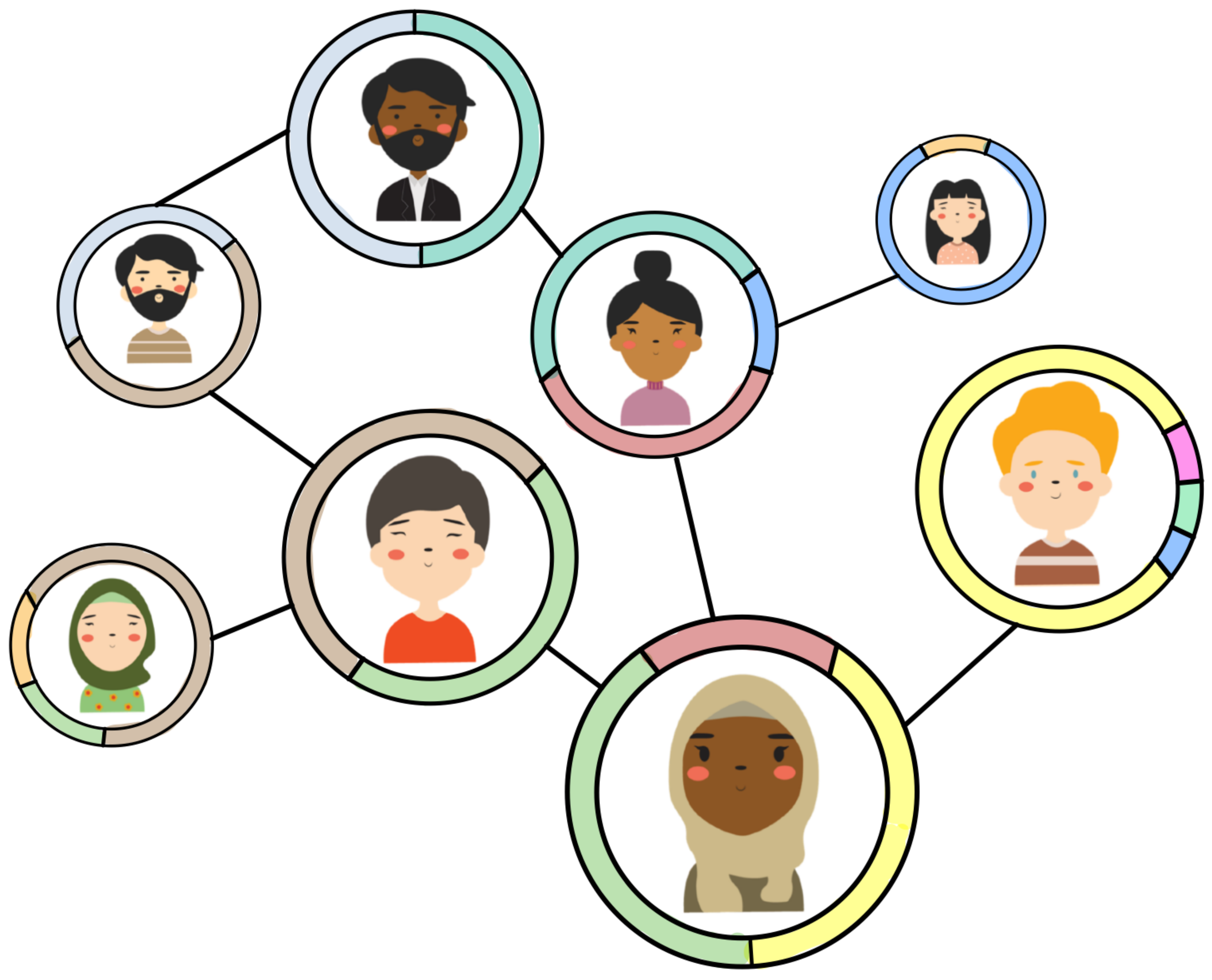
Thinned random measures for sparse networks with overlapping communities

Federica Zoe Ricci

University of California Irvine



Overview

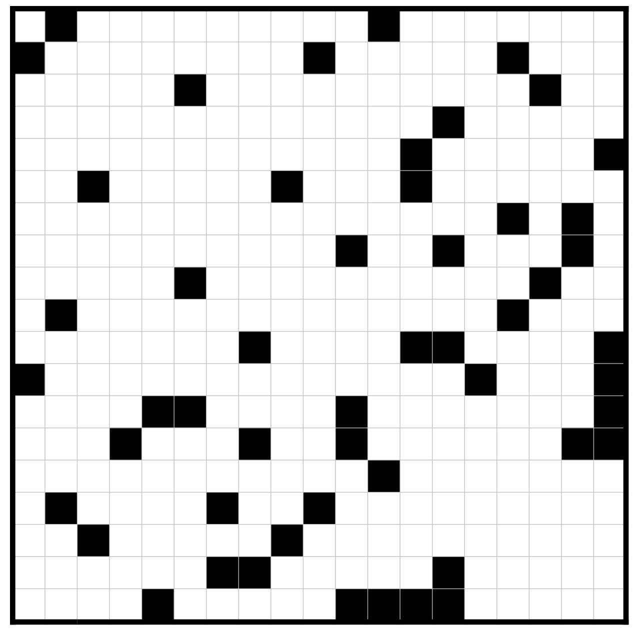
Motivation → Model → Results → Discussion

Overview

Motivation

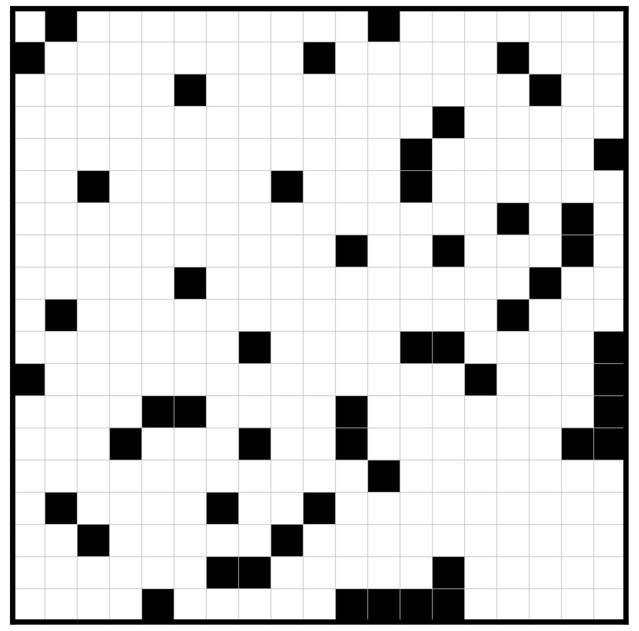
Structure of interest network of **edges** between pairs of **nodes**

<u>Structure of interest</u> network of **edges** between pairs of **nodes**

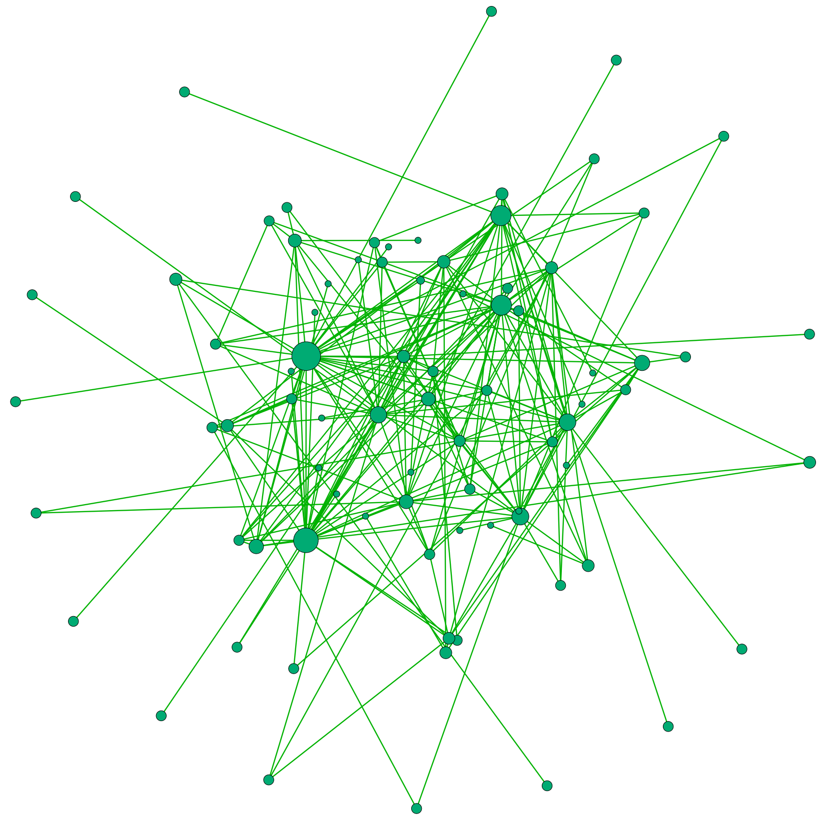


adjacency matrix

Structure of interest network of **edges** between pairs of **nodes**



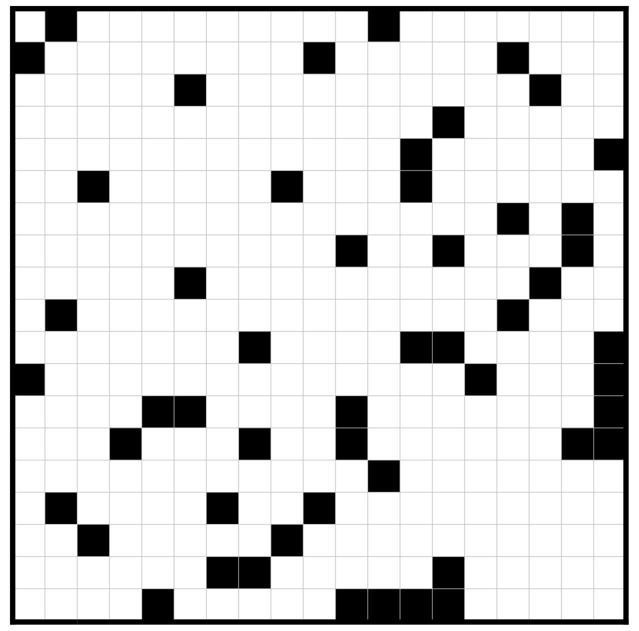
adjacency matrix



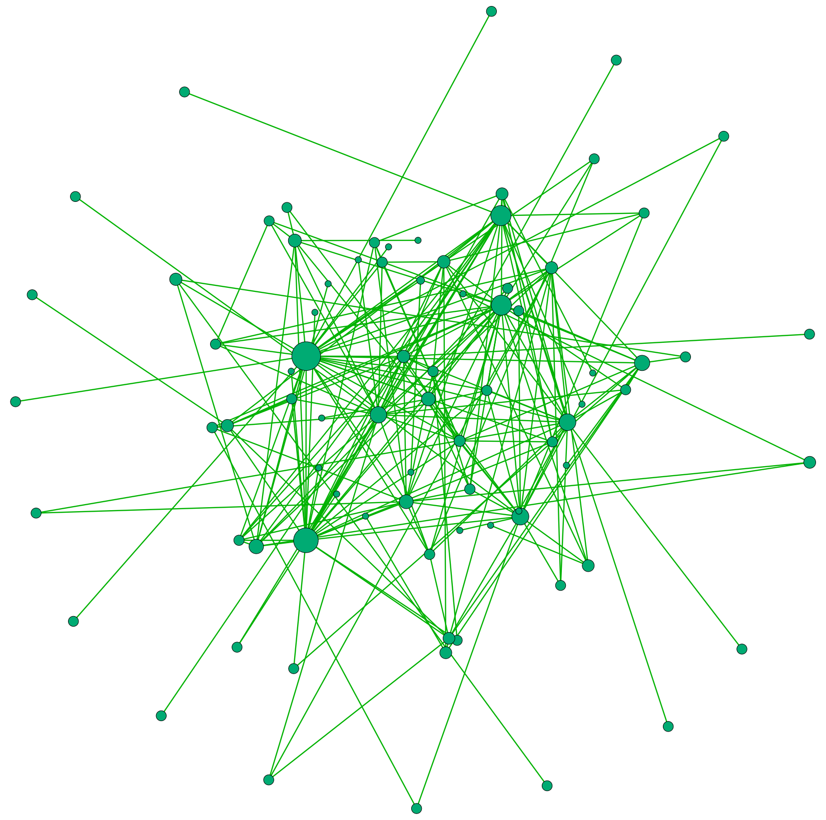
edge-node diagram

Structure of interest

network of **edges** between pairs of **nodes**



adjacency matrix



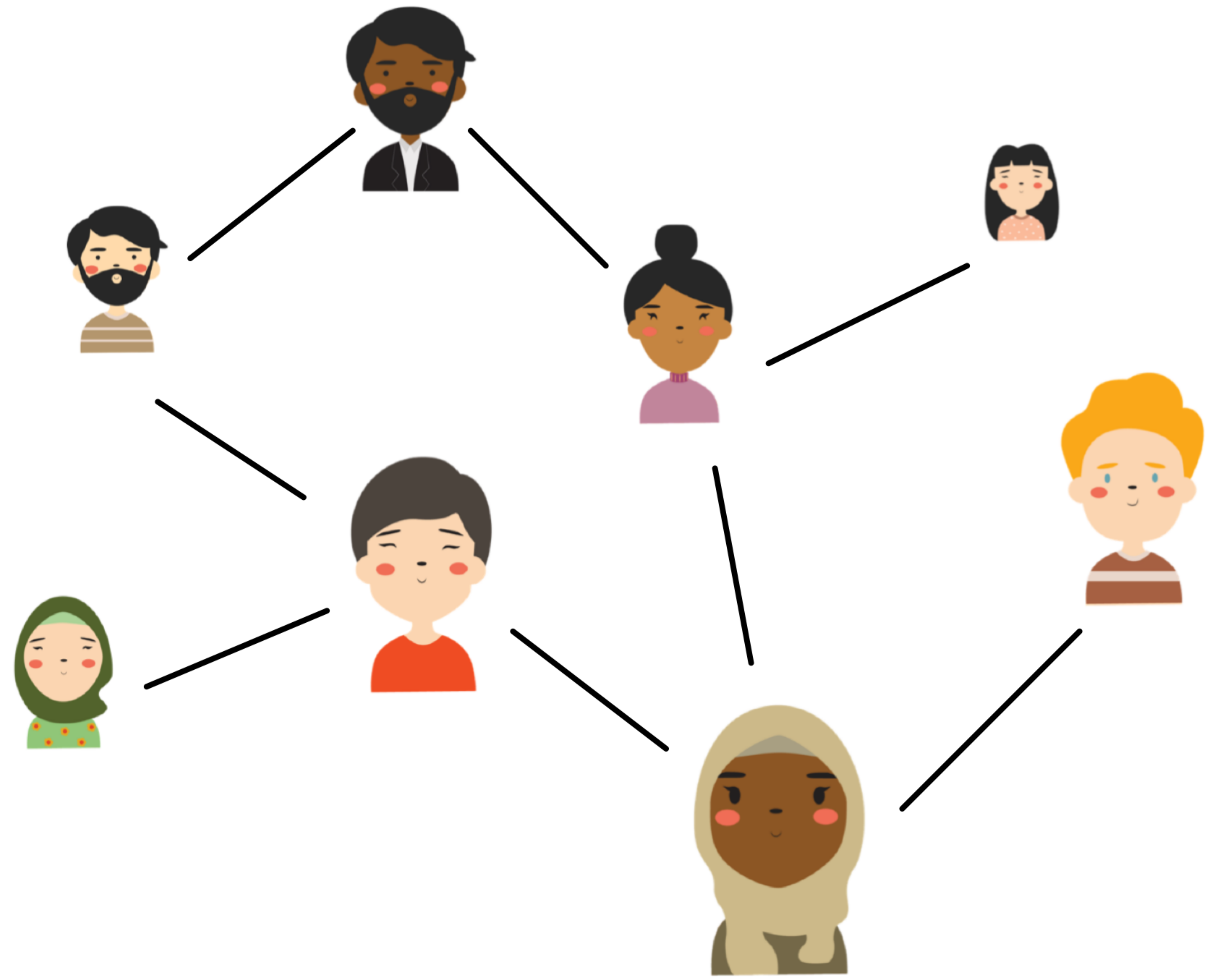
edge-node diagram

nodes	**edges**
people	friendships, emails, collaborations…
neurons	co-activation
proteins	interactions
…	…

