of structure, beyond classic communities



(a) Community structure.



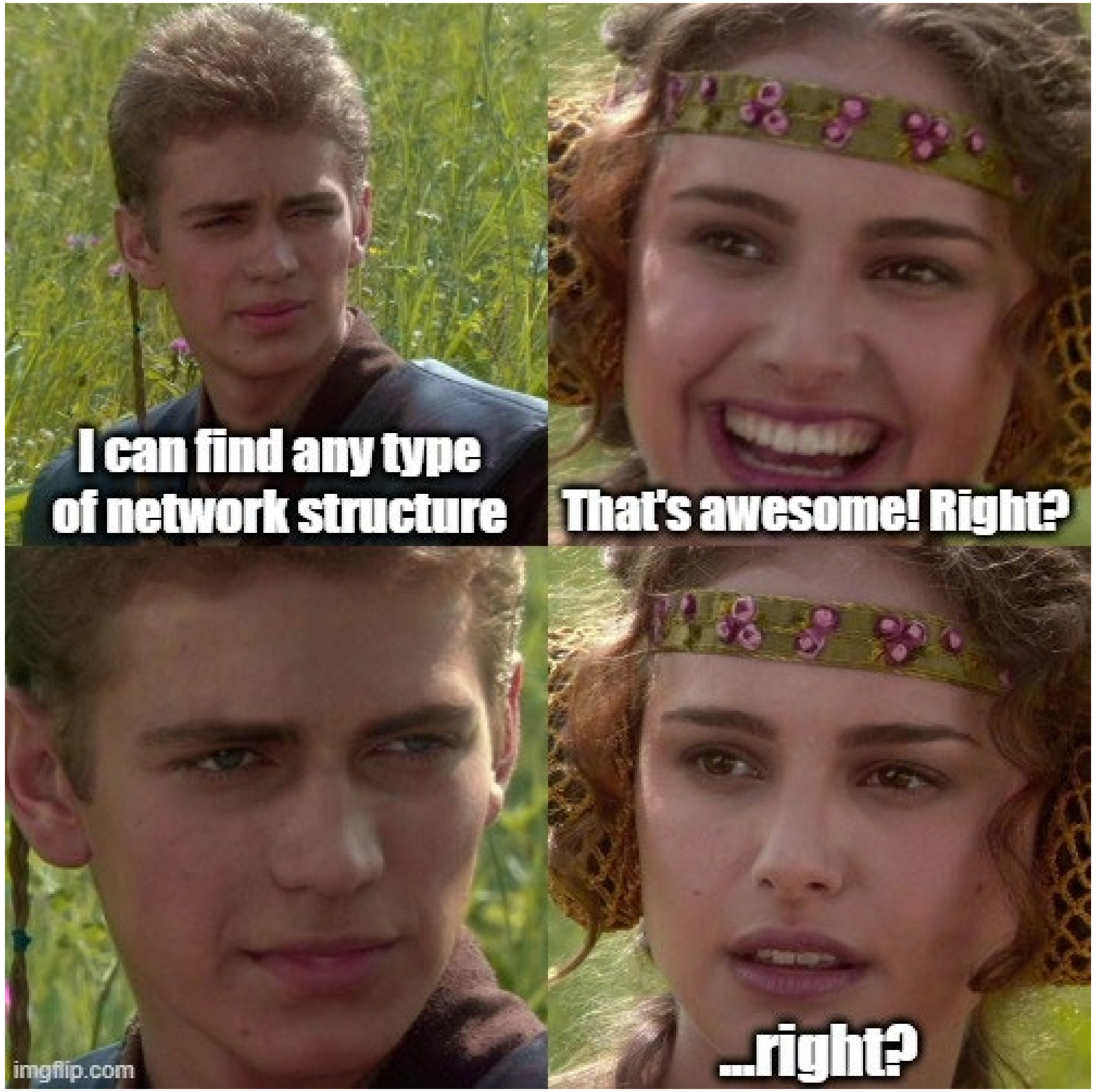
(b) Disassortative structure.



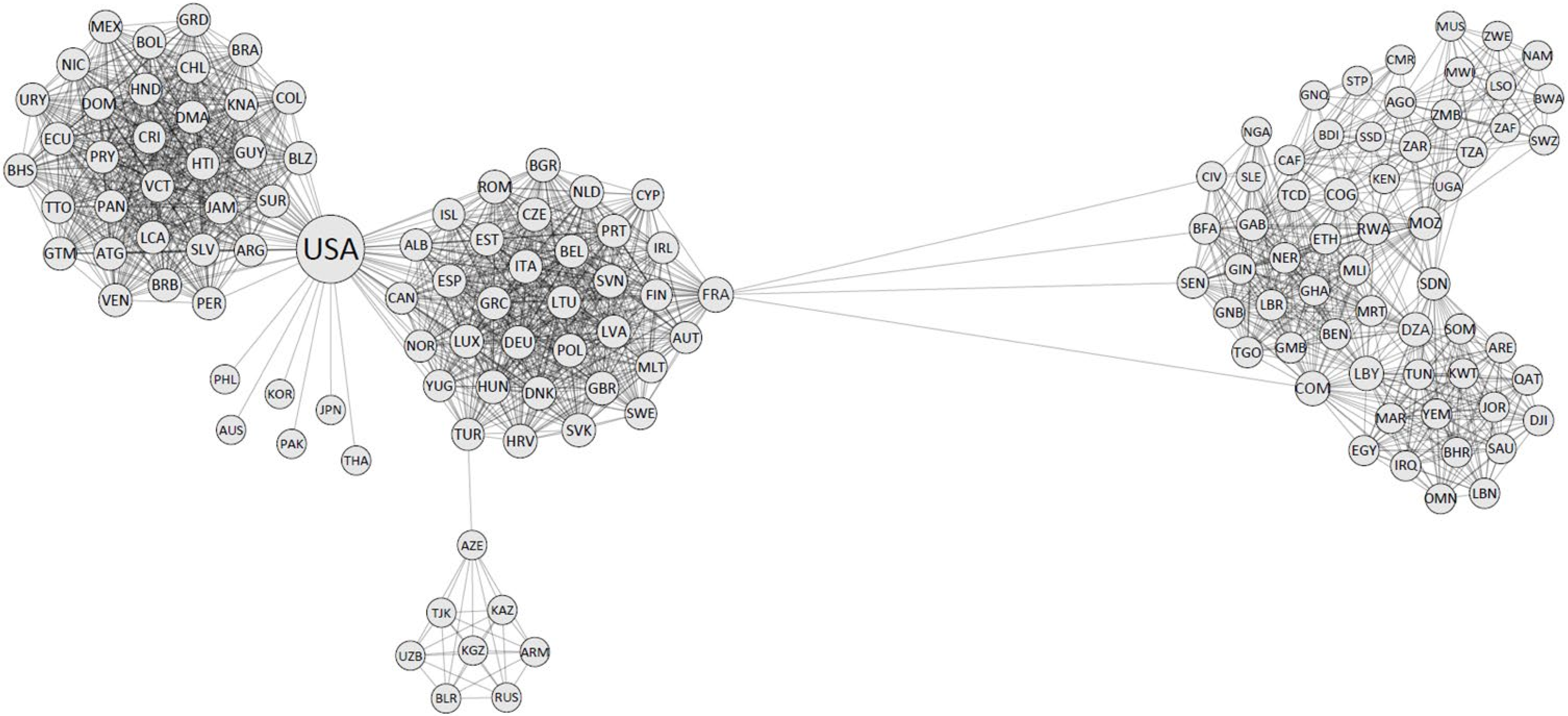
(c) Core–periphery structure.

From Fortunato & Hric, 2016

…but is it always a good thing?

