From data to models

Tutors: Jan Bachmann and Lisette Espín-Noboa

ECML PKDD 2024 VILNIUS

bit.ly/snma-2024

Overview

Time: 11:20 - 13:00

11:20 - 12:05	Mitigating Biased Node Rankings ● A pre-processing intervention
12:05 - 12:50	Model selection ● A Bayesian approach
12:50 - 13:00	Closing remarks

Mitigating biased node rankings

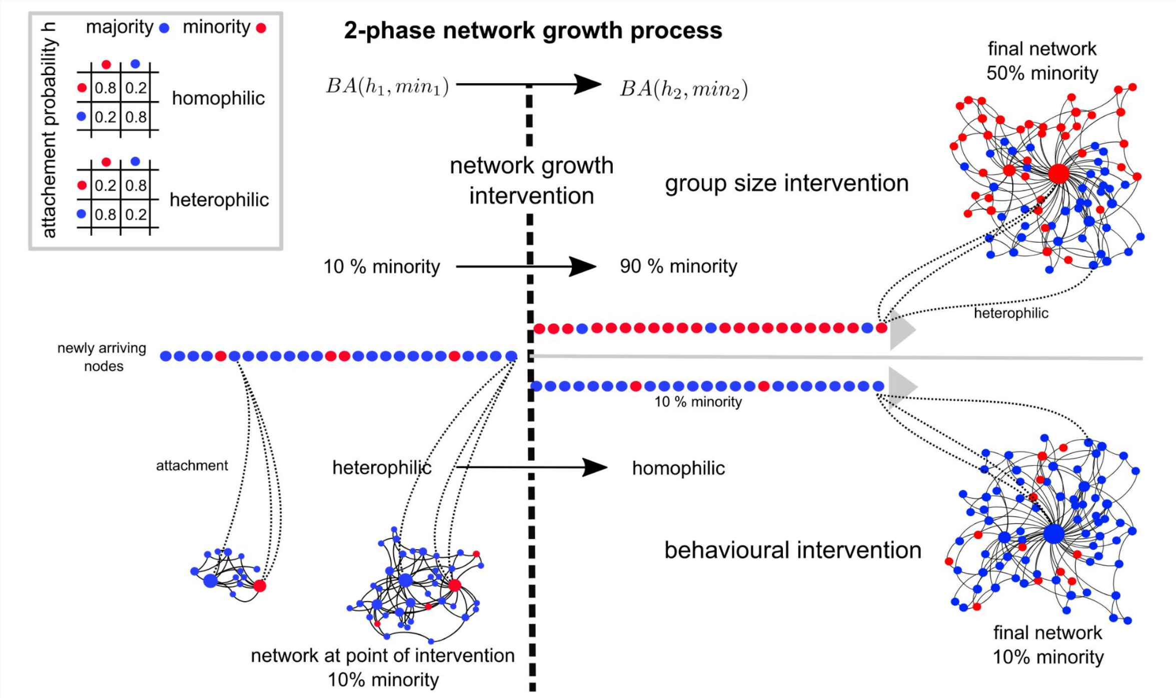
Literature

Non-exhaustive list of material covered in this section.

1. DiMaggio, P. & Garip, F. Network effects and social inequality. Ann. Rev. Sociol. 38,93–118 (2012).
2. Karimi, F., Oliveira, M. & Strohmaier, M. Minorities in networks and algorithms. Preprint at https://arxiv.org/abs/2206.07113 (2022).
3. Espín-Noboa, L., Wagner, C., Strohmaier, M. & Karimi, F. Inequality and inequity in network-based ranking and recommendation algorithms. Sci. Rep. 12, 2012 (2022).
4. Eccles, J. S. Bringing young women to math and science. In Gender and thought: psychological perspectives (eds. Crawford, M. & Gentry, M.) 36–58 (Springer, New York, NY, 1989).
5. Armstrong, M. A. & Jovanovic, J. Starting at the crossroads: intersectional approaches to institutionally supporting underrepresented minority women stem faculty. J. Women Minor. Sci. Eng. 21

Mitigating biased node rankings

- Social network structures contribute to the marginalization of minority groups,
- Impacts access to resources and visibility
- Historical underrepresentation and changing systems
    - Behavioral change
    - Increasing representation



Neuhäuser et al. "Improving the visibility of minorities through network growth interventions". *Commun Phys* **6**, (2023).