examples

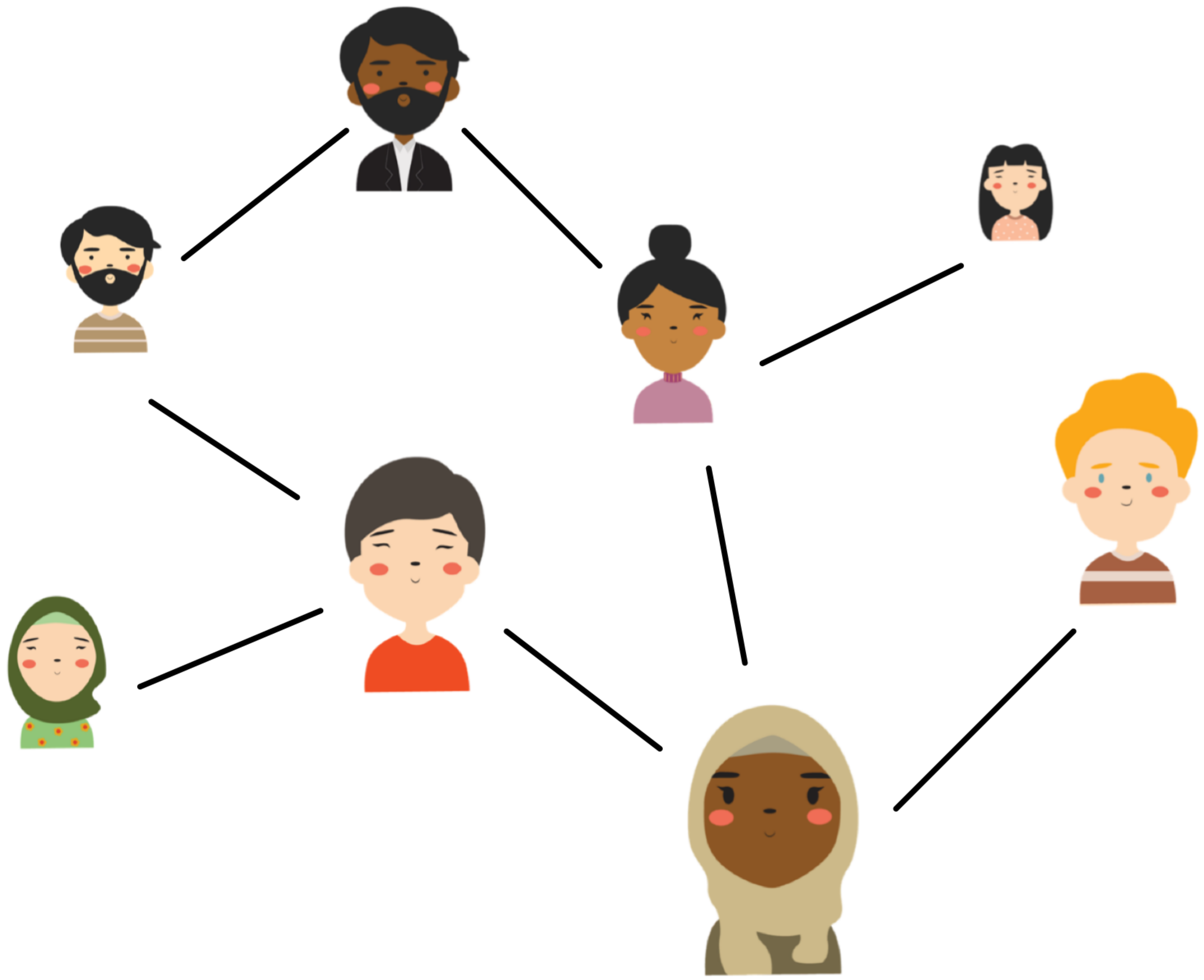
Desired characteristics of our model:

1. sparsity

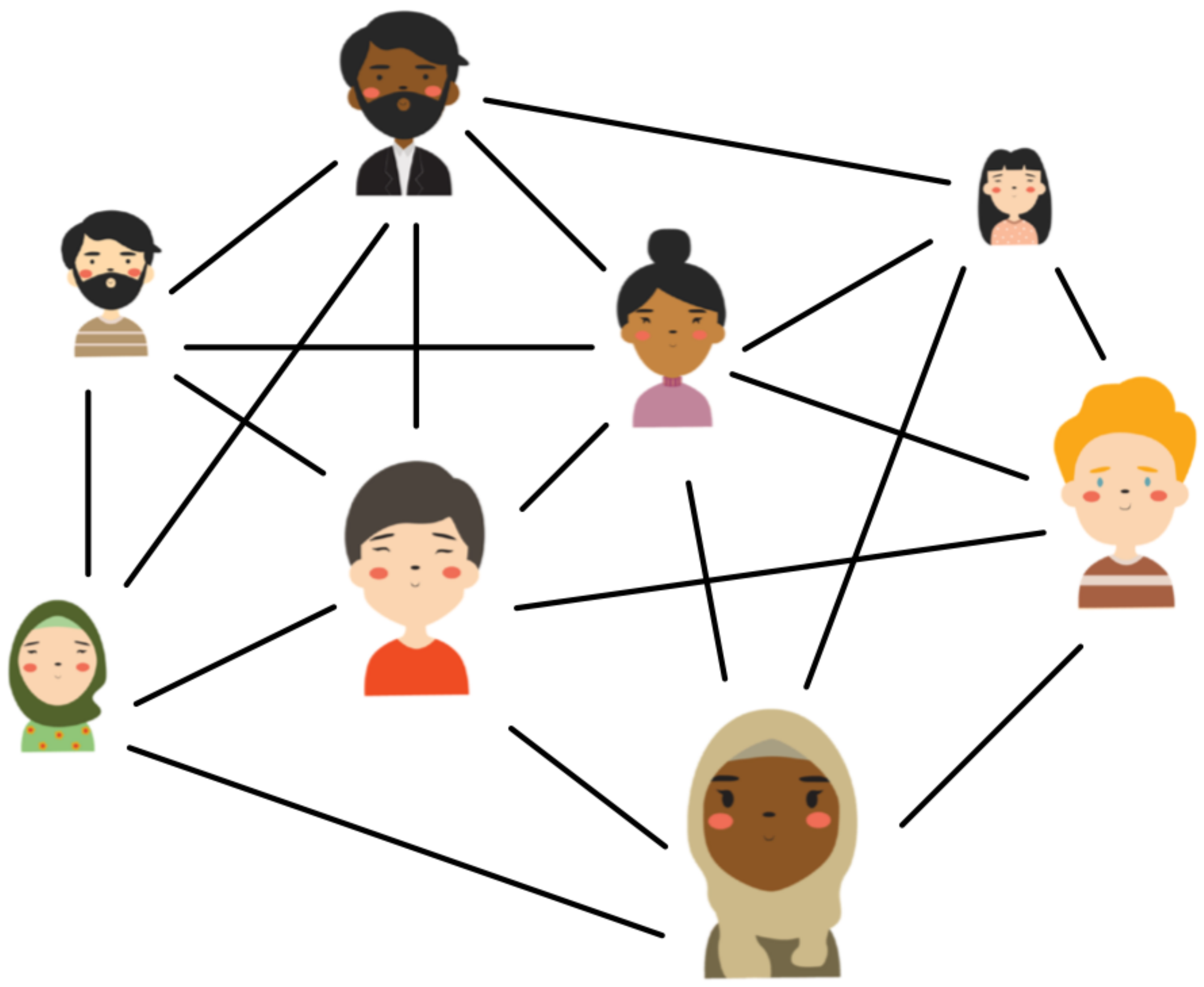


Desired characteristics of our model:

1. sparsity

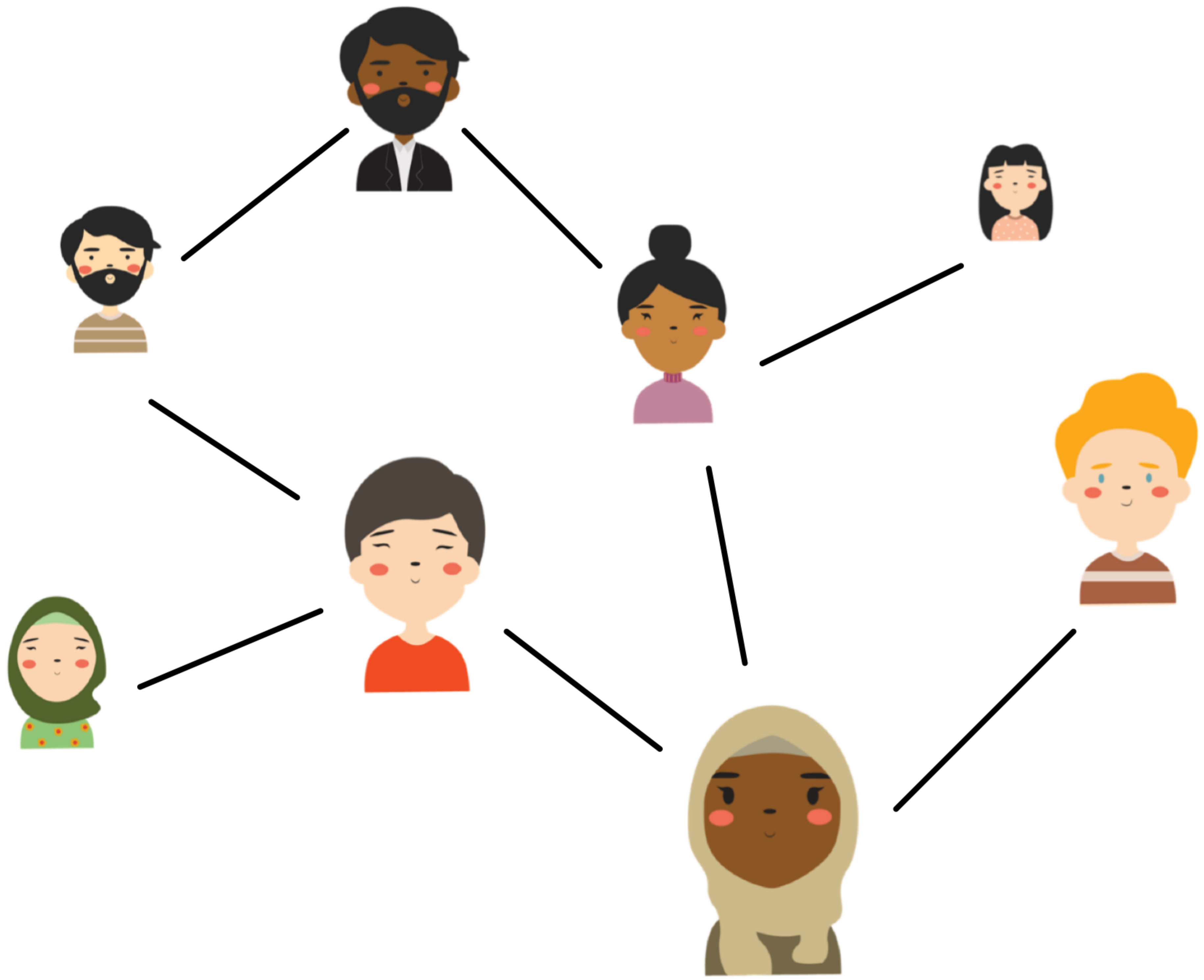


***vs.* density**



Desired characteristics of our model:

1. sparsity



Why is sparsity important?

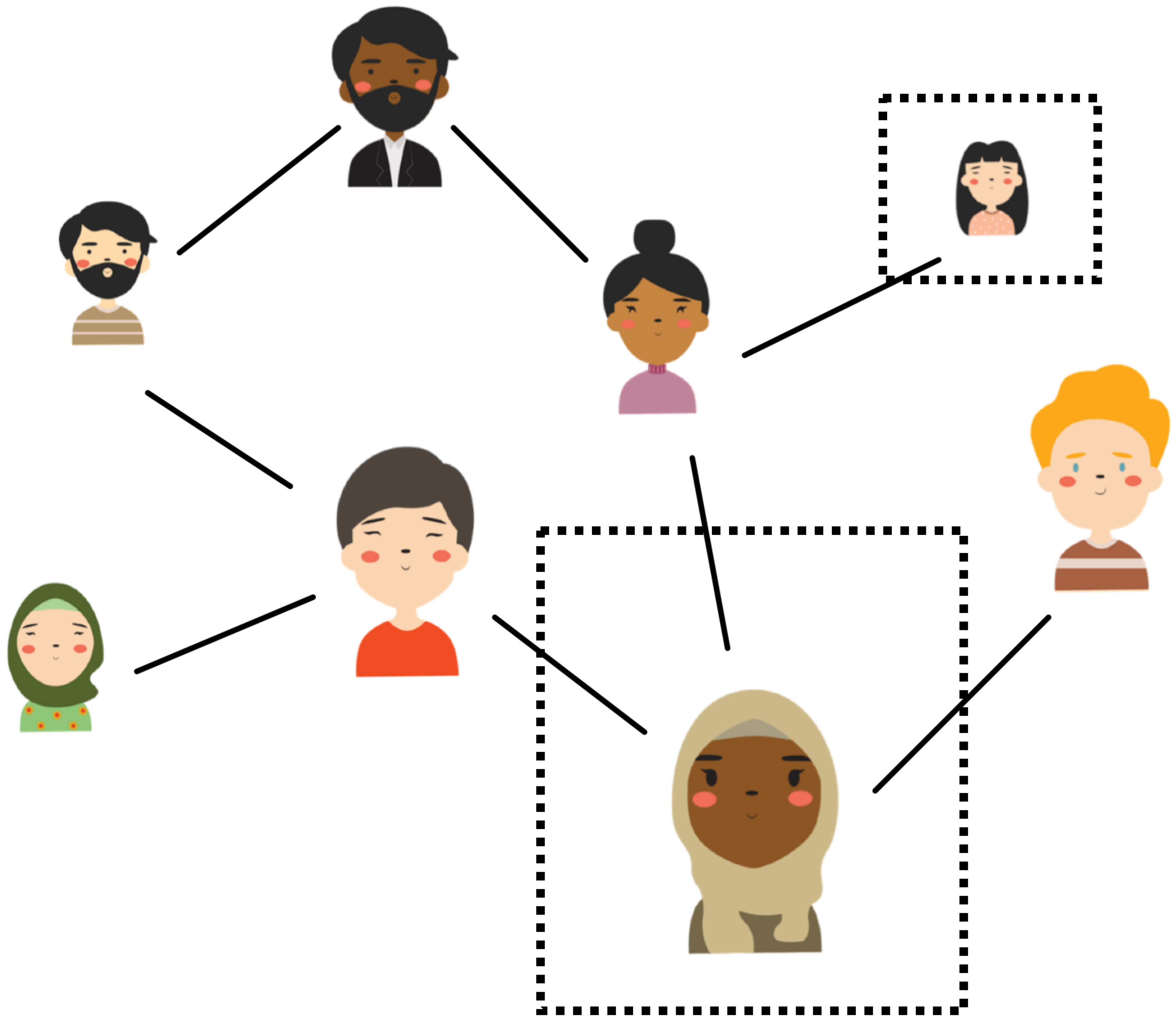
It is a common feature of real world networks

e.g. average number of friends does not grow linearly with population size!