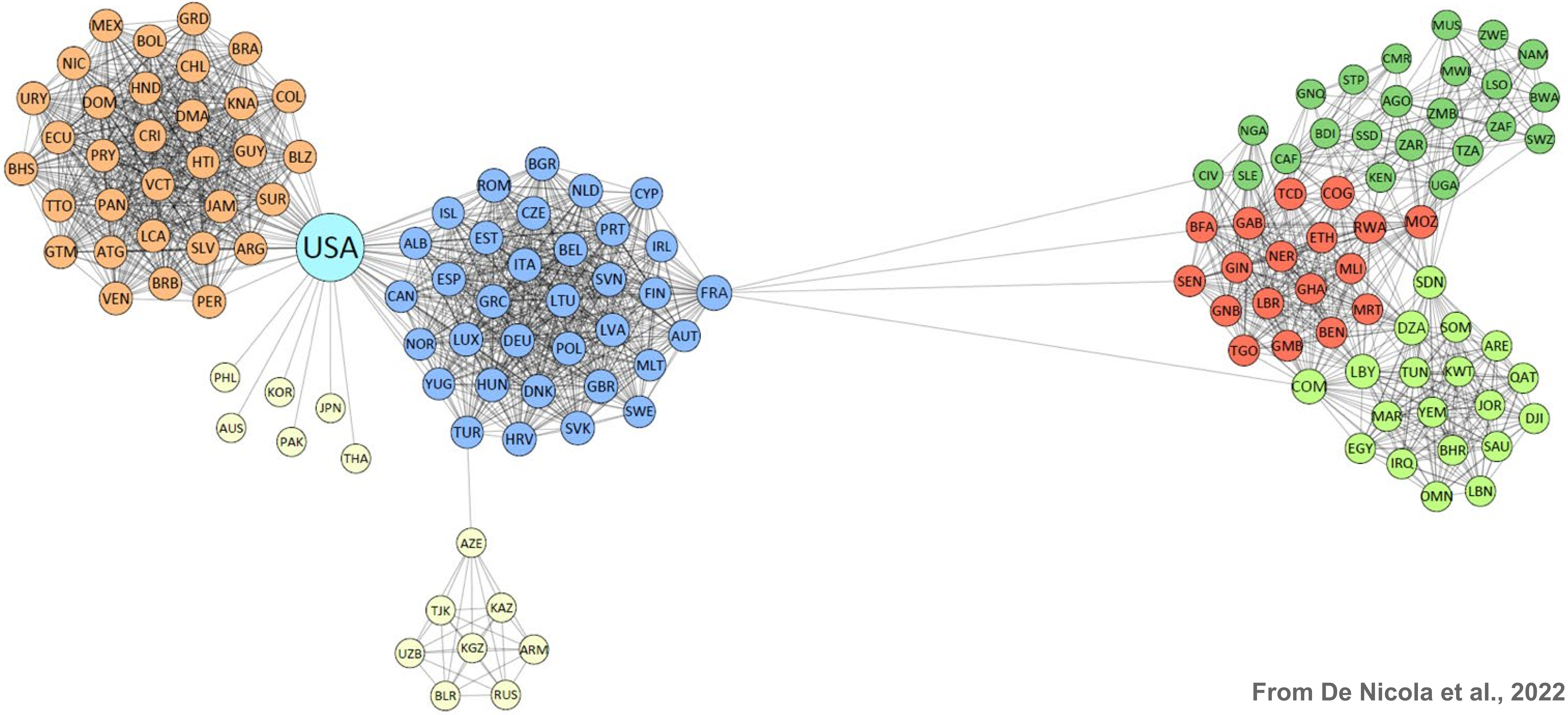


Example 1: Alliances Network



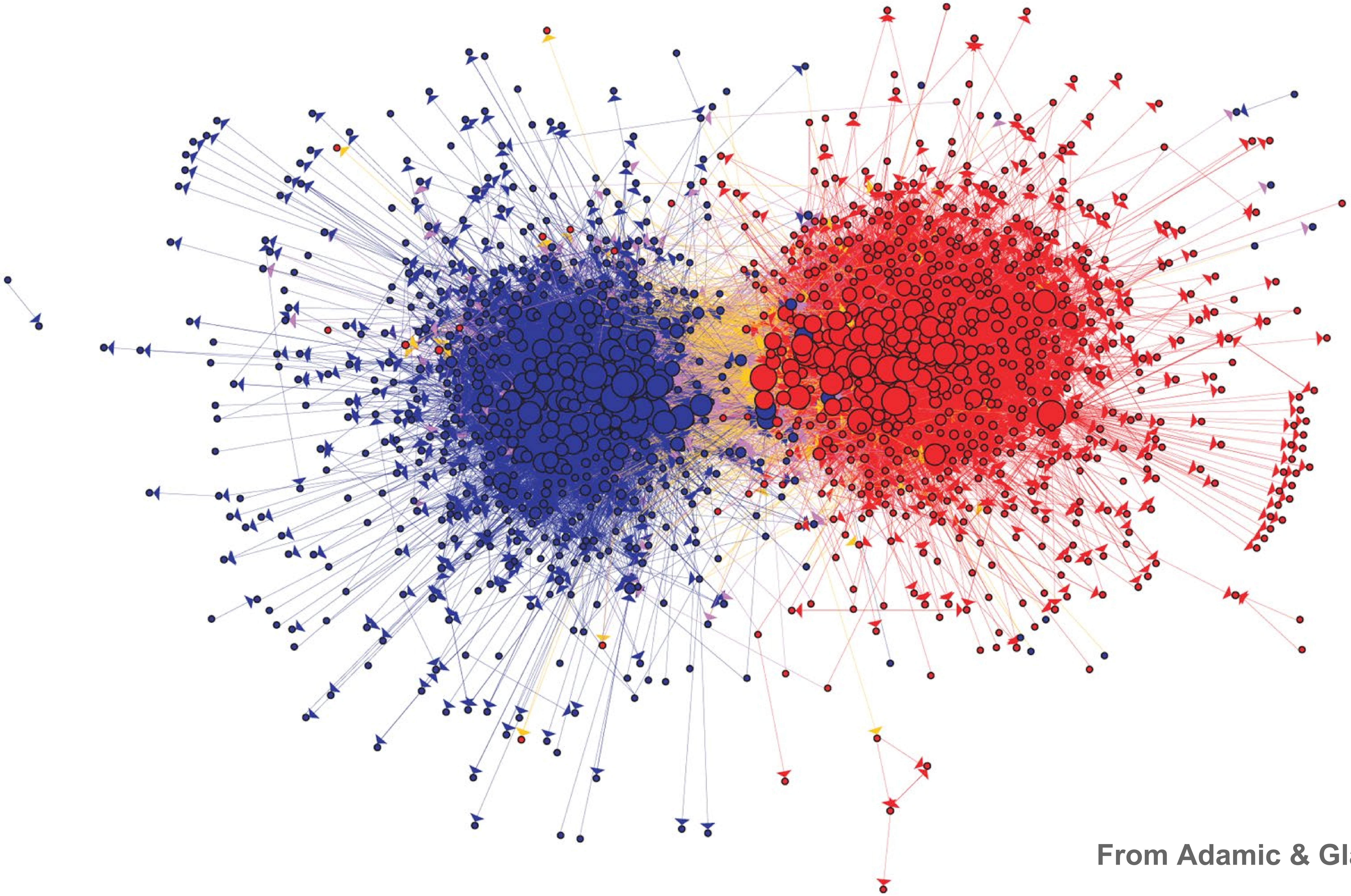
What will the SBM find? (K=7)

…(almost) classical community structure!



From De Nicola et al., 2022

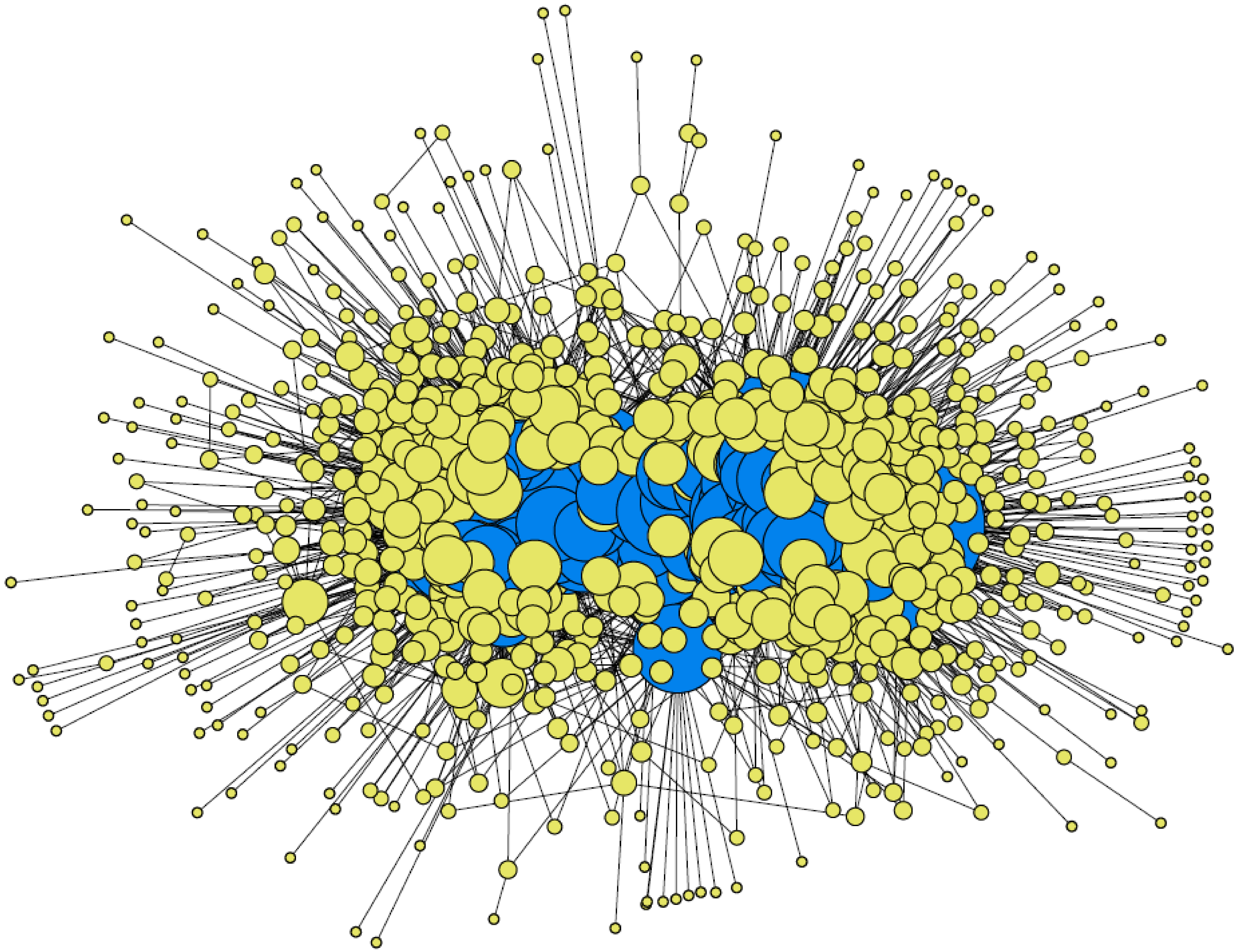
Example: Political blogs network



From Adamic & Glance, 2005

What will the SBM find? (K=2)

….a core-periphery structure?!?



From Karrer & Newman, 2012

The feature/bug of SBM for social networks

- The classical SBM implicitly assumes the degree structure *within blocks* to be relatively homogeneous
- But many real world social networks exhibit extremely skewed degree distributions
- This leads the SBM to very often find core-periphery structures, as opposed to classical assortative communities

From Karrer & Newman, 2012

Degree-corrected SBM

- Karrer & Newman (2012) introduced the idea of degree correction
- The probability of an edge depends not only on block-membership, but also explicitly on node-specific heterogeneity parameters (i.e. node degree):