Exercise #3

Mitigating biased rankings

Task:

1. Implement a custom modeling class that implements two minority group and visualize the simulated networks.
2. Implement the model of Neuhäuser et al. Analyze and visualize how various parameter impact the visibility of the minority.

(30 min)

Open `3_exercise.ipynb`

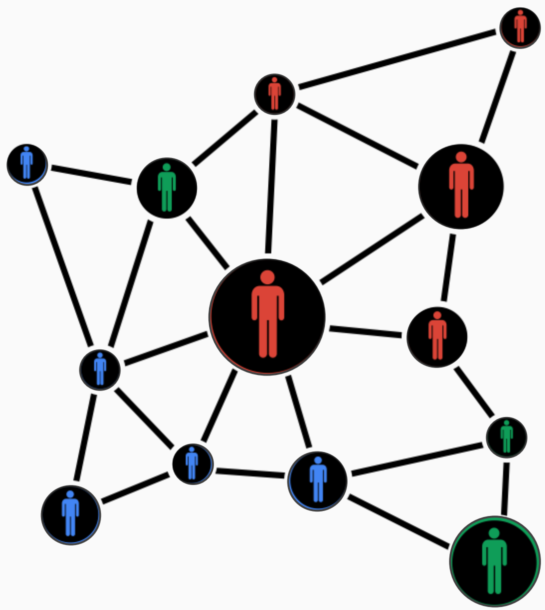
1. Alternatively, you can open the notebook from Google Colab (you need a Google account):
   bit.ly/snma2024-notebooks

Model selection

Literature

Non-exhaustive list of material covered in this section.

1. Espín-Noboa, L., Lemmerich, F., Strohmaier, M., & Singer, P. (2017). JANUS: A hypothesis-driven Bayesian approach for understanding edge formation in attributed multigraphs. Applied Network Science, 2, 1-20.
2. Contisciani, M., Hobbhahn, M., Power, E. A., Hennig, P., & De Bacco, C. (2024). Flexible inference in heterogeneous and attributed multilayer networks. arXiv preprint arXiv:2405.20918.
3. Safdari, H., Contisciani, M., & De Bacco, C. (2021). Generative model for reciprocity and community detection in networks. Physical Review Research, 3(2), 023209.