Desired characteristics of our model:

1. sparsity

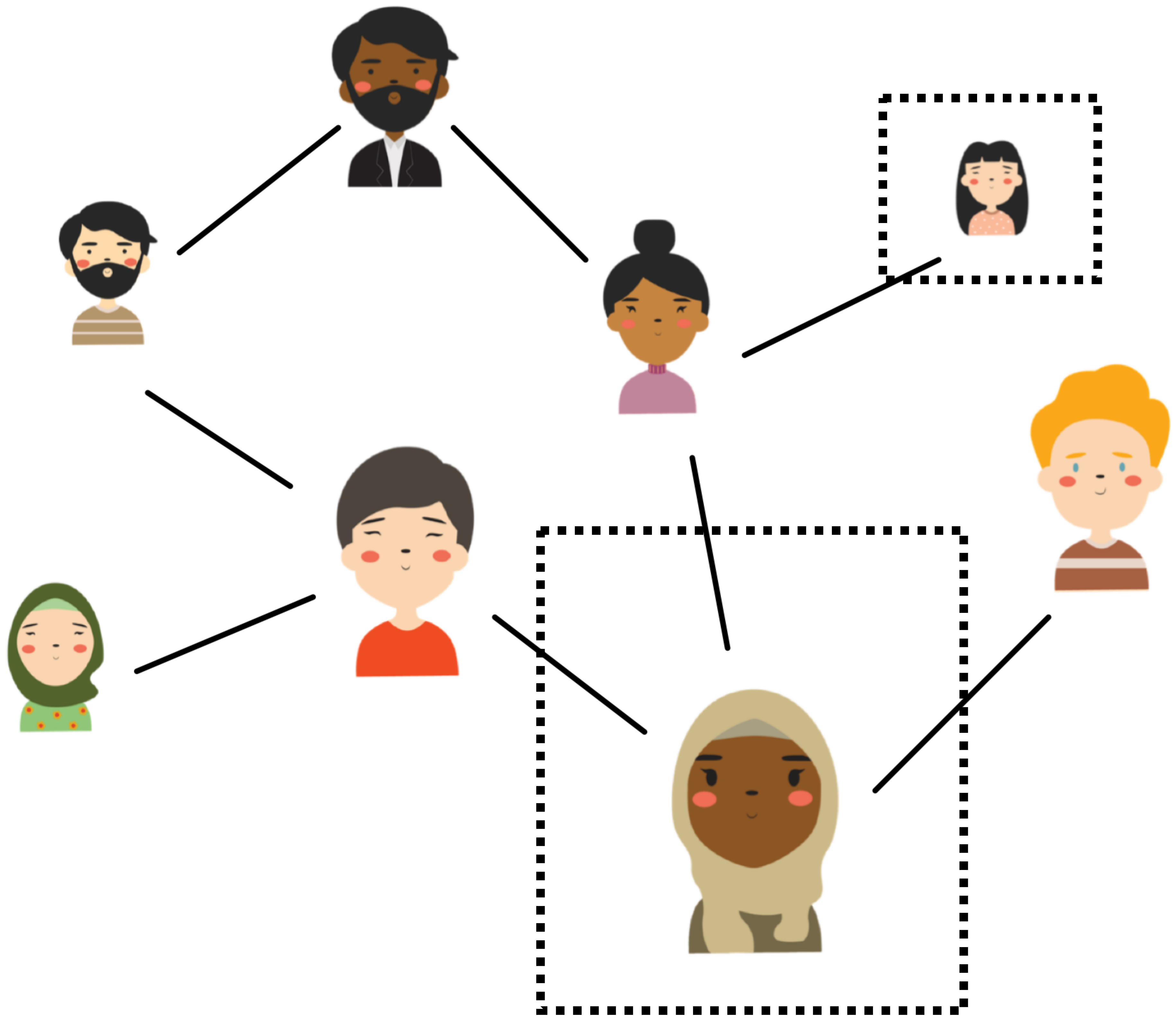
2. degree heterogeneity



Desired characteristics of our model:

1. sparsity

2. degree heterogeneity



Why is degree heterogeneity important?

In real world networks, some nodes have many more edges than others

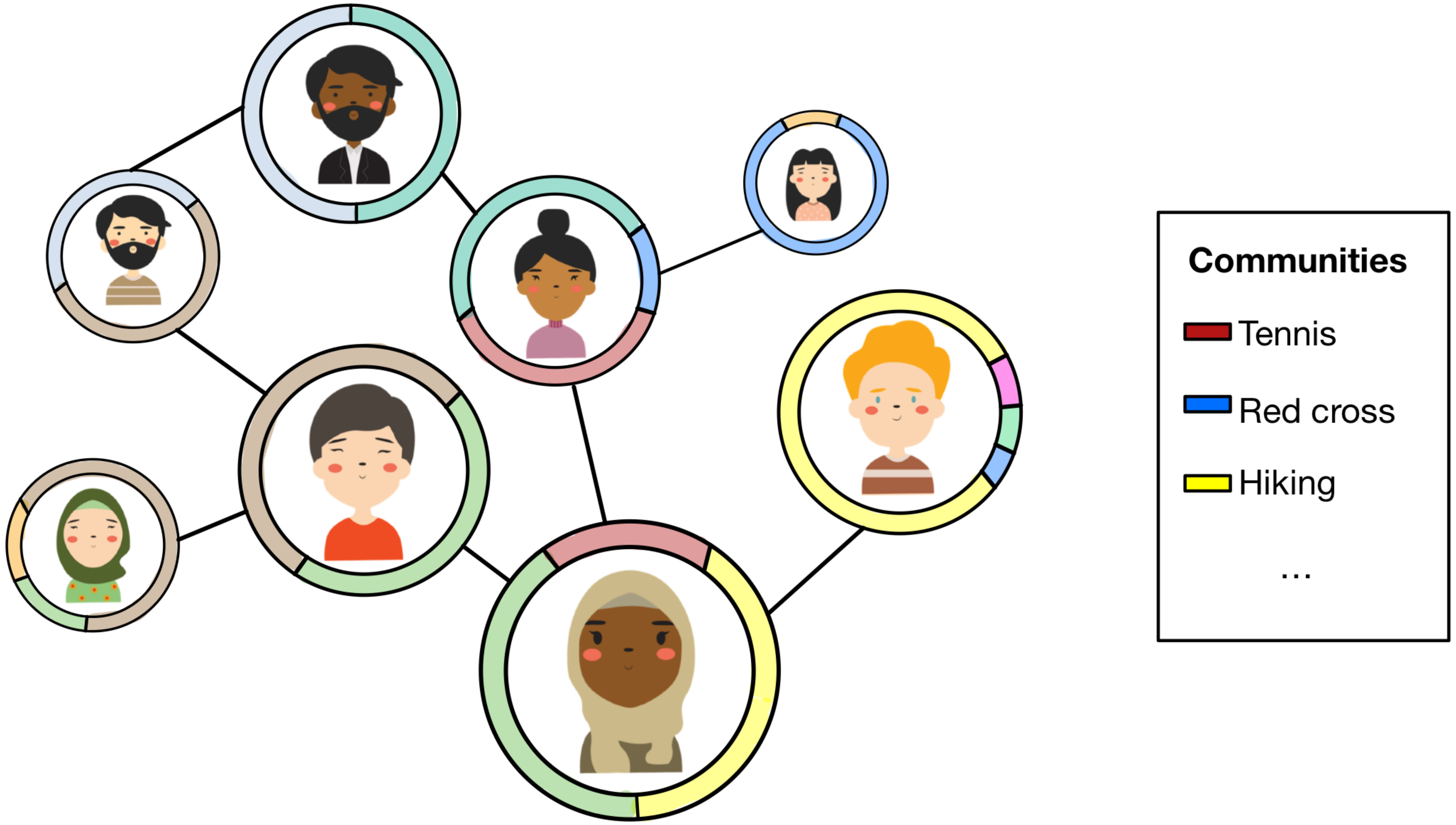
e.g. number of Beyoncé followers on Twitter vs. me

Aim: design probabilistic model featuring:

1. sparsity

2. degree heterogeneity

3. mixed community memberships

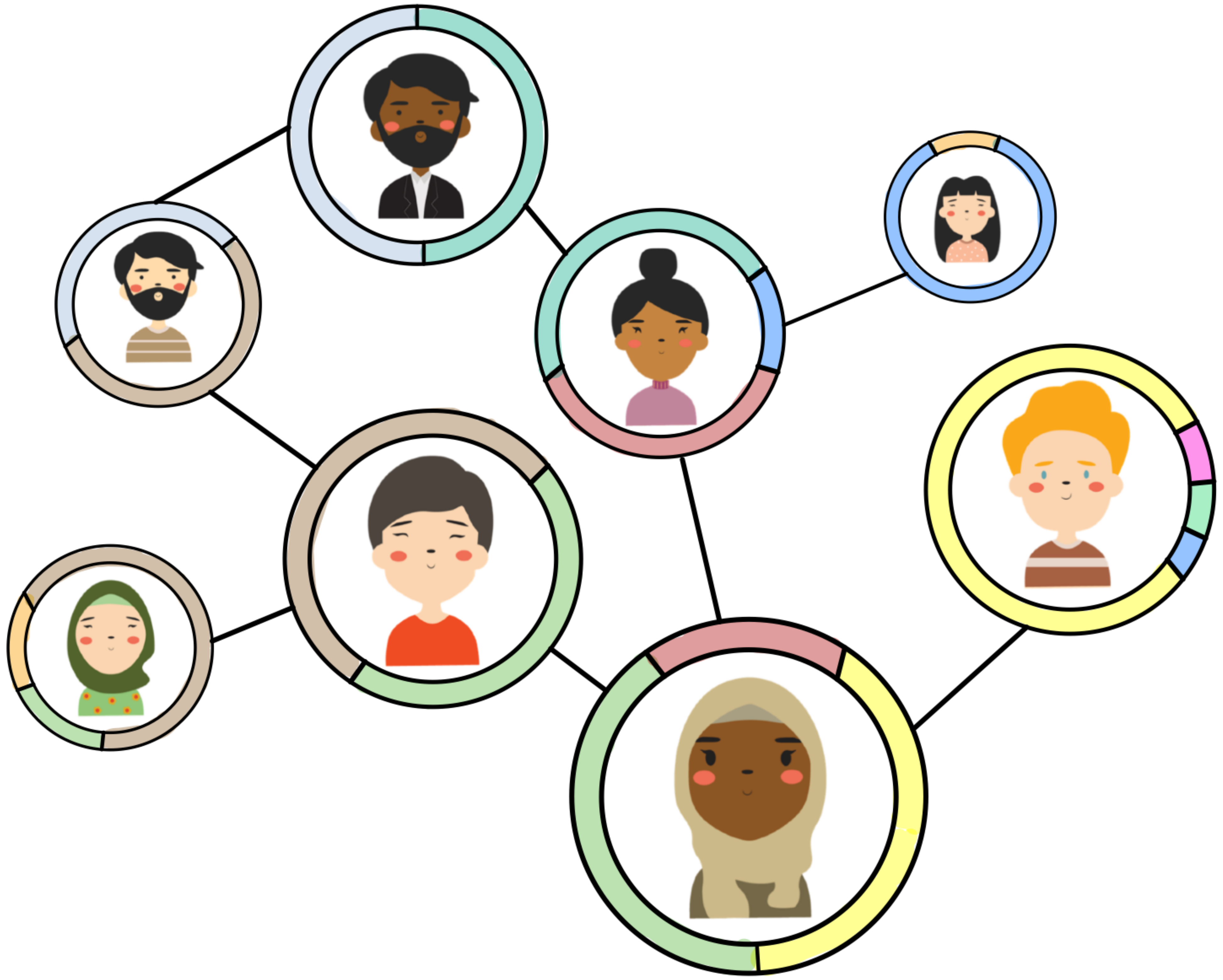


Aim: **design probabilistic model** featuring:

1. sparsity

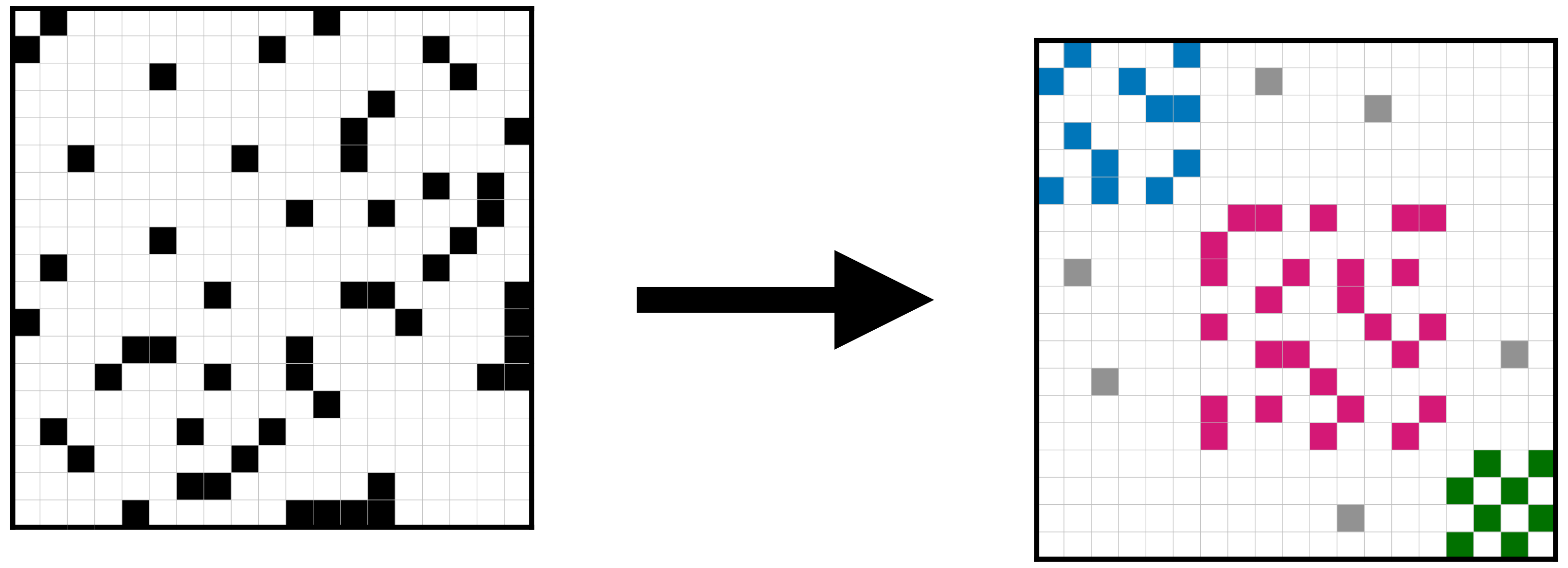
2. degree heterogeneity

3. mixed community memberships



Why are communities important?