From Karrer & Newman, 2012

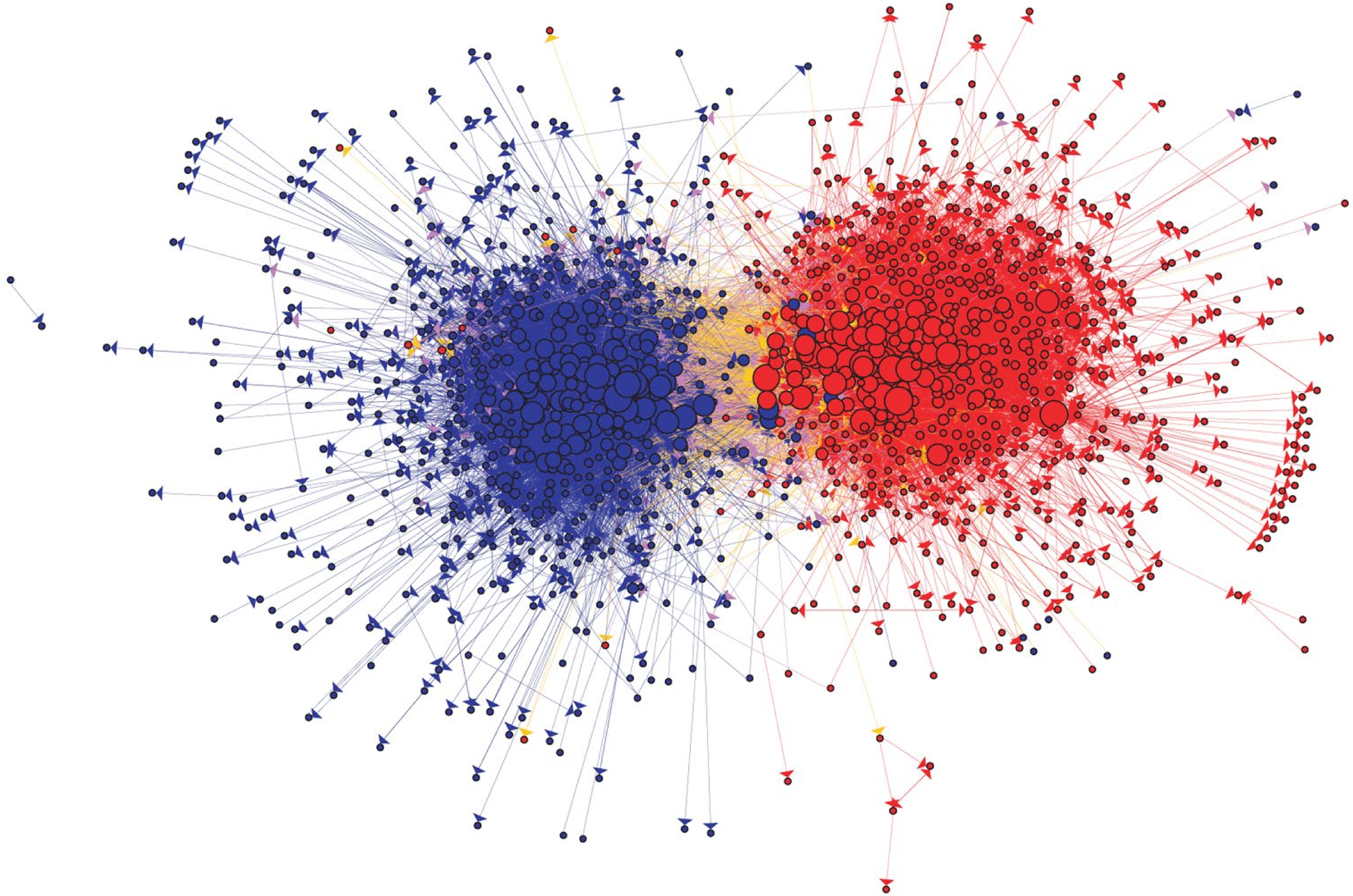
Degree-corrected SBM

- Karrer & Newman (2012) introduced the idea of degree correction
- The probability of an edge depends not only on block-membership, but also explicitly on node-specific heterogeneity parameters (i.e. node degree):

$$\lambda_{ij} = \exp\{\gamma_i + \gamma_j + \omega_{z_i z_j}\}$$

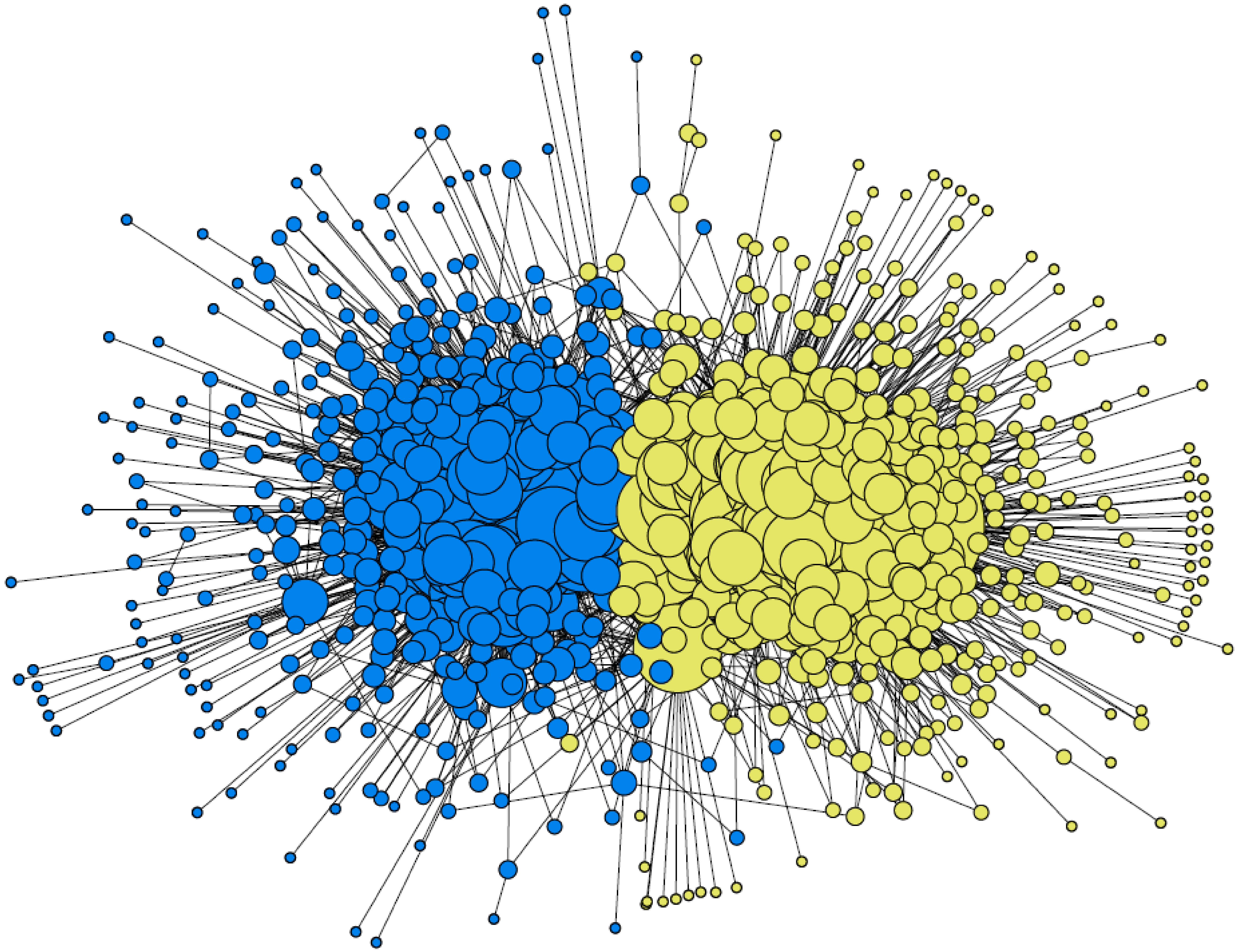
From Karrer & Newman, 2012

Back to political blogs



What will the degree-corrected SBM find?

…assortative communities!



From Karrer & Newman, 2012

SBM vs DCSBM

From Karrer & Newman, 2012

SBMs: Variants and Extensions

- Classical SBM is a very simple model, many other variants and extensions exist
- Variants aimed at finding specific types of network structures
- Some of the most prominent ones:
    - Mixed membership SBM (Airoldi et al., 2008)
    - Hierarchical SBM (Peixoto, 2012)
    - Mixture of experts SBM (Gormley & Murphy, 2010)

SBM as a model class: Features

- Good for finding different types of community structure
- Principle, likelihood based methods, with all the perks that come with it
- Relatively fast estimation routines exist
- A lot of software available openly available

SBM as a model class: Features

- Good for finding different types of community structure
- Principle, likelihood based methods, with all the perks that come with it
- Relatively fast estimation routines exist
- A lot of software available openly available

All in all, a solid tool for finding discrete structures in different types of networks

SBM as a model class: Limitations

- Discrete → too simplistic
- Not straightforward to include covariates
- Number of communities K needs to be inputted
    - Several ways to estimate it data-driven
    - Still requires some prior assumptions (far from being solved)

SBM as a model class: Limitations

- Discrete → too simplistic
- Not straightforward to include covariates
- Number of communities K needs to be inputted
    - Several ways to estimate it data-driven
    - Still requires some prior assumptions (far from being solved)

Can we address these?

Latent Space Models

Continuous Latent Variables

- It is quite natural to generalize the idea of discrete communities into continuous ones

- Hoff et al. (2002) propose to “map” the network into a Euclidean latent social space, where the distance between two nodes determines their probability of being connected

The Latent Distance Model

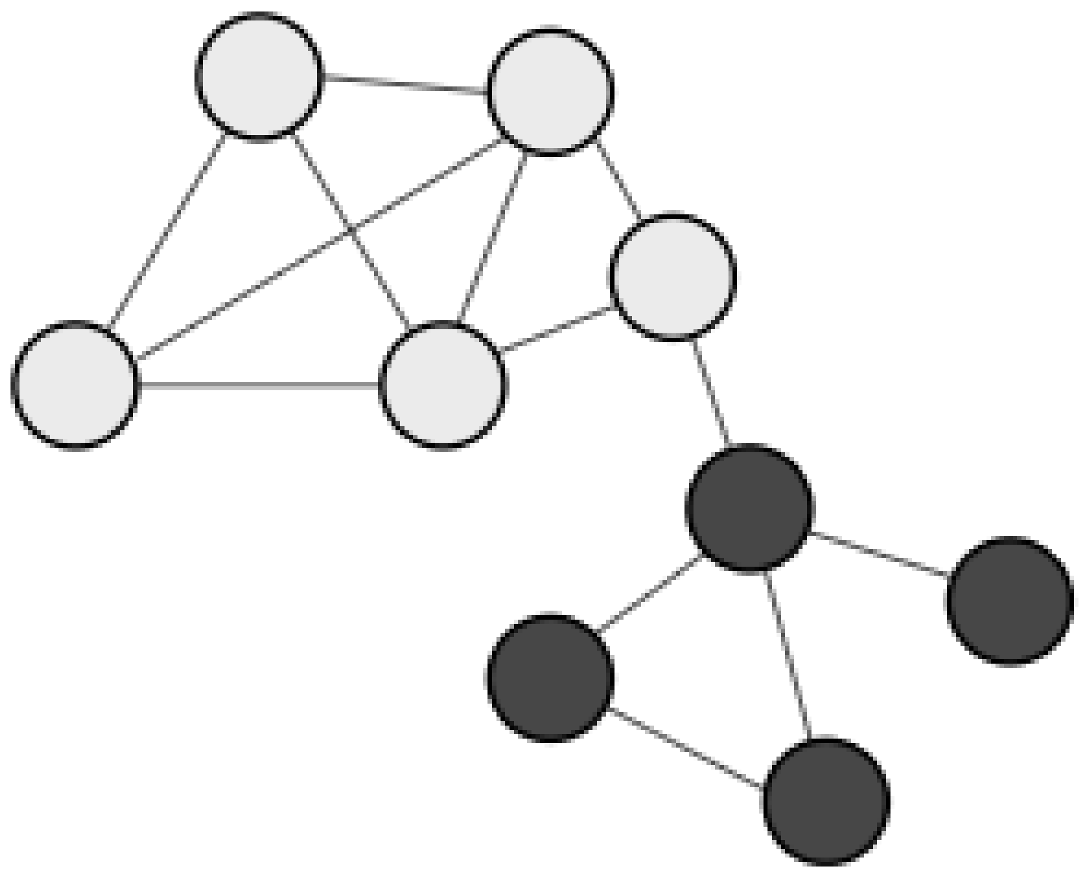
- Postulates that the actors are located in a latent social space
- The closer they are in this space, the more likely they are to connect
- Specifically, log-odds of a tie between nodes *i* and *j* given by:

$$\begin{aligned}\eta_{i,j} &= \log \text{odds}(y_{i,j} = 1 | z_i, z_j, x_{i,j}, \alpha, \beta) \\ &= \alpha + \beta' x_{i,j} - |z_i - z_j|.\end{aligned}$$