To explain and learn edge generating process

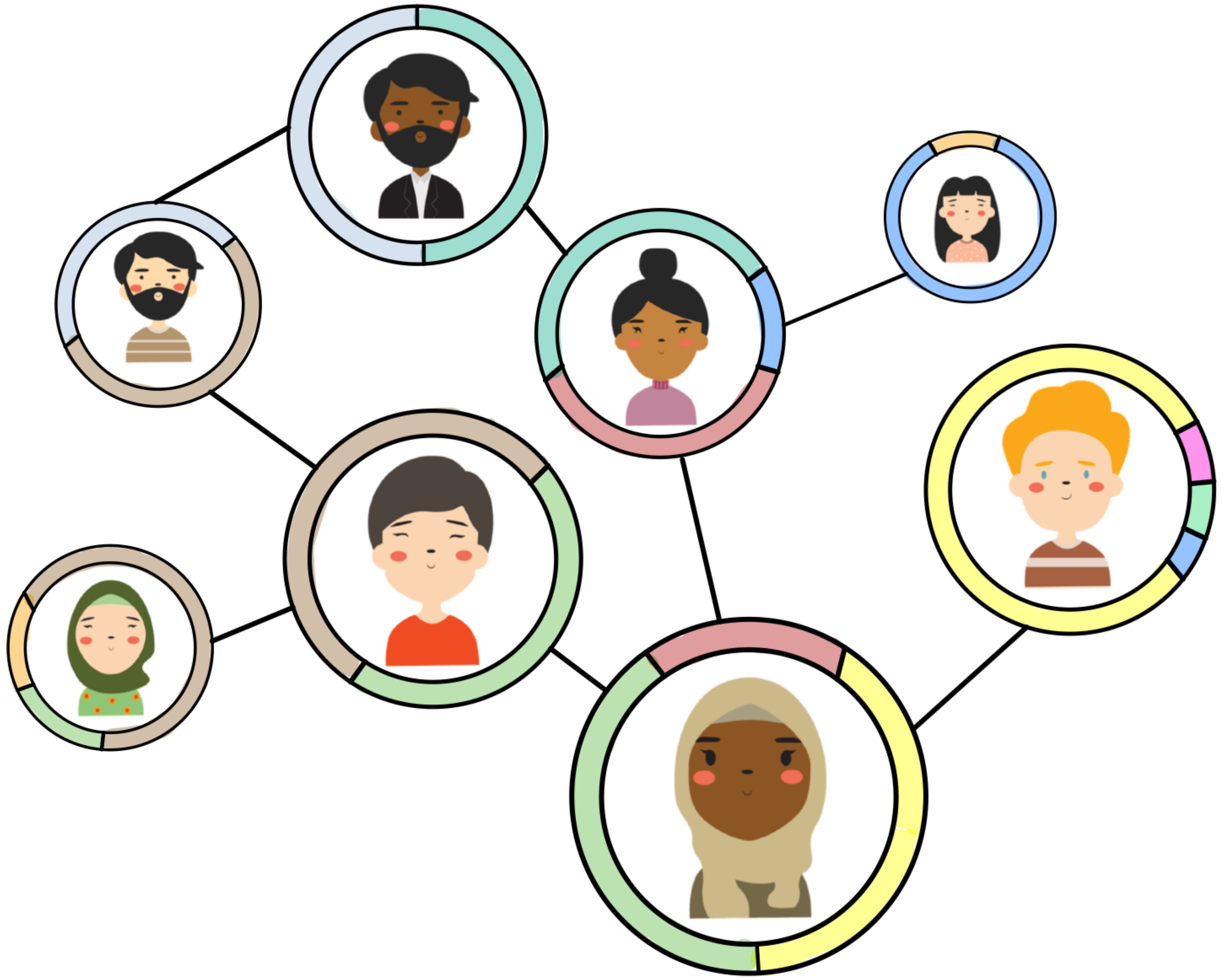


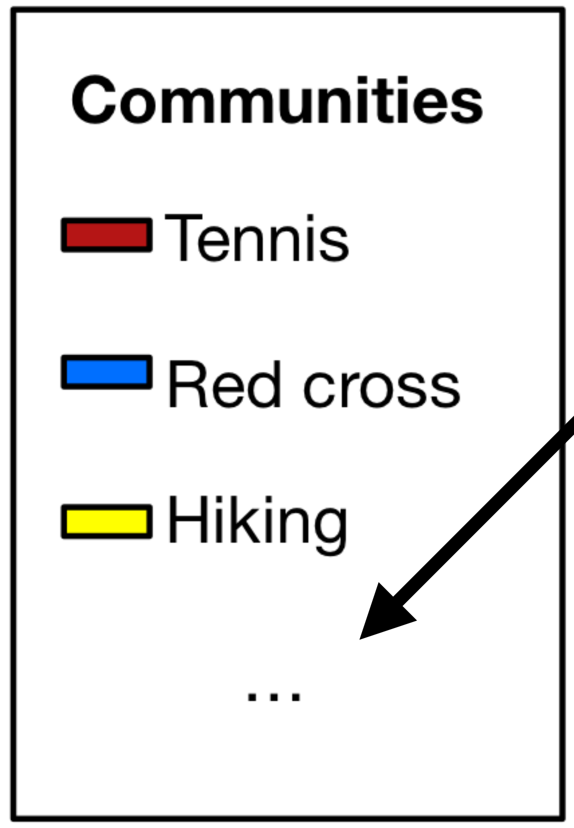
Aim: **design probabilistic model** featuring:

1. sparsity

2. degree heterogeneity

3. mixed community memberships

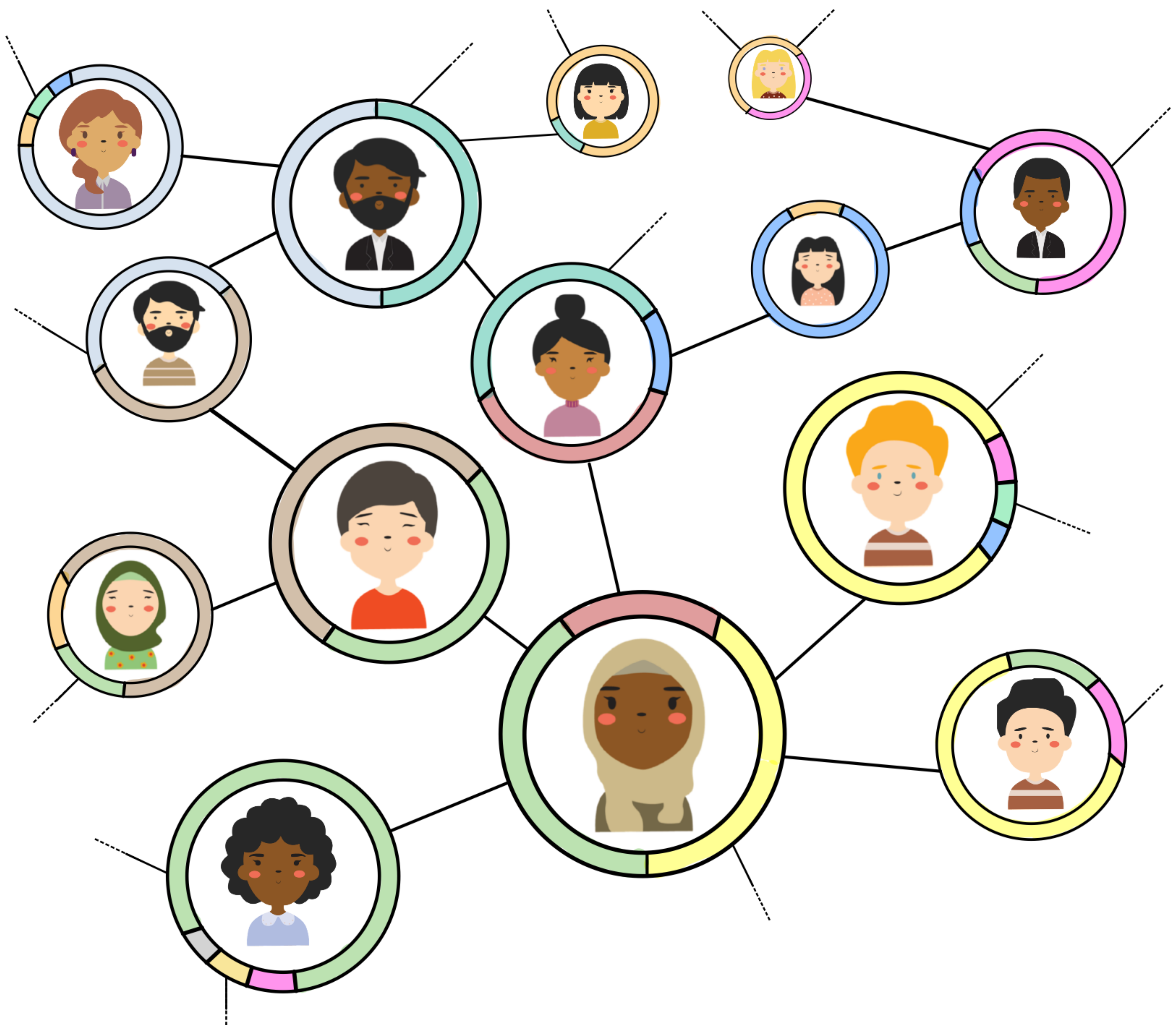




4. learning number of communities

Aim: **design probabilistic model** featuring:

1. **sparsity**
2. **degree heterogeneity**
3. **mixed community memberships**



4. **learning number of communities**
5. **can scale to large networks**

Overview

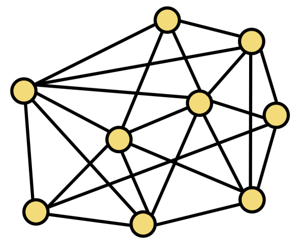
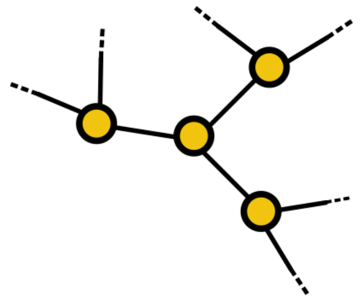
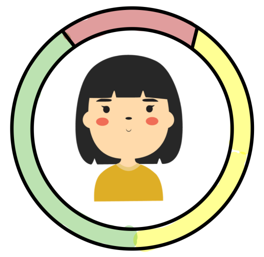
Motivation

Model

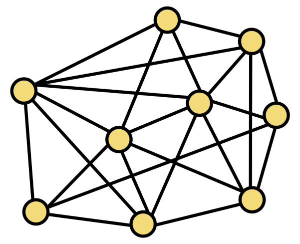
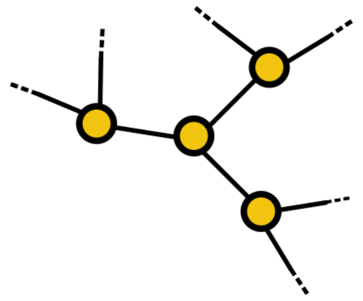
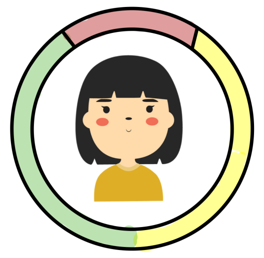
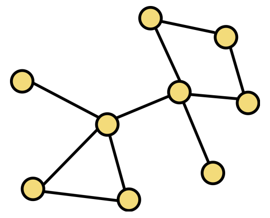
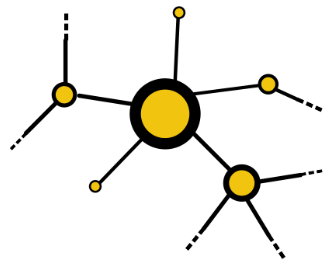
BACKGROUND

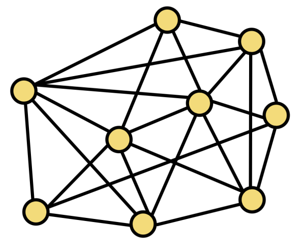
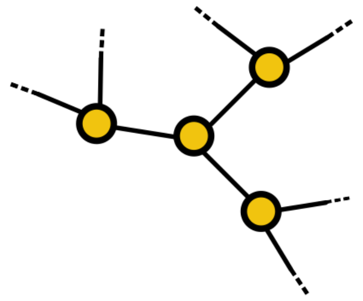
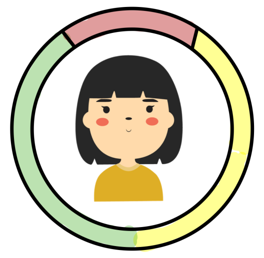
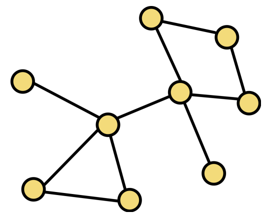
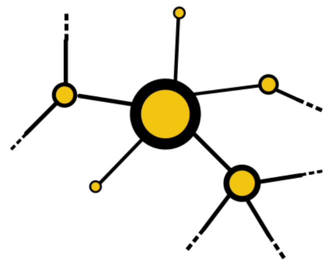
Mixed membership block models
(Airoldi et al. 2008)

Models based on completely random measures
(Caron and Fox 2017)

BACKGROUND	**Mixed membership block models (Airoldi et al. 2008)**	Dense 	No degree heterogeneity 	Mixed community membership 
	Models based on completely random measures (Caron and Fox 2017)			

BACKGROUND

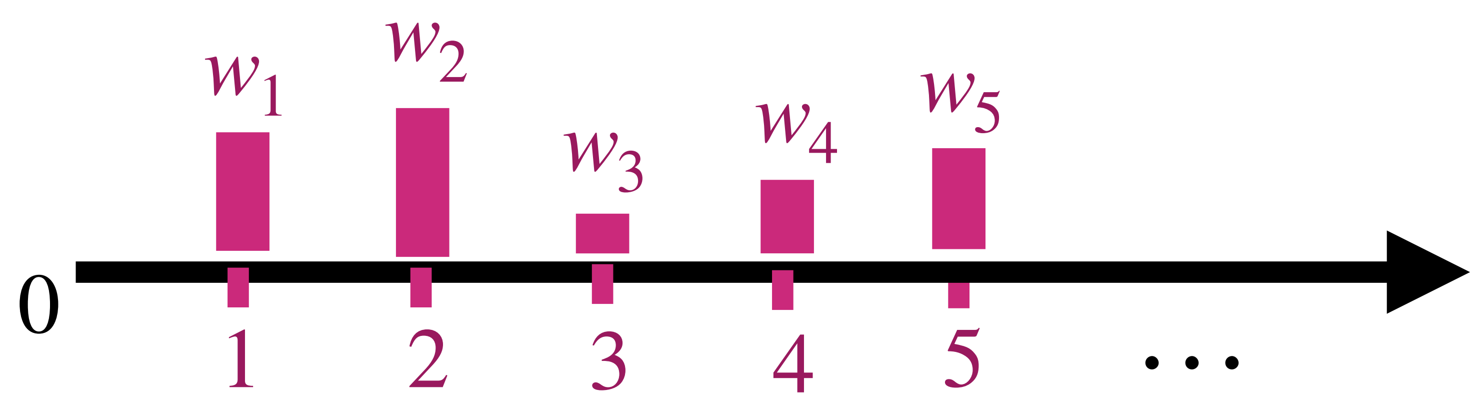
Mixed membership block models (Airoldi et al. 2008)	Dense 	No degree heterogeneity 	Mixed community membership 
Models based on completely random measures (Caron and Fox 2017)	Sparse 	Degree heterogeneity 	No community membership 

BACKGROUND			
Mixed membership block models (Airoldi et al. 2008)	Dense 	No degree heterogeneity 	**Mixed community membership** 
Models based on completely random measures (Caron and Fox 2017)	**Sparse** 	**Degree heterogeneity** 	No community membership 

BACKGROUND	**Mixed membership block models (Airoldi et al. 2008)**	Dense	No degree heterogeneity	**Mixed community membership**
	Models based on completely random measures (Caron and Fox 2017)	**Sparse**	**Degree heterogeneity**	No community membership
PROPOSED MODEL	**Models based on thinned completely random measures (proposed models)**	**Sparse**	**Degree heterogeneity**	**Mixed community membership**

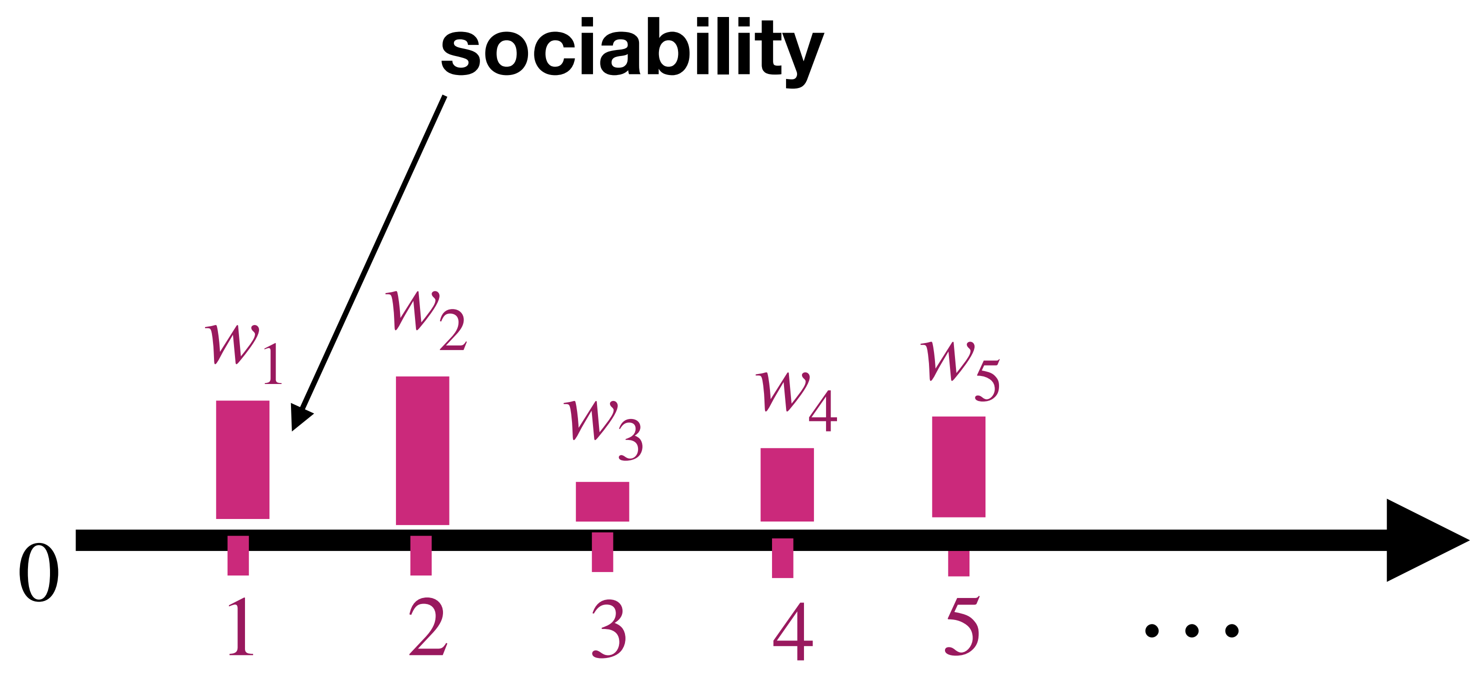
Proposed model

1. Draw set of ***potential nodes and edges*** with the Generalized Gamma Process (as Caron and Fox (2017))



Proposed model

1. Draw set of ***potential nodes and edges*** with the Generalized Gamma Process (as Caron and Fox (2017))

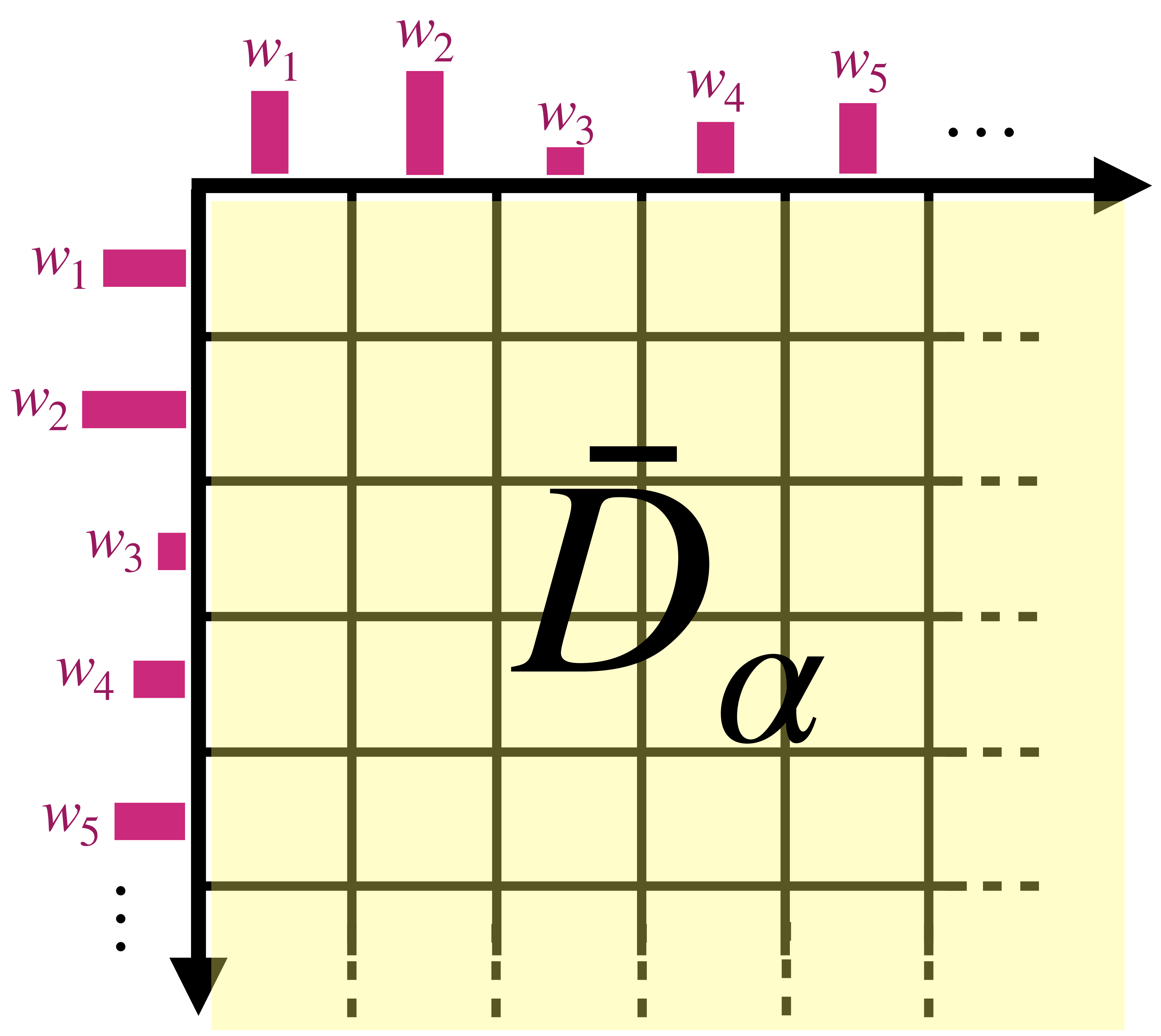


Proposed model

1. Draw set of ***potential nodes and edges*** with the Generalized Gamma Process (as Caron and Fox (2017))

1.1 Draw total number $\bar{D}_\alpha$ of (directed) edges:

$$\bar{D}_\alpha \sim \text{Poisson}\left(\bar{W}_\alpha^2\right) \qquad \bar{W}_\alpha = \sum_i w_i$$



Proposed model

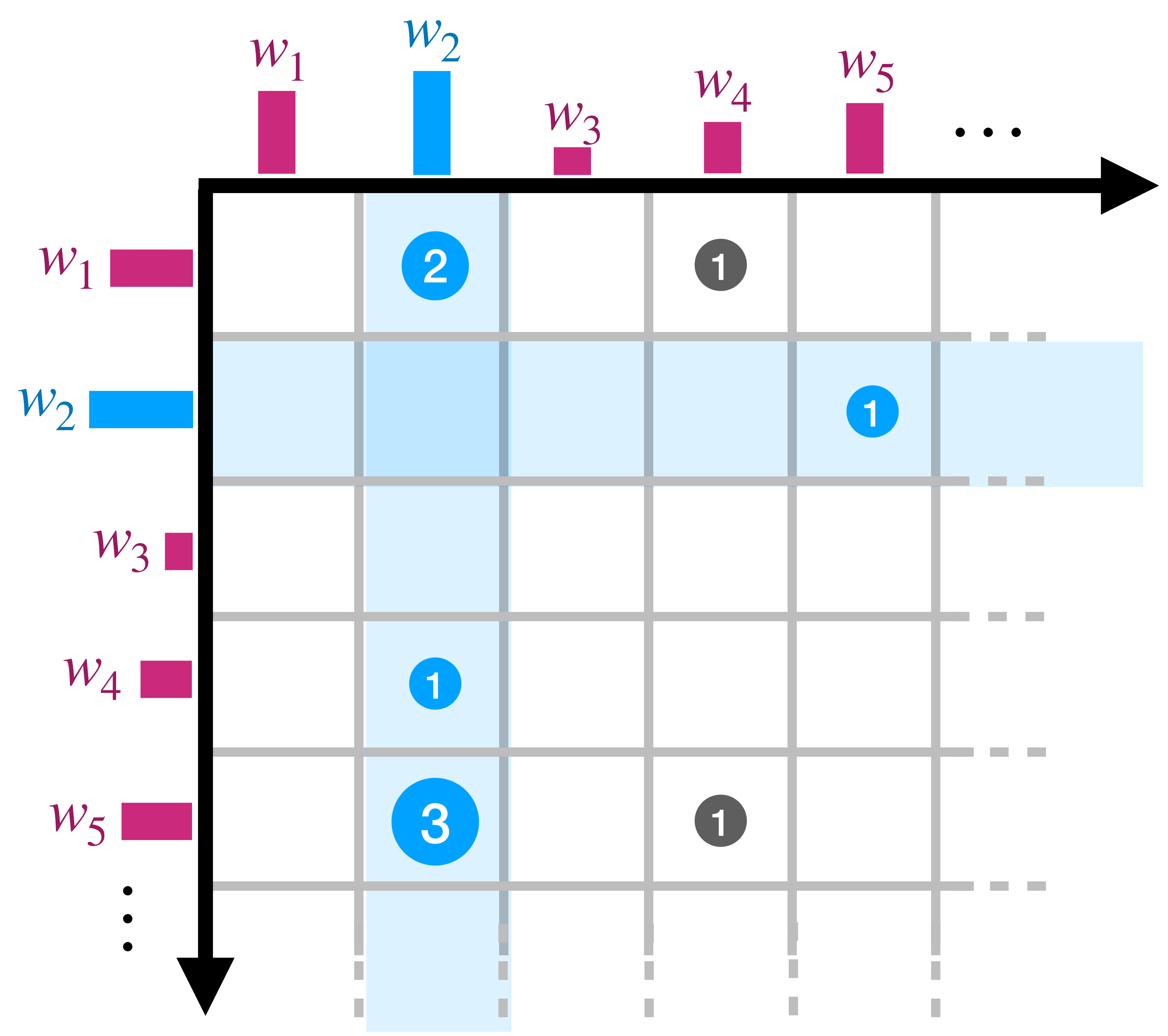
1. Draw set of ***potential nodes and edges*** with the Generalized Gamma Process (as Caron and Fox (2017))

1.1 Draw total number $\bar{D}_\alpha$ of (directed) edges:

$$\bar{D}_\alpha \sim \text{Poisson}\left(\bar{W}_\alpha^2\right) \qquad \bar{W}_\alpha = \sum_i w_i$$

1.2 Assign each edge to a node pair based on sociabilities:

$$P(x_{e1} = i) = \frac{w_i}{\bar{W}_\alpha}, \qquad P(x_{e2} = j) = \frac{w_j}{\bar{W}_\alpha}$$



high sociability
—>
high degree

Proposed model

2. Assign nodes to (possibly multiple) communities:

Proposed model

2. Assign nodes to (possibly multiple) communities:

example with $K = 4$ communities

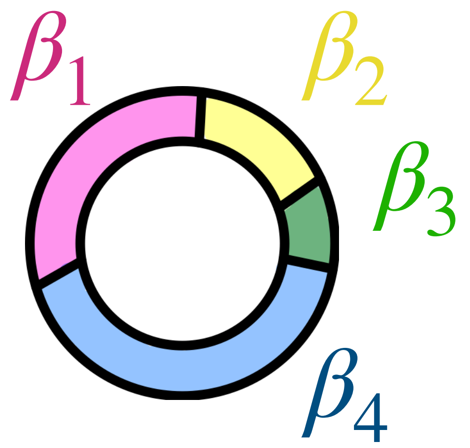
Proposed model

2. Assign nodes to (possibly multiple) communities:

example with $K = 4$ communities

2.1 Draw *global* frequency of each of K communities (as in Mixed Membership Stochastic Blockmodels):

$$(\beta_1, \ldots, \beta_K) \sim \text{Dirichlet}\left(\frac{\gamma}{K}, \ldots, \frac{\gamma}{K}\right)$$