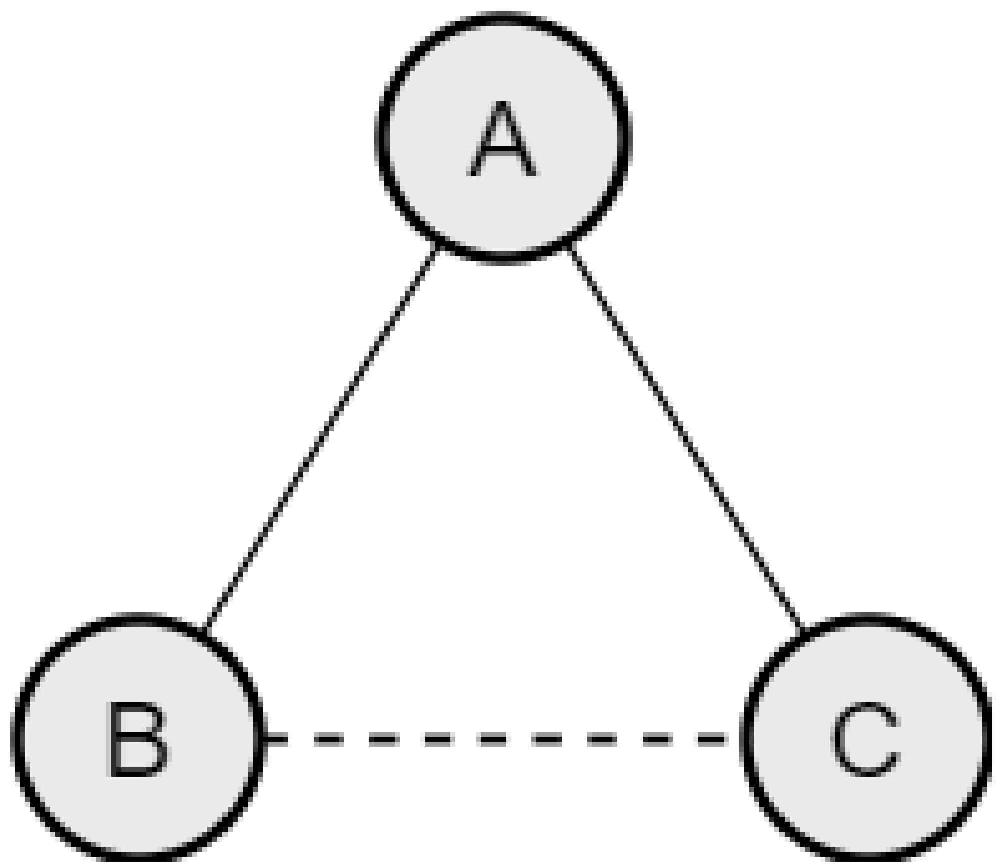
Properties

- Does a good job at representing patterns that are typical of social networks, such as:

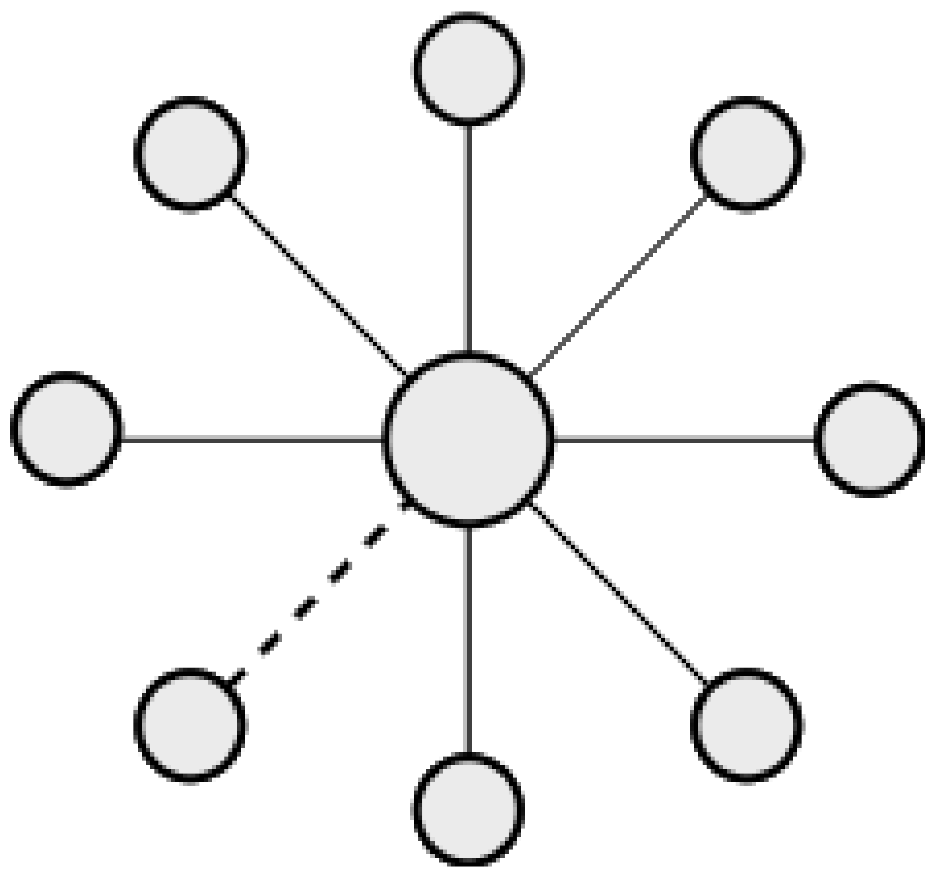
Homophily



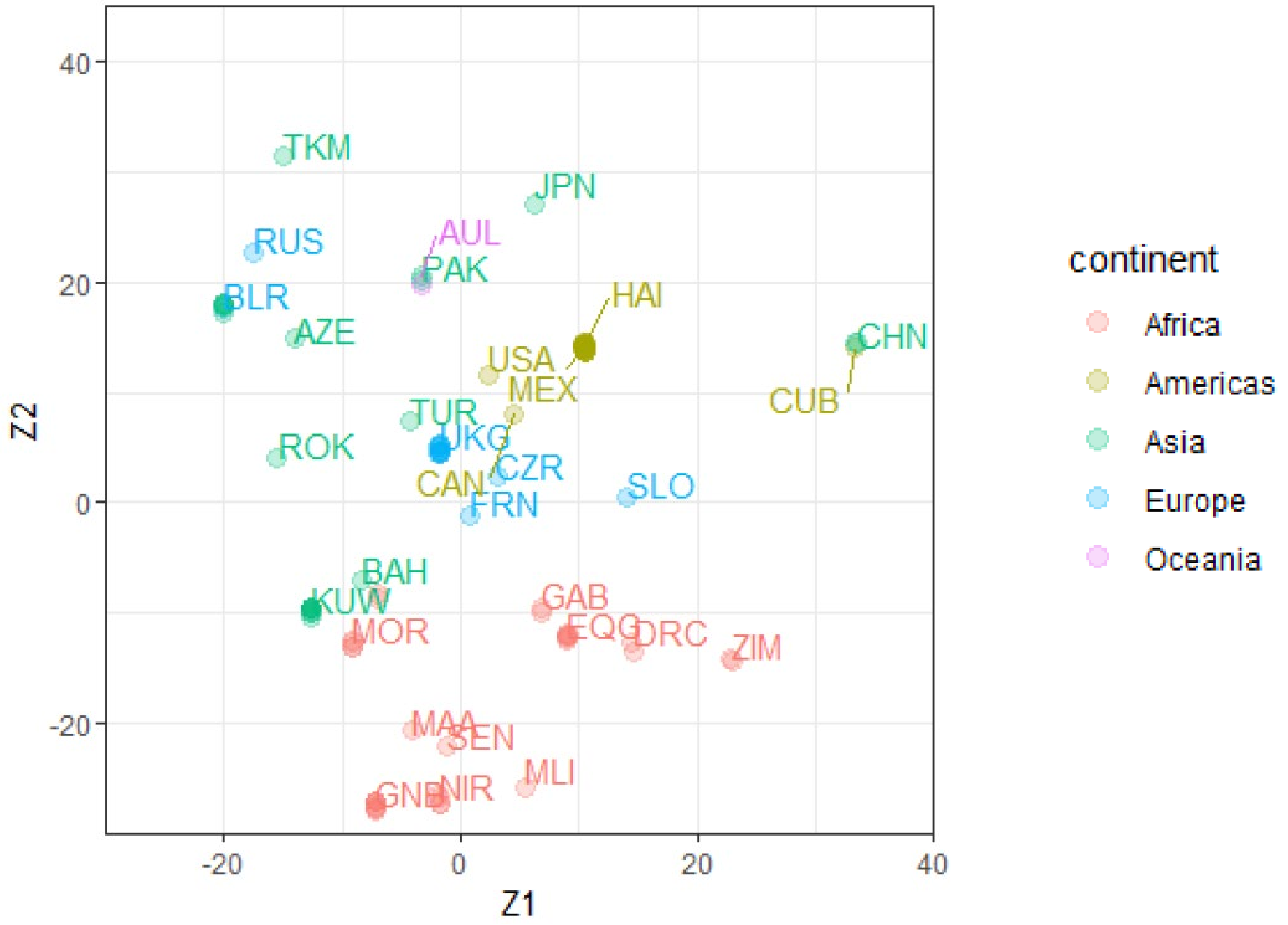
Triadic Closure



Preferential attachment



Example: Alliances network



The latent position cluster model

- We can allow for model-based clustering of the latent positions, to also get communities (see Handcock et al., 2007)
- Assume the positions to come from a mixture distribution:

$$Z_i \overset{\text{i.}i.d.}{\sim} \sum_{g=1}^{G} \lambda_g \, \mathrm{MVN}_d\left(\mu_g, \sigma_g^2 I_d\right) \quad i = 1, \ldots, n$$

Further extension

- We can also control for the actors' different propensity to form ties (see Krivitsky et al., 2009)
- Add node-specific random effects:

$$\eta_{i,j} = \sum_{k=1}^{p} \beta_k x_{k,i,j} - \|Z_i - Z_j\| + \delta_i + \gamma_j$$

Further extension

- We can also control for the actors' different propensity to form ties (see Krivitsky et al., 2009)
- Add node-specific random effects:

$$\eta_{i,j} = \sum_{k=1}^{p} \beta_k x_{k,i,j} - \|Z_i - Z_j\| + \delta_i + \gamma_j, \quad \text{with}$$

$$\delta_i \overset{\text{i.}i.d.}{\sim} \mathrm{N}(0, \sigma_\delta^2) \quad i = 1, \ldots, n$$

$$\gamma_i \overset{\text{i.}i.d.}{\sim} \mathrm{N}(0, \sigma_\gamma^2) \quad i = 1, \ldots, n$$

Network of COVID-19 Twitter elites

- Start from database with all tweets about COVID-19
- Rank tweets by their popularity (likes + retweets + replies)
- A user is “elite” if they have a tweet on COVID-19 with popularity > 2000
- Result: 1024 tweets by a total of 363 users

Application to COVID-19 Twitter elites

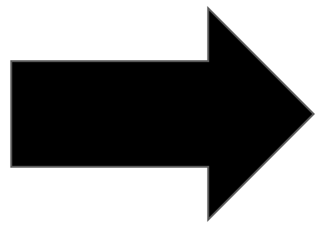
- Start from database with all tweets in German about COVID-19
- Rank tweets by their popularity (likes + retweets + replies)
- A user is “elite” if they have a tweet on COVID-19 with popularity > 2000
- Result: 1024 tweets by a total of 363 users

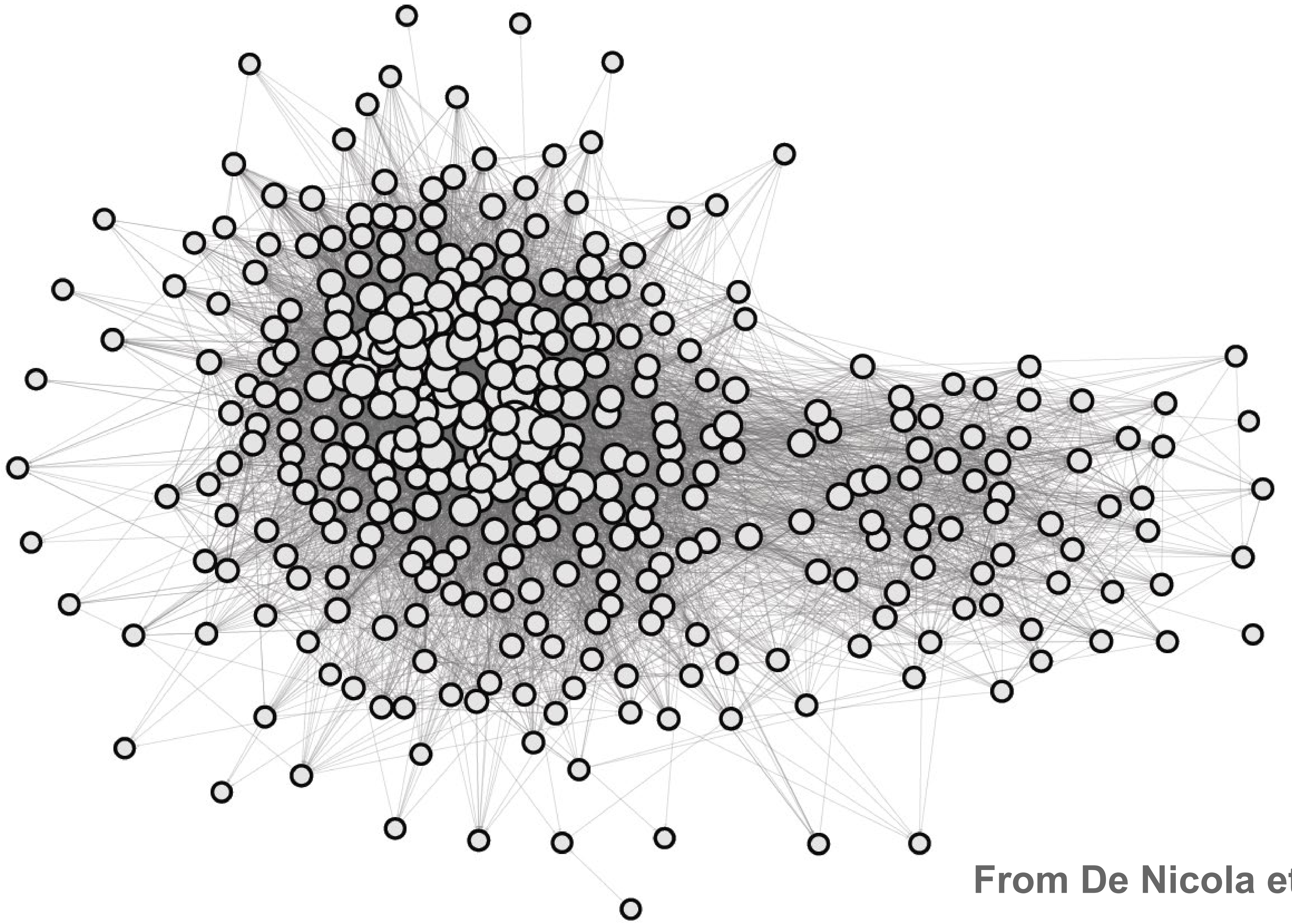
text	author	tweet_popularity
Wir haben keinen einzigen #COVID19 Pati ...	Ricardo Lange	29,422
Der Bundesgesundheitsminister fordert so ...	Jens Clasen	25,852
Über Freiheit und Eigenverantwortung spr ...	Dunja Hayali ❤️🇩🇪🇪🇺🧠	25,832
Kosten einer BioNTech-Impfdosis: 19,95K ...	Krankenpflegel	25,725
Das Letzte, was das Coronavirus sieht, b ...	Fabian Köster	25,205
"Der Weg hierher und hier raus ist ein h ...	Christian Drosten	21,368
Wir stecken tief in der Schuld unserer P ...	Prof. Karl Lauterbach	21,208
Echt stark, wie gut wir Covid-19 im Grif ...	Cornelius W. M. Oettle	20,367
(1) Nachdem ich mich heute bei der dpa z ...	Carsten Watzl	19,434
Um das noch einmal ganz klar zu sagen: ...	Jens Clasen	19,415

Network of COVID-19 Twitter elites

- We naturally define an edge from user A to user B if A follows B on Twitter
- Resulting network **of 363 users has** 12182 directed edges **(9.2% density)**

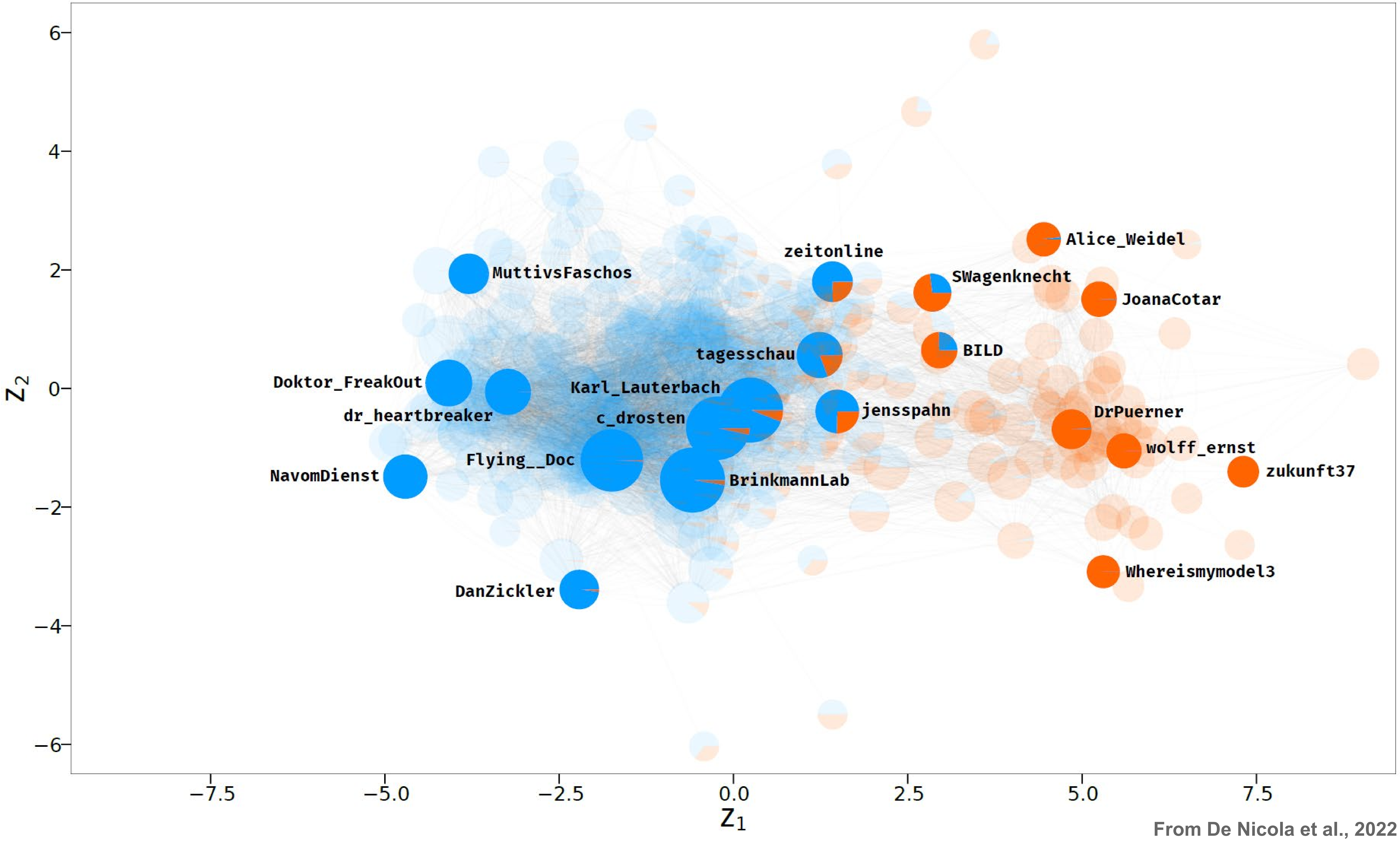
SENDER	RECEIVER
c_drosten	Karl_Lauterbach
Karl_Lauterbach	c_drosten
jensspahn	c_drosten
BrinkmannLab	Flying__Doc
Alice_Weidel	JoanaCotar
……	……