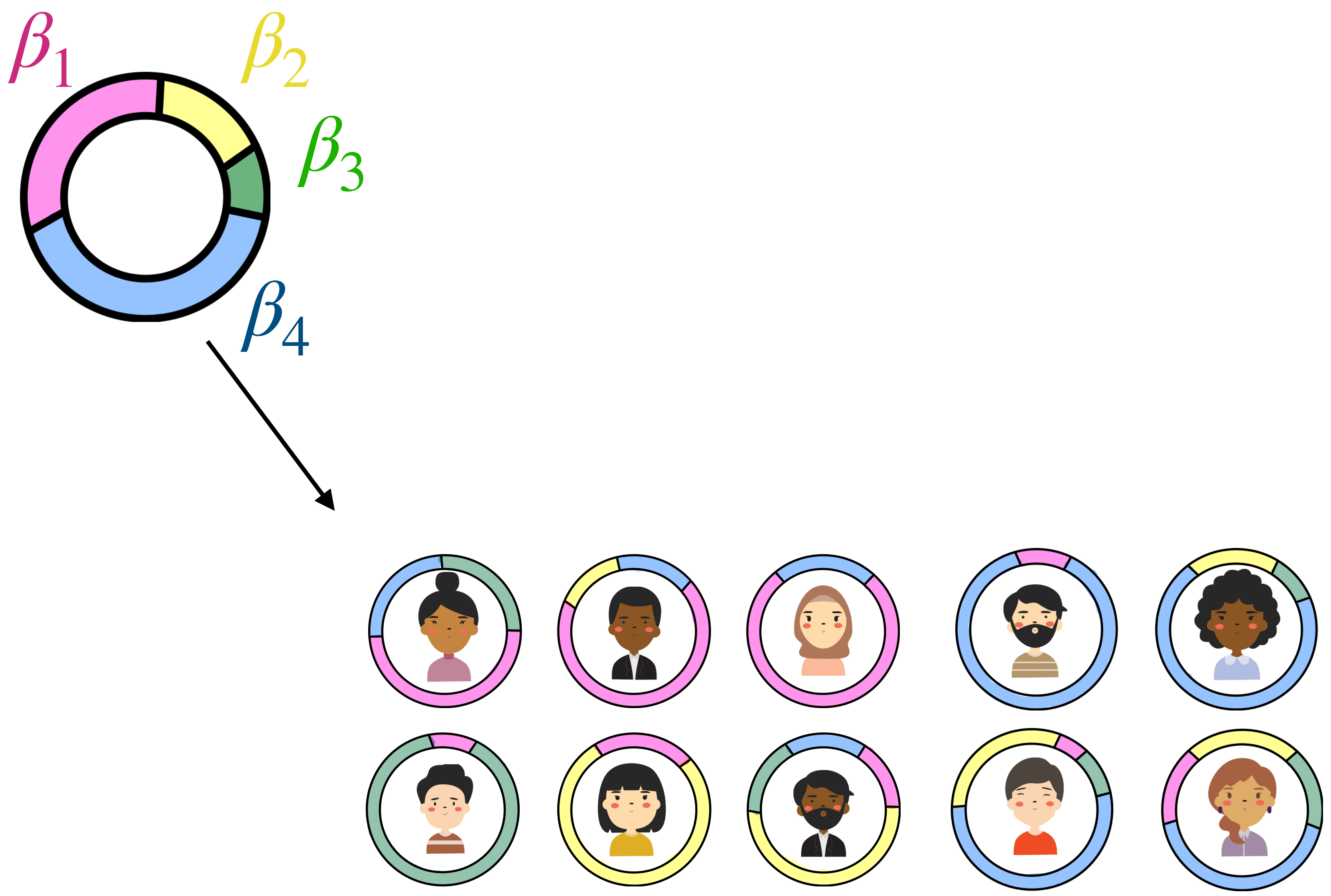
Proposed model

2. Assign nodes to (possibly multiple) communities:

example with $K = 4$ communities

2.1 Draw *global* frequency of each of K communities (as in Mixed Membership Stochastic Blockmodels):

$$(\beta_1, \ldots, \beta_K) \sim \text{Dirichlet}\left(\frac{\gamma}{K}, \ldots, \frac{\gamma}{K}\right)$$



2.2 Assign each node i to a distribution over communities π_i :

$$\pi_i = (\pi_{i1}, \ldots, \pi_{iK}) \mid \beta \overset{\text{ind}}{\sim} \text{Dirichlet}\left(\zeta\beta_1, \ldots, \zeta\beta_K\right)$$

Proposed model

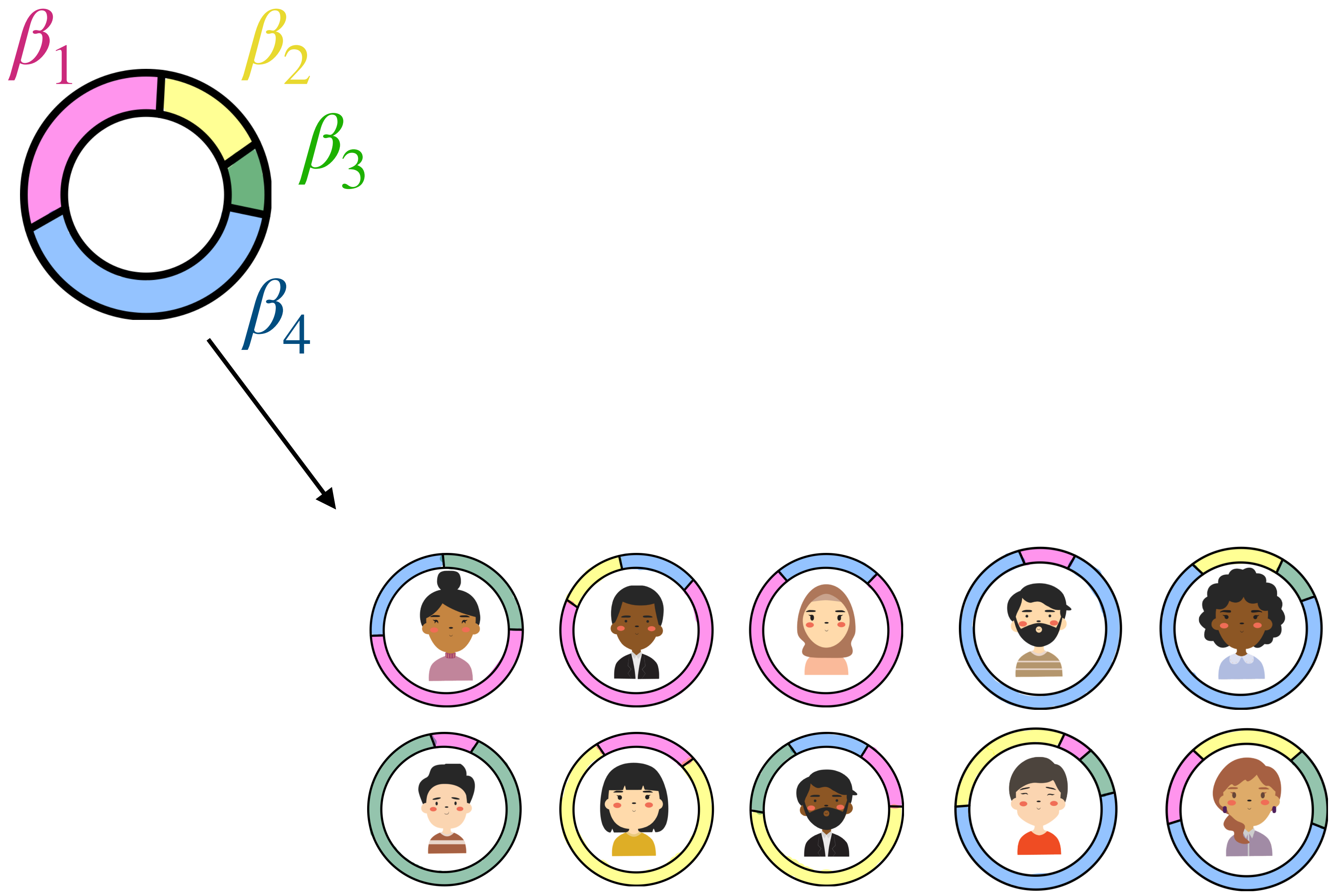
Large K and $\gamma < K$
$\rightarrow$ can learn $K_{\textbf{true}} < K$

2. Assign nodes to (possibly multiple) communities:

Approximates Dirichlet Process as $K \rightarrow \infty$

2.1 Draw *global* frequency of each of K communities (as in Mixed Membership Stochastic Blockmodels):

$$(\beta_1, \ldots, \beta_K) \sim \text{Dirichlet}\left(\frac{\gamma}{K}, \ldots, \frac{\gamma}{K}\right)$$



2.2 Assign each node i to a distribution over communities π_i :

$$\pi_i = (\pi_{i1}, \ldots, \pi_{iK}) \mid \beta \overset{\text{ind}}{\sim} \text{Dirichlet}\left(\zeta\beta_1, \ldots, \zeta\beta_K\right)$$