From De Nicola et al., 2022

The latent social space of COVID-19 elites



From De Nicola et al., 2022

LSM as a model class: Features

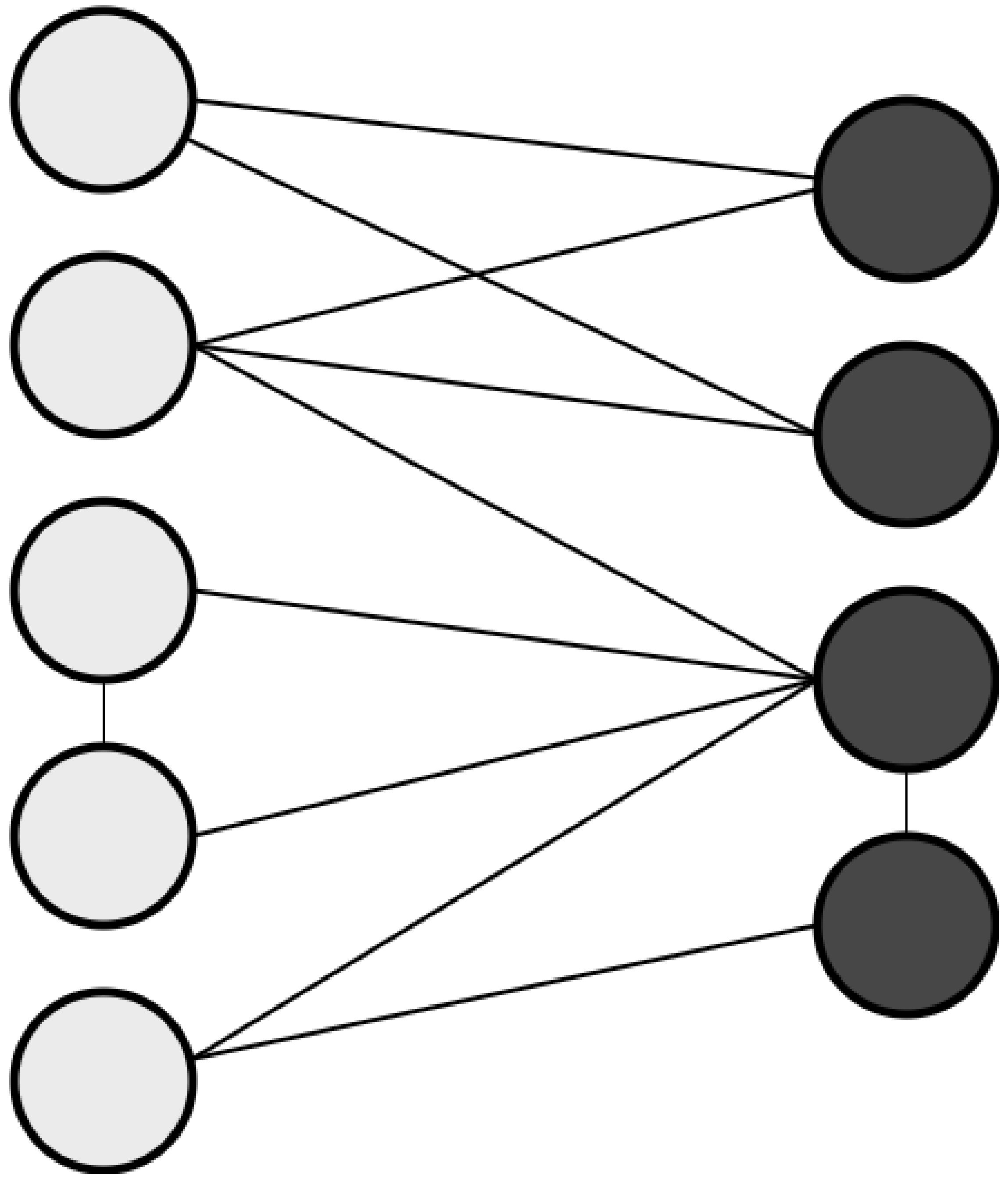
- Good for a more nuanced view than discrete SBM
- In our example: Beyond polarization on social media
- Great for graphical representation: positions have a probabilistic meaning
- Uncertainty quantification
- Possible to incorporate covariates (but interpretation changes greatly)

LSM as a model class: Chief limitation

- Fails at capturing disassortative structures

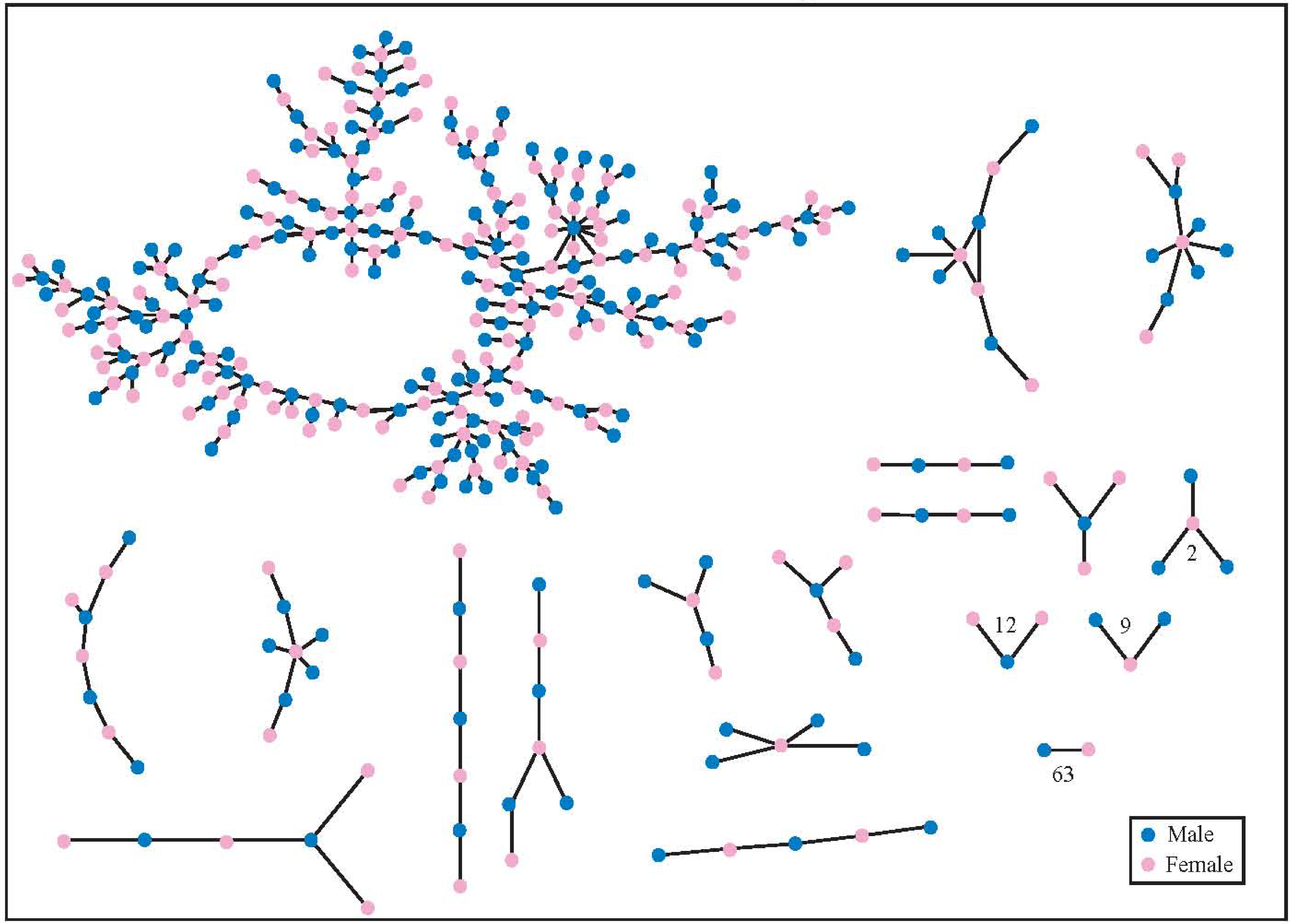
LSM as a model class: Chief limitation

- Fails at capturing disassortative structures
  - **Example:**



A heterophilic networks

The Structure of Romantic and Sexual Relations at "Jefferson High School"



Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).