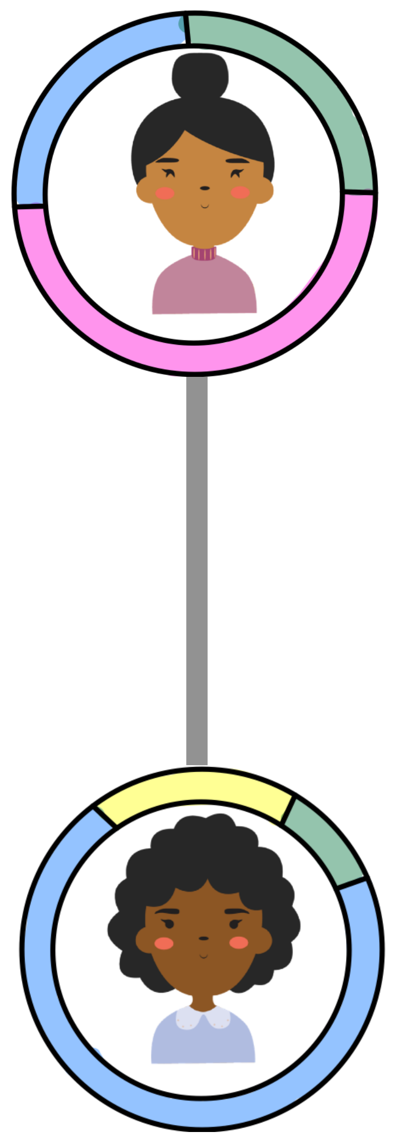
Proposed model

3. For each edge, we assign nodes to communities:

Proposed model

3. For each edge, we assign nodes to communities:

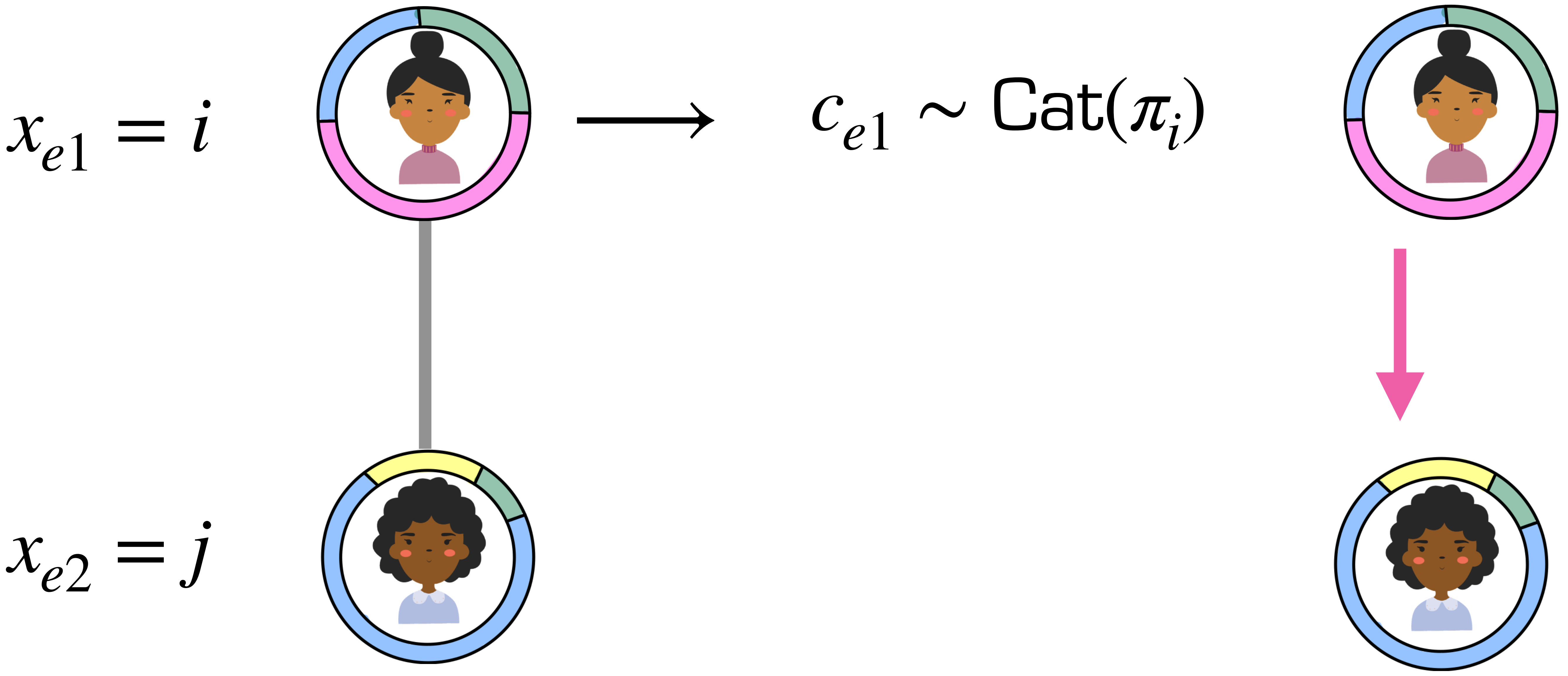
$x_{e1} = i$



$x_{e2} = j$

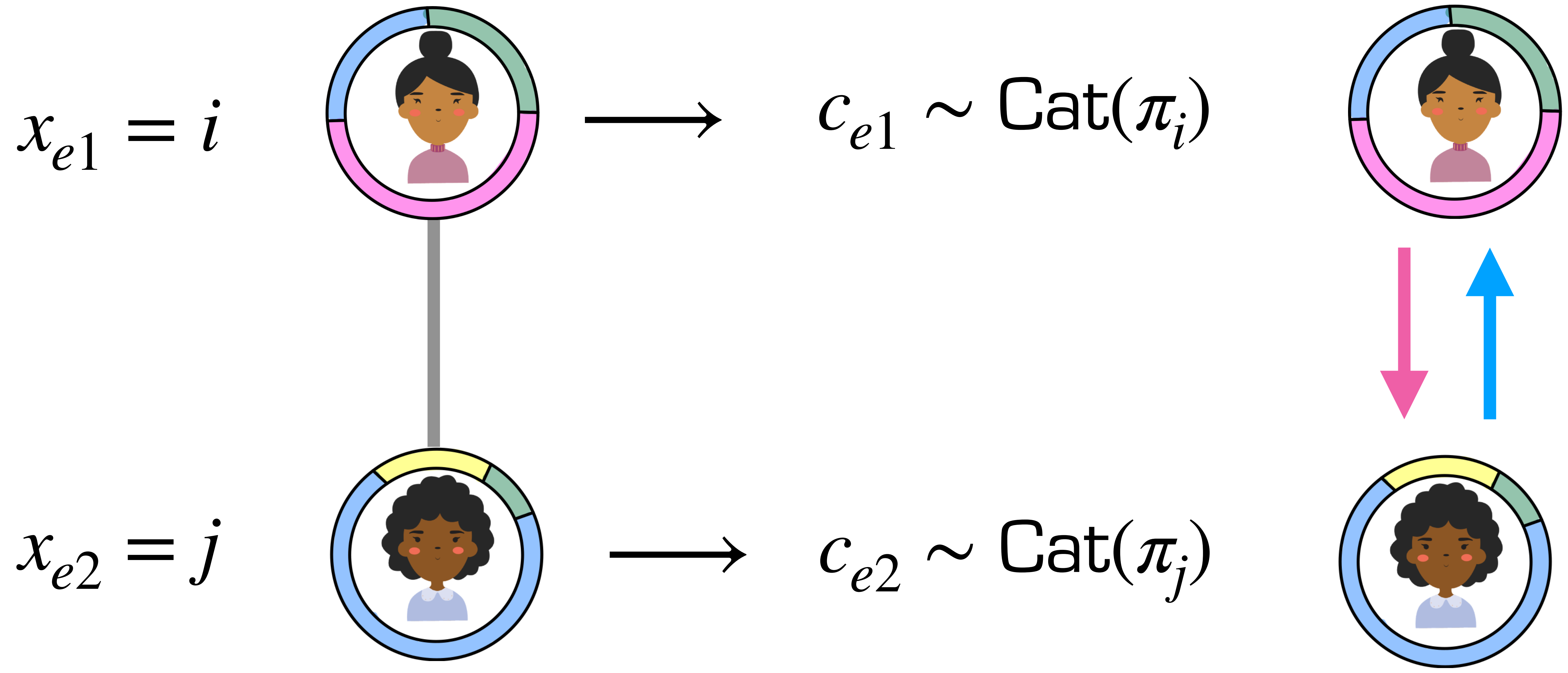
Proposed model

3. For each edge, we assign nodes to communities:



Proposed model

3. For each edge, we assign nodes to communities:



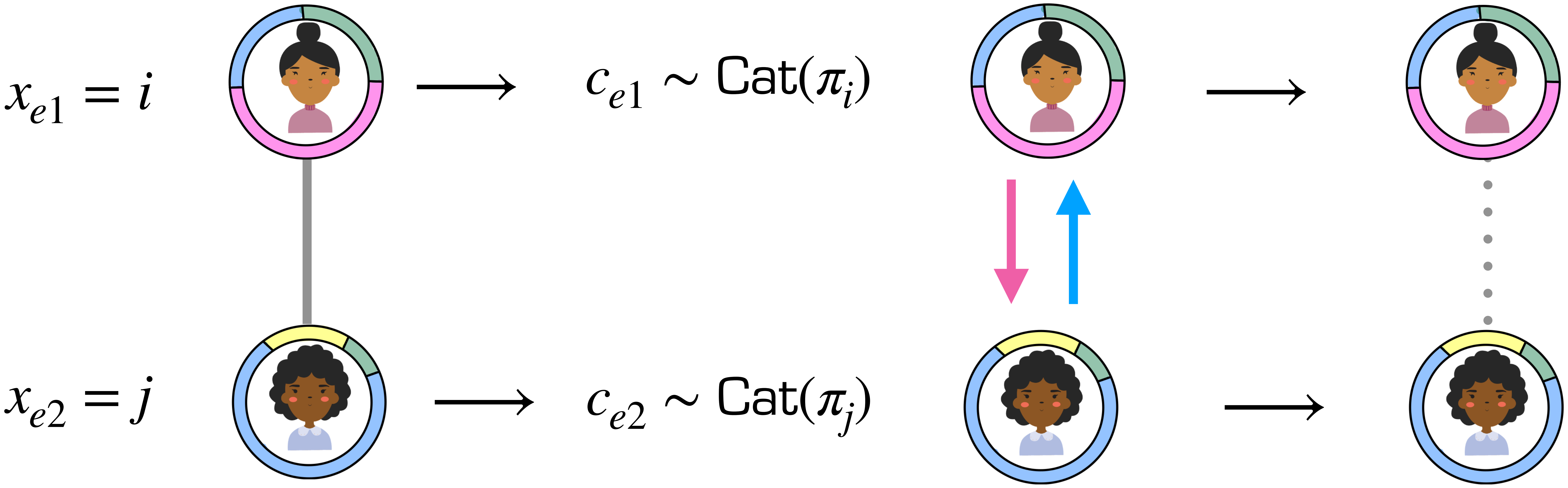
$$x_{e1} = i \longrightarrow c_{e1} \sim \text{Cat}(\pi_i)$$

$$x_{e2} = j \longrightarrow c_{e2} \sim \text{Cat}(\pi_j)$$

Proposed model

3. For each edge, we assign nodes to communities:

3.1 **Thin (remove) edges** between nodes assigned to **different communities:**

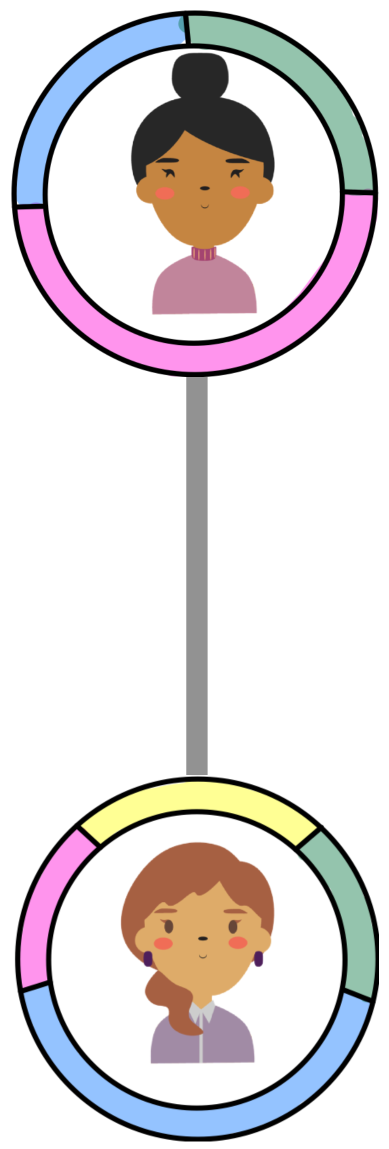


Proposed model

3. For each edge, we assign nodes to communities:

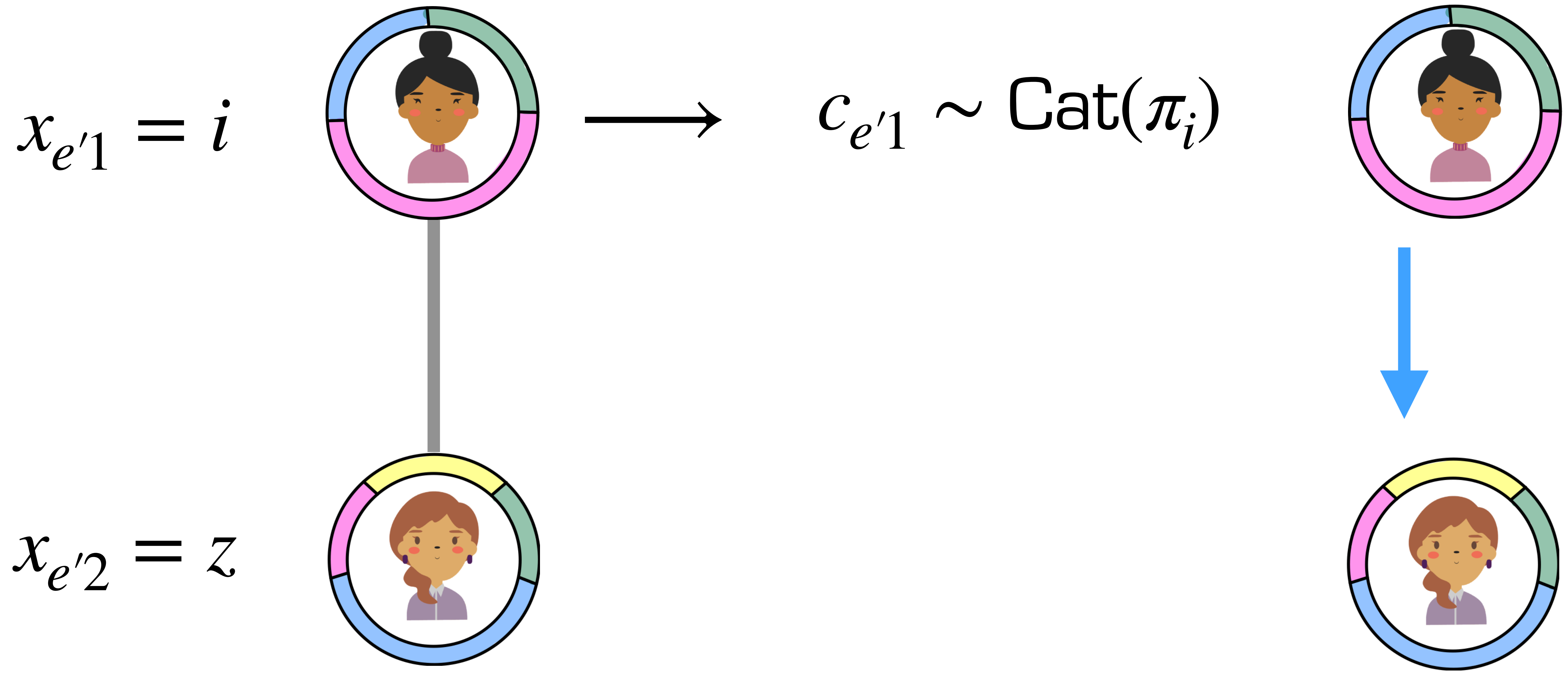
$$x_{e'1} = i$$

$$x_{e'2} = z$$



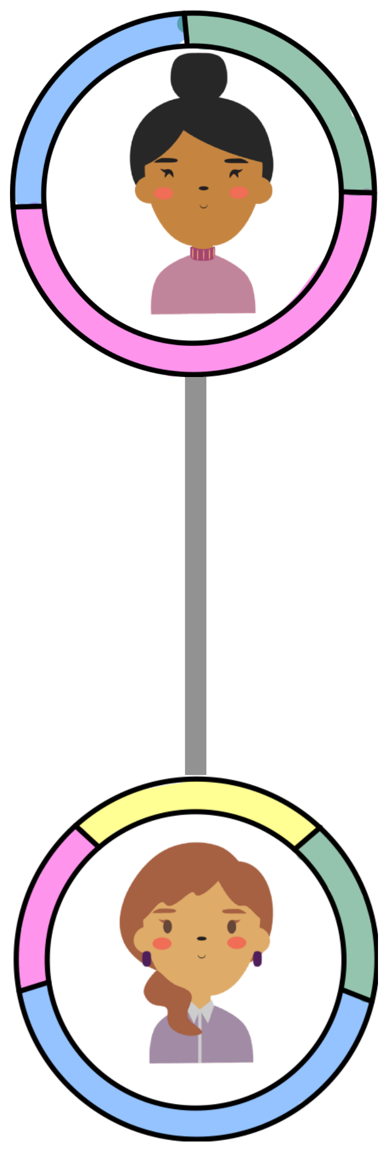
Proposed model

3. For each edge, we assign nodes to communities:



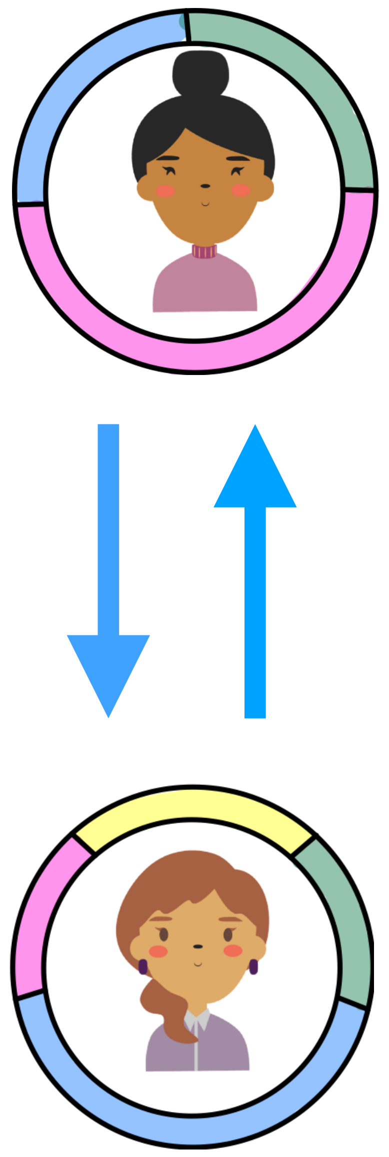
Proposed model

3. For each edge, we assign nodes to communities:



$$x_{e'1} = i \longrightarrow c_{e'1} \sim \text{Cat}(\pi_i)$$

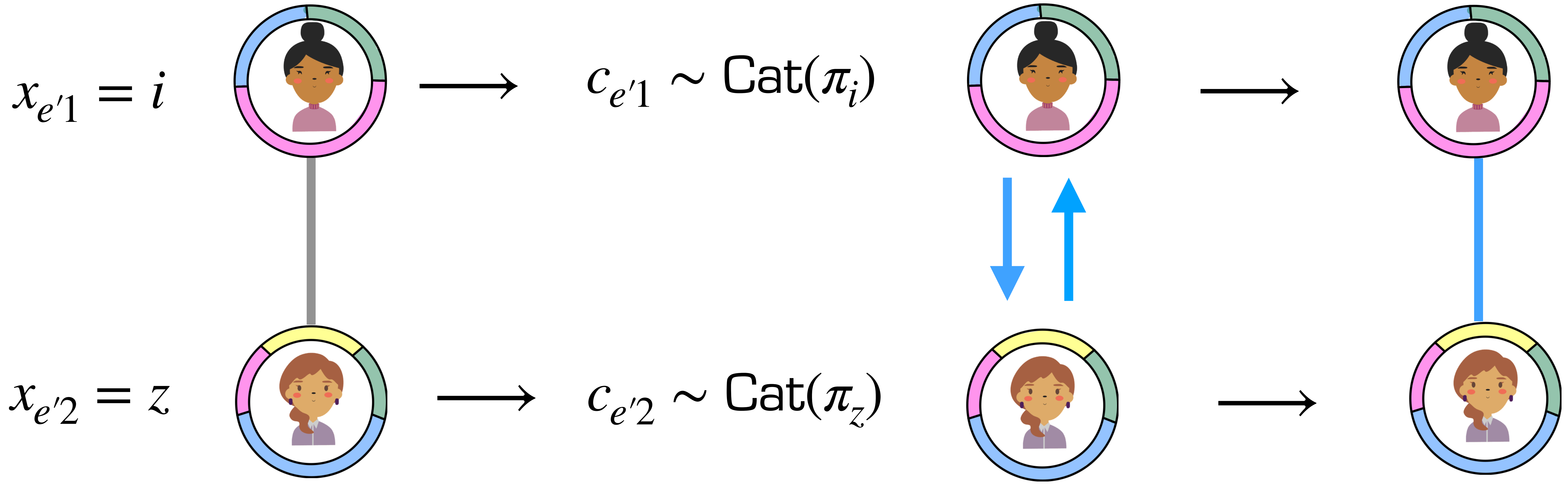
$$x_{e'2} = z \longrightarrow c_{e'2} \sim \text{Cat}(\pi_z)$$