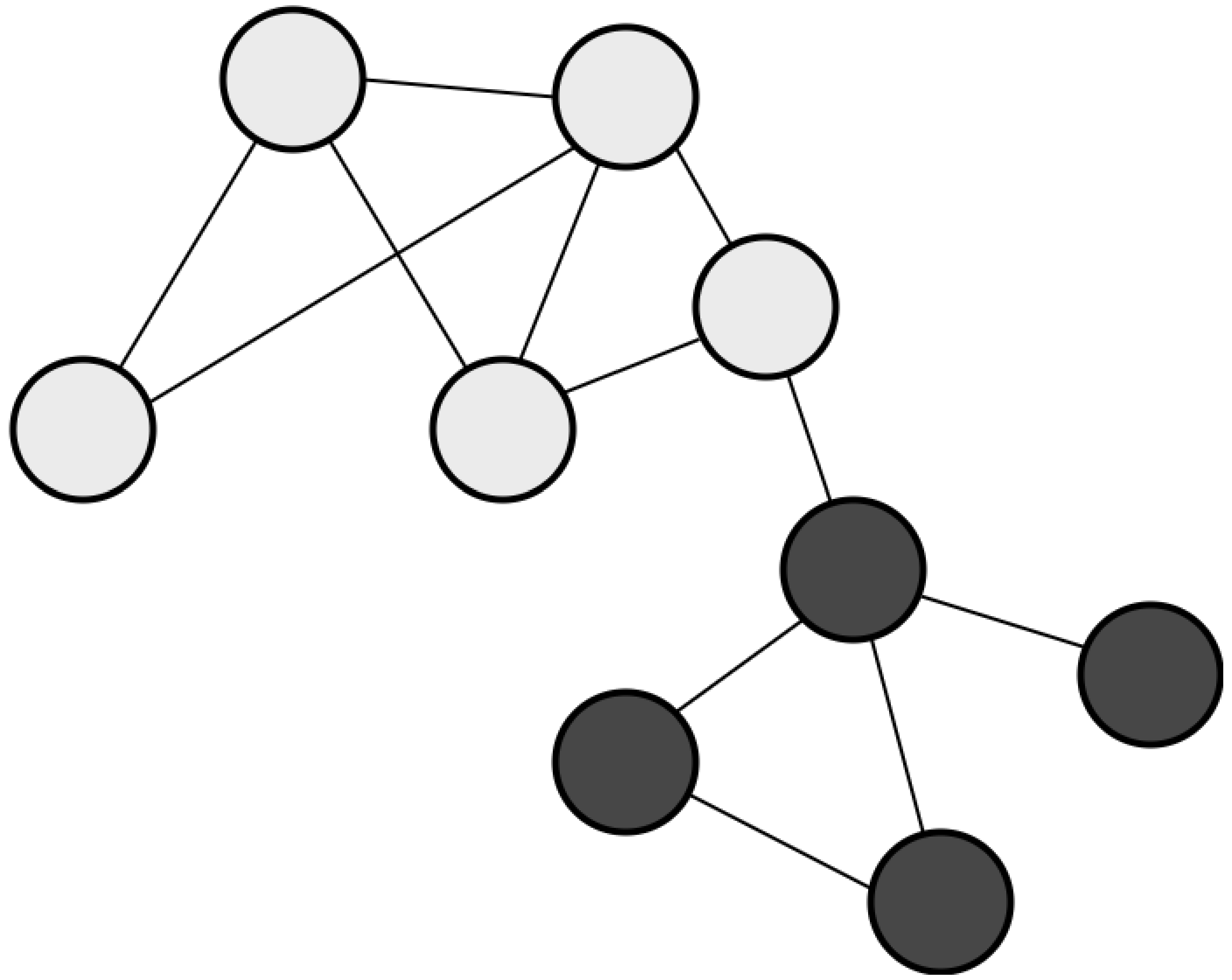
From Bearman et al., 2004

LSM: Terms of use

- Best to use LSMs only when reasonable to believe that the network is (mostly) homophilic and triadic
- That is not always known a priori
- Real world network can also display mixed patterns:

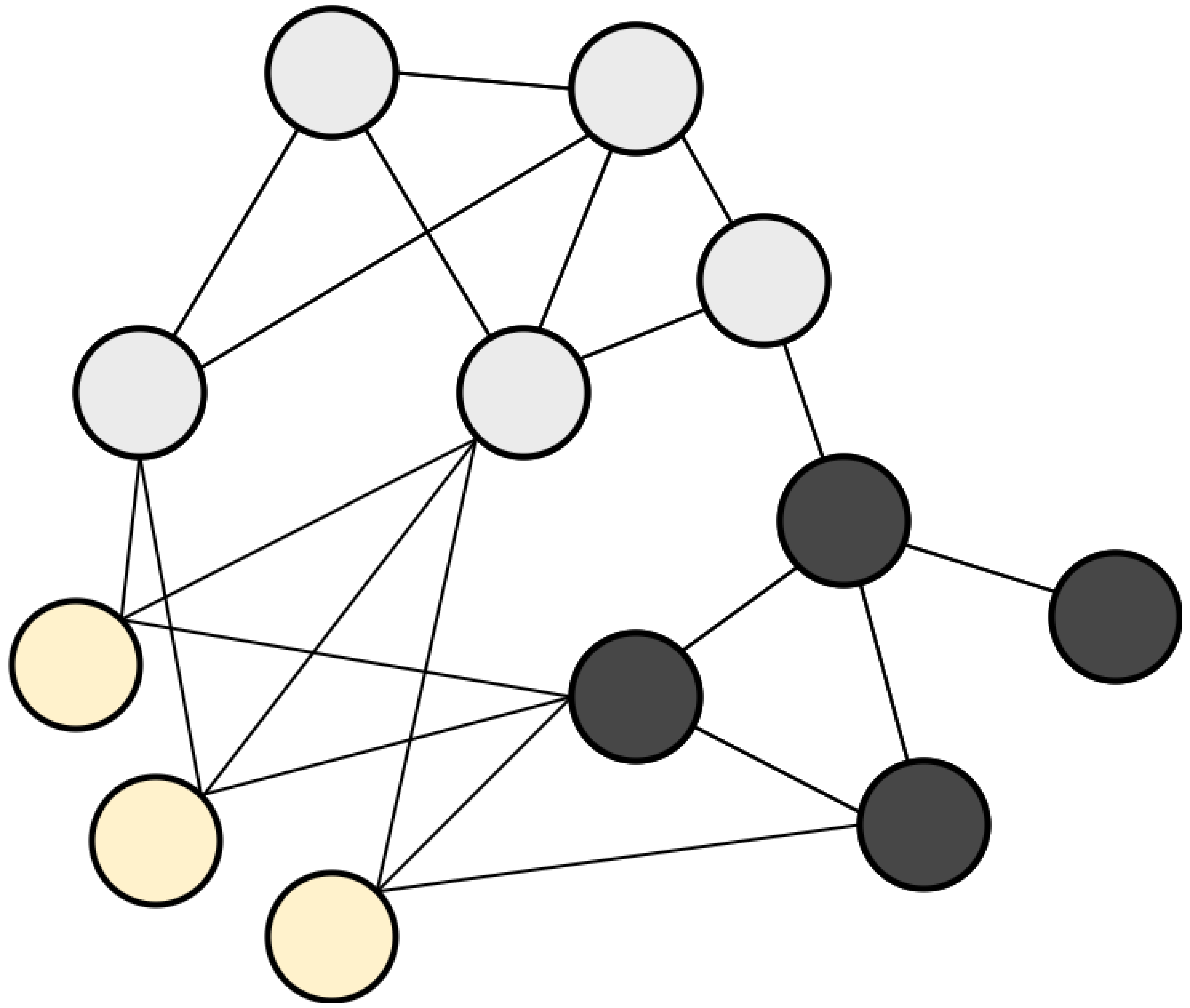
LSM: Terms of use

- Best to use LSMs only when reasonable to believe that the network is (mostly) homophilic and triadic
- That is not always known a priori
- Real world network can also display mixed patterns:



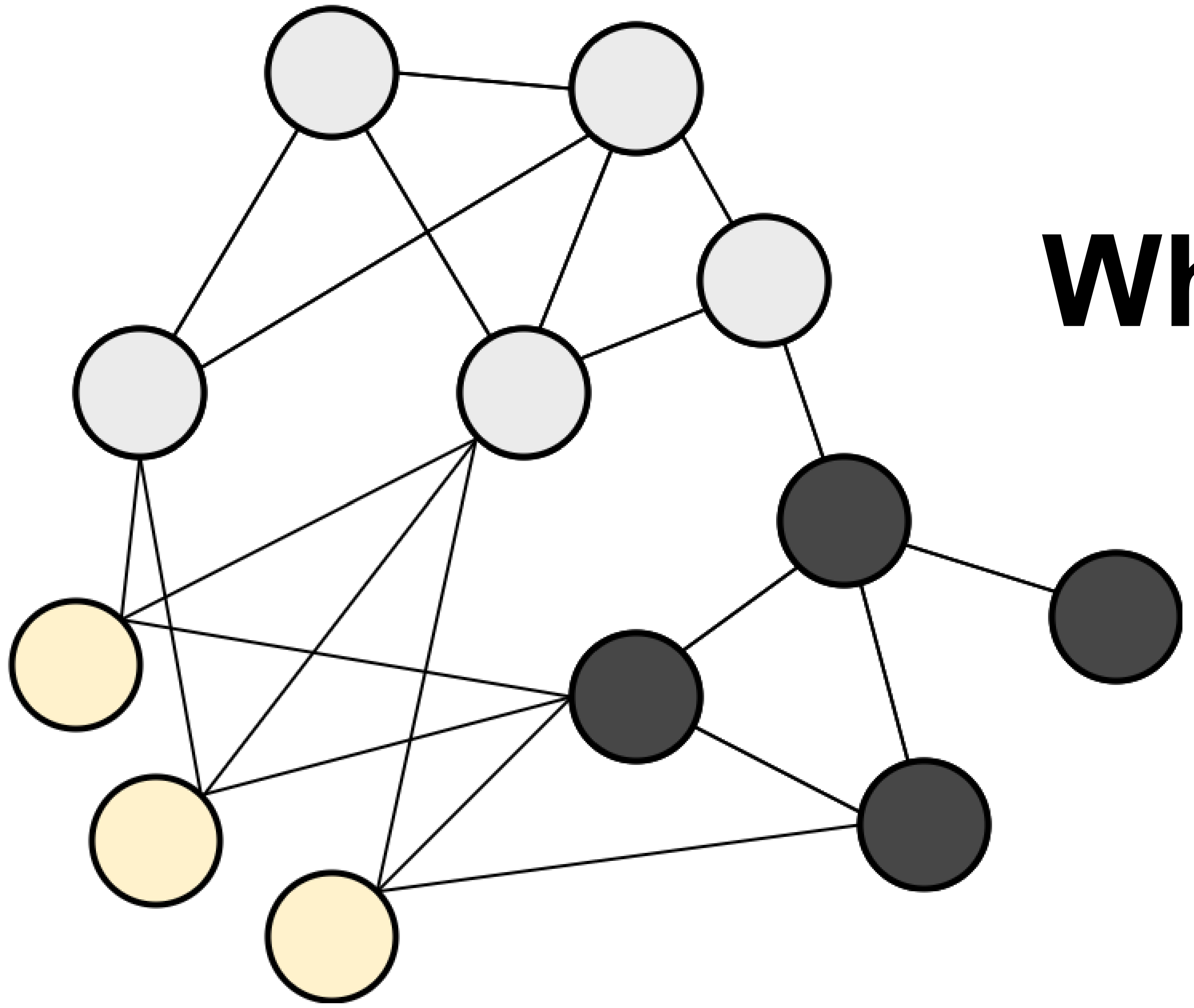
LSM: Terms of use

- Best to use LSMs only when reasonable to believe that the network is (mostly) homophilic and triadic
- That is not always known a priori
- Real world network can also display mixed patterns:



LSM: Terms of use

- Best to use LSMs only when reasonable to believe that the network is (mostly) homophilic and triadic
- That is not always known a priori
- Real world network can also display mixed patterns:



What can we do about this?