Proposed model

3. For each edge, we assign nodes to communities:

3.1 **Keep edges** between nodes assigned to **the same communities:**



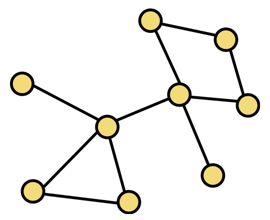
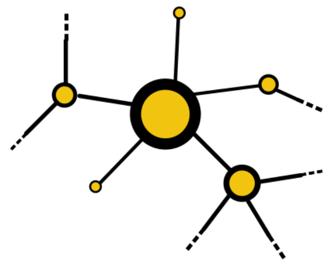
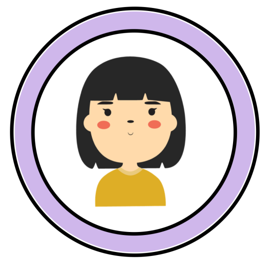
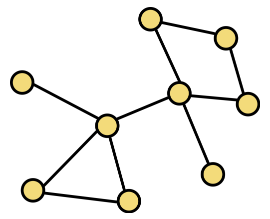
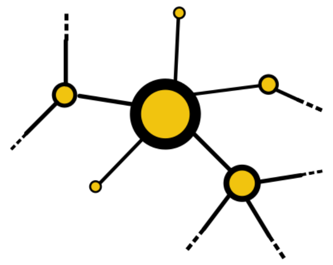
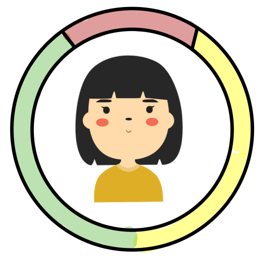
Overview

Motivation

Model

Results

RELATED MODELS

Model	Sparsity		Degree		Membership	
Sparse block models (Herlau et al. 2016)	**Sparse**		**Degree heterogeneity**		Single community membership	
Sparse mixed membership (Todeschini et al. 2020)	**Sparse**		**Degree heterogeneity**		**Mixed community membership**	

Posterior predictive results:

1. Data

Network name	Type	# nodes	# edges
Reed	online social network	962	18812
Simmons	online social network	1510	32984
SmaGri	co-authorship network	1024	4916
Yeast	Protein interaction	2224	6609

Posterior predictive results:

1. Data

Network name	Type	# nodes	# edges
Reed	online social network	962	18812
Simmons	online social network	1510	32984
SmaGri	co-authorship network	1024	4916
Yeast	Protein interaction	2224	6609

2. Models

- Thinned GGP (proposed)
- Sparse block model
- Sparse mixed membership
- Dense block model
- Dense mixed membership

Posterior predictive results:

1. Data

Network name	Type	# nodes	# edges
Reed	online social network	962	18812
Simmons	online social network	1510	32984
SmaGri	co-authorship network	1024	4916
Yeast	Protein interaction	2224	6609

2. Models

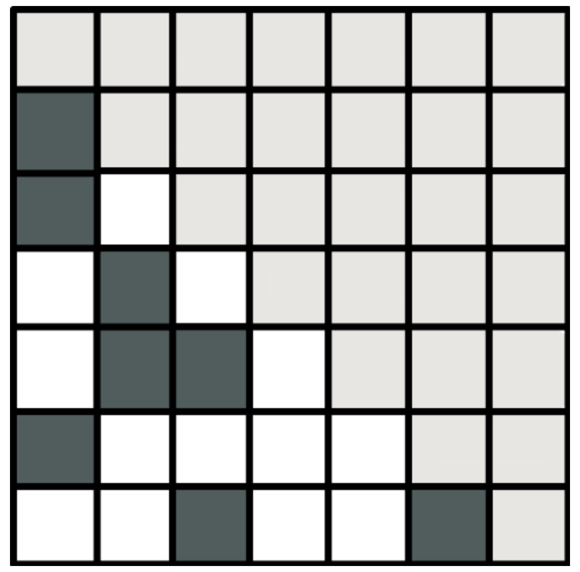
- Thinned GGP (proposed)
- Sparse block model
- Sparse mixed membership
- Dense block model
- Dense mixed membership

3. Evaluation

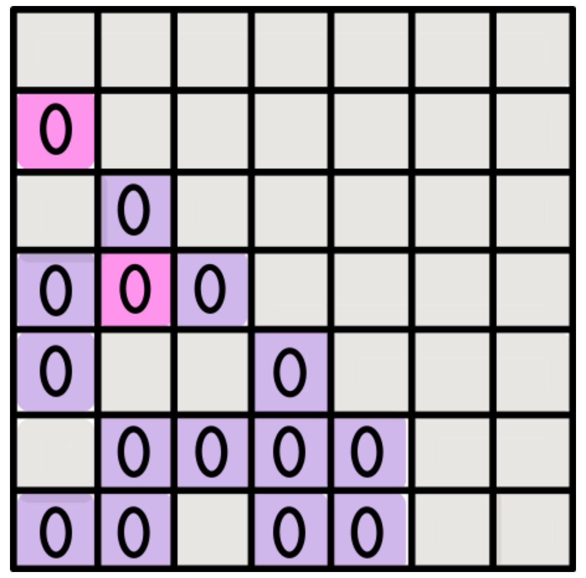
1. Fit model on fully observed data
2. Learn node-specific interaction parameters (e.g. nodes sociabilities and community memberships)
3. Use node-specific interaction parameters to predict edges (two prediction tasks)

Posterior predictive results: - Predict $Y_{ij} = 1$ among $Y_{ij} = 0$ (5% mislabeled)

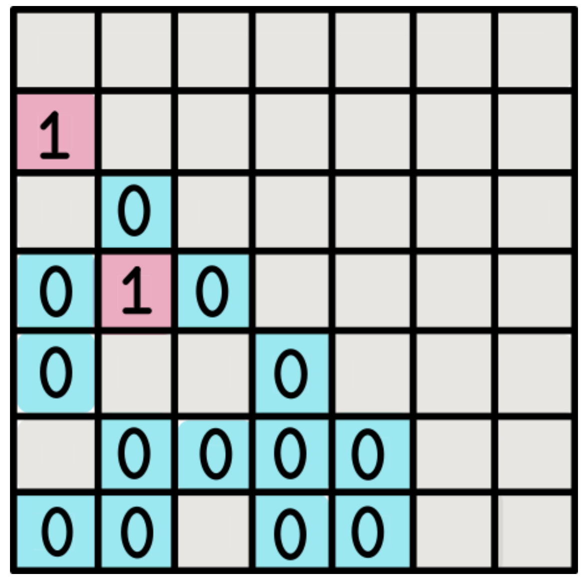
True



Task



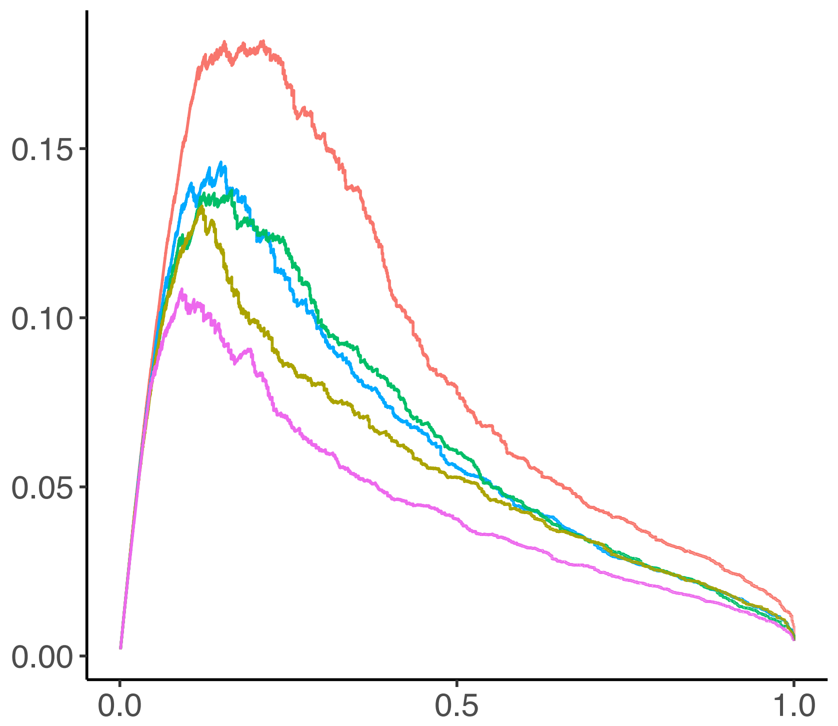
Perfect prediction



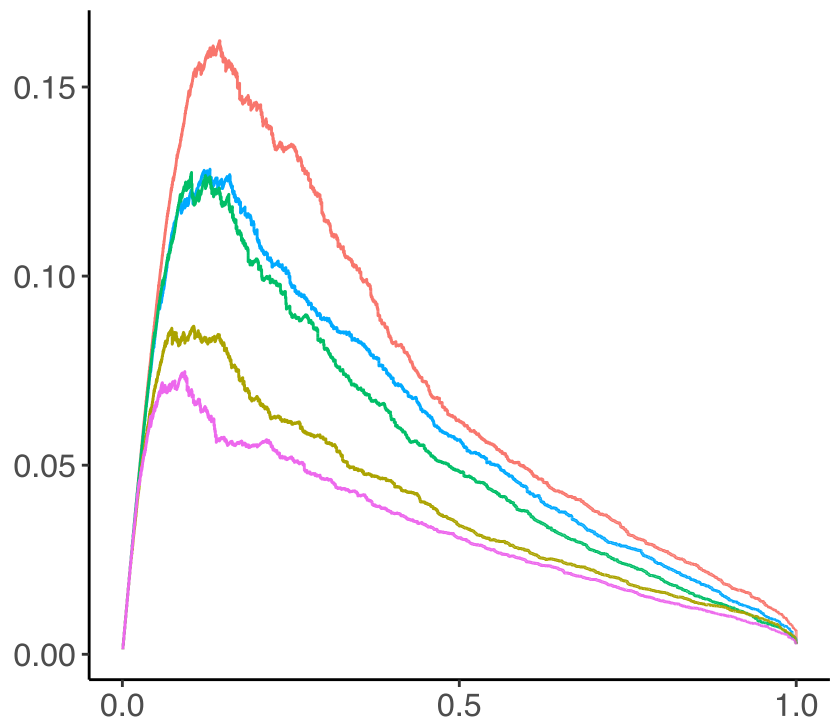
F-score vs. recall

- Thinned GGP (proposed)
- Sparse block model
- Sparse mixed membership
- Dense block model
- Dense mixed membership

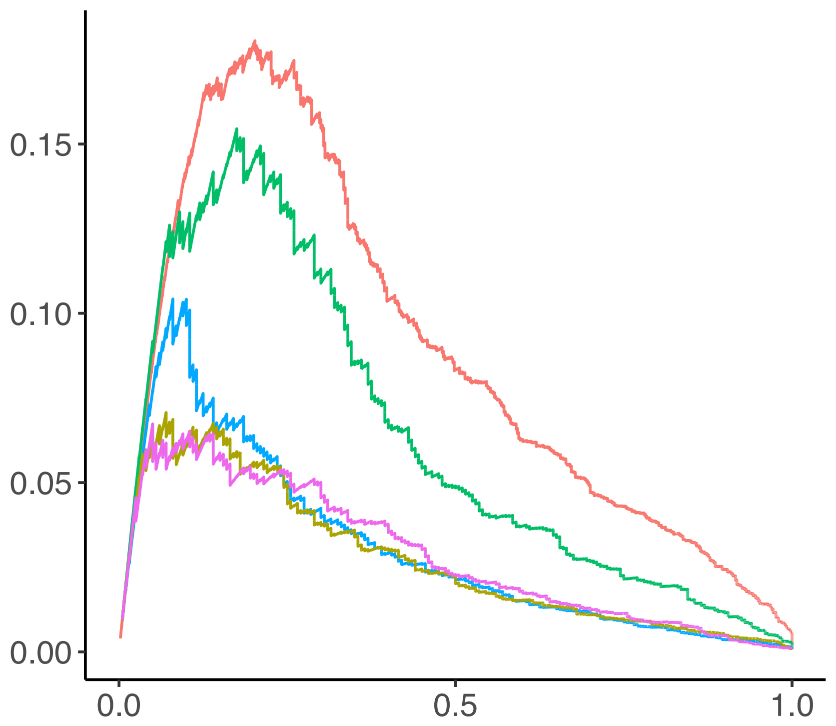
Reed



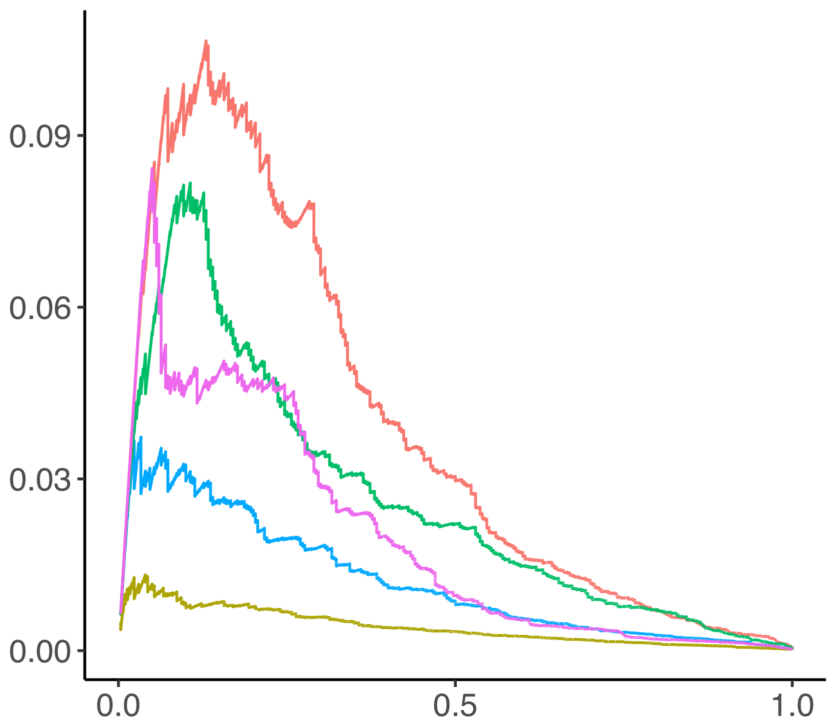
Simmons



SmaGri



Yeast



Overview

Motivation → Model → Results → Discussion

Limitations:

- Posterior inference sub-quadratic in number of nodes but too slow for very large networks (e.g. 100,000 nodes)
- Node-centered vs. edge-centered network models

Limitations:

- Posterior inference sub-quadratic in number of nodes but too slow for very large networks (e.g. 100,000 nodes)
- Node-centered vs. edge-centered network models

Future directions:

- Approximate posterior inference (for large networks)
- Model dynamically evolving networks

Acknowledgements

My PhD advisors

Erik Sudderth

University of California, Irvine

Michele Guindani

University of California, Los Angeles

(and the amazing people in their research groups)

Funding

Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

HPI Research Center in Machine Learning and Data Science at UCI

References

- **F.Z. Ricci, M. Guindani, E. Sudderth. Thinned completely random measures for sparse graphs with overlapping communities. Advances in Neural Information Processing Systems, 2022 (forthcoming).**
- F. Caron, and E.B. Fox. Sparse graphs using exchangeable random measures. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2017.
- Y.J. Wang, and G.Y. Wong. Stochastic blockmodels for directed graphs. Journal of the American Statistical Association, 1987.
- E.M. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. Advances in Neural Information Processing Systems, 2008.
- A. Todeschini, X. Miscouridou, and F. Caron. Exchangeable random measures for sparse and modular graphs with overlapping communities. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2020.
- T. Herlau, M.N. Schmidt, and M. Mørup. Completely random measures for modelling block-structured sparse networks. Advances in Neural Information Processing Systems, 2016.
- Y.W. Teh, M. Jordan, and M. Beal. Hierarchical Dirichlet processes. JASA, 2006.
- D.I. Kim, P.K. Gopalan, D. Blei, and E. Sudderth. Efficient online inference for bayesian nonparametric relational models. Advances in Neural Information Processing Systems, 2013.