Additive and Multiplicative Effect Models

AME: Motivation and framework

- Network data often exhibit dependencies of different orders:
  - First order: Node-specific heterogeneity
  - Second order: Reciprocity
  - Third order: Triadic effects
- The AME network model (Hoff, 2021) is designed to capture all of these type of dependencies simoultaneously.

The AME Network Model

- The AME Probit model specifes the probability of a tie as:

$$\mathbb{P}(Y_{ij} = 1|W) = \boldsymbol{\Phi}(\theta^\top x_{ij} + e_{ij}),$$

- Where:
  - $\boldsymbol{\Phi}$ is the standard normal cumulative distribution function
  - $\theta^\top x_{ij}$ accommodates the inclusion of covariates
  - e_{ij} can be viewed as a structured residual

The AME Network Model

- The structured error e_{ij} is a function of the latent variables:

$$e_{ij} = a_i + b_j + u_i v_j + \varepsilon_{ij}$$

The AME Network Model

- The structured error e_{ij} is a function of the latent variables:

$$e_{ij} = a_i + b_j + u_i v_j + \varepsilon_{ij}$$

- a_i and b_j are zero-mean additive effects for sender i and receiver j, which account for first order dependencies
  - More specifically:

$$(a_1, b_1), \ldots, (a_n, b_n) \overset{\text{i.i.d.}}{\sim} N_2(0, \Sigma_1), \quad \text{with} \quad \Sigma_1 = \begin{pmatrix} \sigma_a & \sigma_{ab} \\ \sigma_{ab} & \sigma_b \end{pmatrix}$$

The AME Network Model

- The structured error e_{ij} is a function of the latent variables:

$$e_{ij} = a_i + b_j + u_i v_j + \varepsilon_{ij}$$

The AME Network Model

- The structured error e_{ij} is a function of the latent variables:

$$e_{ij} = a_i + b_j + u_i v_j + \varepsilon_{ij}$$

- ε_{ij} is a zero-mean residual term, accounting for second order dependency, i.e. reciprocity
  - More specifically:

$$\{(\varepsilon_{ij}, \varepsilon_{ji}) : i < j\} \overset{\text{i.i.d.}}{\sim} N_2(0, \Sigma_2), \quad \text{with} \quad \Sigma_2 = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

The AME Network Model

- The structured error e_{ij} is a function of the latent variables:

$$e_{ij} = a_i + b_j + u_i v_j + \varepsilon_{ij}$$

The AME Network Model

- The structured error e_{ij} is a function of the latent variables:

$$e_{ij} = a_i + b_j + u_i v_j + \varepsilon_{ij}$$

- u_i and v_j are d-dimensional multiplicative “latent positions” accounting for third order dependencies, with

$$(u_1, v_1), \ldots, (u_n, v_n) \sim \mathcal{N}_{2d}(0, \Sigma_3)$$

AME: Pros and cons

Advantages:

- Incredibly flexible, able to represent many network structures
- Has been shown to generalize both SBM and LSM

Disadvantages:

- Incredibly complex, estimation slow
- Multiplicative latent space is not as interpretable nor good for representation as the LSM

AME: How to use

- Given it’s complexity AME is a suboptimal choice when focus is on interpretability and visualization of the latent structure
- To the contrary, it is an ideal fit when underlying network dependencies are unkown, and the focus is on estimating covariate effects controlling for the network structure
- Many such cases, especially in the social sciences

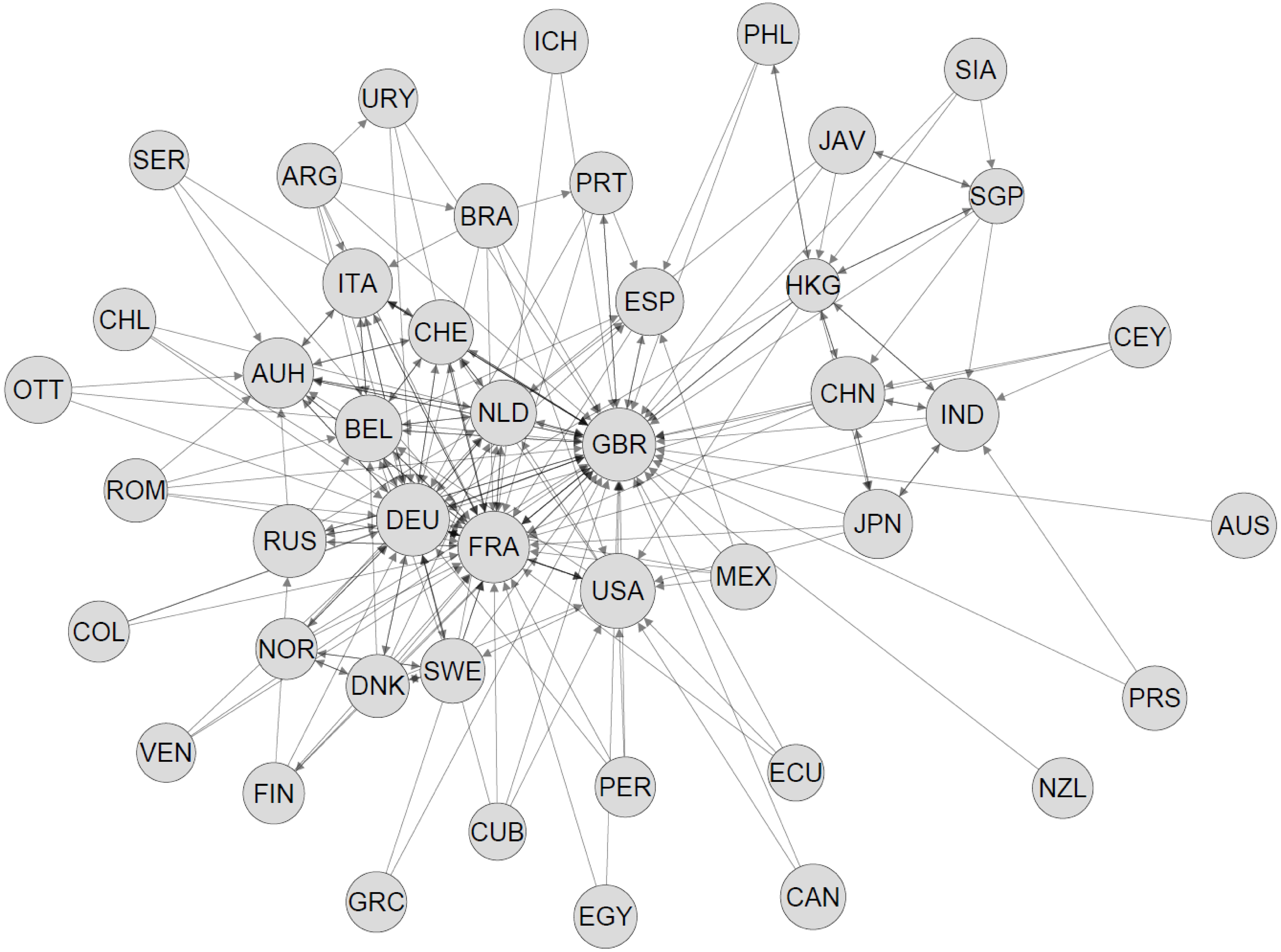
Application: Historical forex network

- In 1900, every financial center featured a foreign exchange market were bankers bought and sold foreign currency against the domestic one.
- Foreign exchange market activity was monitored in local bulletins, which allowed Flandreau and Jobst (2005) to collect a global dataset.

Application: Historical forex network

- In 1900, every financial center featured a foreign exchange market were bankers bought and sold foreign currency against the domestic one.
- Foreign exchange market activity was monitored in local bulletins, which allowed Flandreau and Jobst (2005) to collect a global dataset.
- This gives rise to a directed network, where:
  - Countries are nodes
  - A directed edge i → j is present if currency from country j is traded within the financial center of country i

Historical forex network



Application: Historical forex network

- Interested in understanding the effect of several factors on currency trade between countries, which is an important indicator of economic influence.
- **Nodal covariates:**
  - Gold standard, GDP per-capita, democracy index
- **Dyadic covariates:**
  - Distance, reciprocal trade volume

Historical forex network

		AME	Classical Probit
	Intercept	−4.845 (5.310)	−3.211 (1.580)*
Sender	Gold standard	−0.629 (0.397)	−0.354 (0.155)*
	log-GDP per-capita	−0.453 (0.419)	−0.259 (0.152)
	Democracy index	−0.033 (0.064)	−0.025 (0.026)
	Currency coverage	1.418 (0.405)***	0.470 (0.137)***
Receiver	Gold standard	−0.599 (0.667)	−0.468 (0.191)*
	log-GDP per-capita	0.426 (0.703)	0.240 (0.159)
	Democracy index	0.121 (0.102)	0.066 (0.019)***
	Currency coverage	2.734 (0.691)***	1.363 (0.181)***
Dyadic	Distance	−1.019 (0.151)***	−0.471 (0.064)***
	log-trade volume	0.488 (0.081)***	0.346 (0.036)***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Takeaways

- **Dependencies matter!**
- Latent variable models are one way to measure them and/or take them into account
- There is no single best: different networks require different approaches
- Different research questions also are answered with different models

Latent Variable Models: Overview

- **Stochastic blockmodels** are good to find group structures of different kinds, and can capture stochastic equivalence

- **Latent distance models** are great for representing and understanding networks in which nodes that behave similarly tend to connect to each other frequently

- **AME models** are optimal when the focus is on measuring the effect of exogenous covariates, while controlling for the network

Thank you for your attention!

Funded by the Deutsche Forschungsgemeinschaft
(DFG, German Research Foundation) – 460037581

BERD Academy

Thank you for your attention!

Questions?

Funded by the Deutsche Forschungsgemeinschaft
(DFG, German Research Foundation) – 460037581