A real network

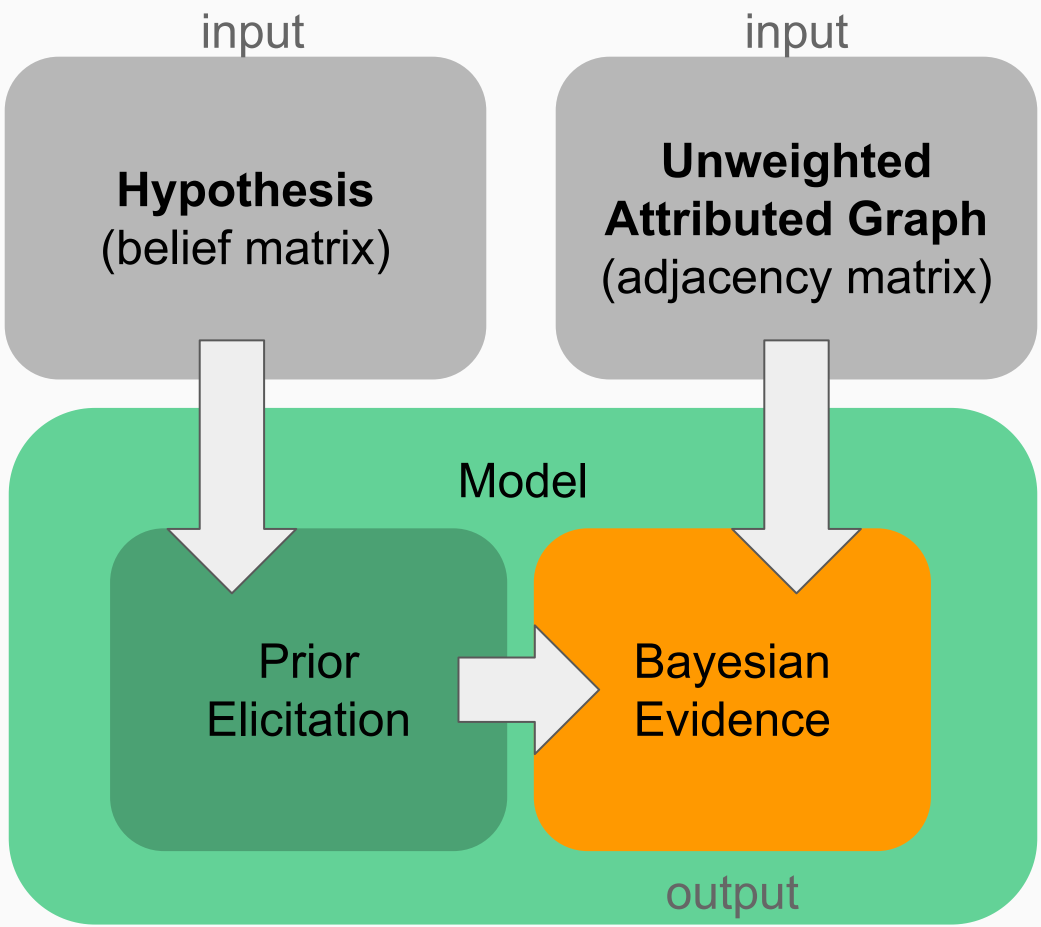
An attributed network

1. How do nodes (people) connect in this network?

2. What if we know some information about these nodes?

3. Can we leverage our "**prior beliefs**" to determine how these nodes connected in this network?

JANUS: A Bayesian approach for model selection

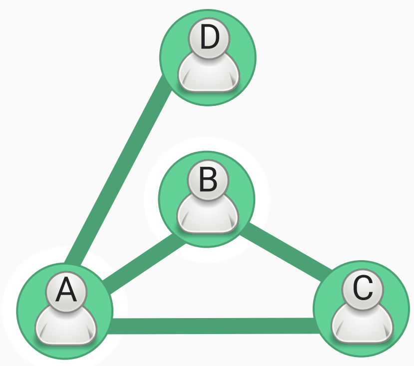


Espín-Noboa et al. 2017

Bayesian modeling

Edge Formation

- **Graph**
  Nodes and edges



	A	B	C	D
A	0	1	1	1
B	1	0	1	0
C	1	1	0	0
D	1	0	0	0

Adjacency matrix

- **Categorical Distribution**
  Each edge is sampled from a *categorical distribution*

$$(v_i, v_j) \sim \text{Categorical}(\theta)$$

Prior Elicitation

Expressing Hypotheses

- **Belief matrix**
  Our beliefs in edge formation as priors over the model parameters θ

B1: researchers from the same country are more likely to coauthor together

	A	B	C	D
A	0	0.9	0.9	0.1
B	0.9	0	0.9	0.1
C	0.9	0.9	0	0.1
D	0.1	0.1	0.1	0

(A) Lithuania
(B) Lithuania
(C) Lithuania
(D) Ecuador

- **Dirichlet Prior**
  Conjugate prior of Categorical distribution.

$$\alpha_{ij} = \frac{b_{ij}}{Z} \times \kappa + 1$$

Z: normalization constant

Espín-Noboa et al. 2017

Bayesian Evidence

Ranking of Hypotheses

- Bayes Factors to compare relative plausibility of hypotheses

$$\mathbf{BF} = \frac{\mathbf{P(D|H_1)}}{\mathbf{P(D|H_2)}}$$

$$\overbrace{P(\theta|D,H)}^{\text{posterior}} = \frac{\overbrace{P(D|\theta,H)}^{\text{likelihood}}\overbrace{P(\theta|H)}^{\text{prior}}}{\underbrace{P(D|H)}_{\text{marginal likelihood}}}$$

Bayes theorem

$$P(D|H) = \prod_{i=1}^{n} \frac{\Gamma(\sum_{j=1}^{n} \alpha_{ij})}{\Gamma(\sum_{j=1}^{n} \alpha_{ij} + m_{ij})} \prod_{j=1}^{n} \frac{\Gamma(\alpha_{ij} + m_{ij})}{\Gamma(\alpha_{ij})}$$

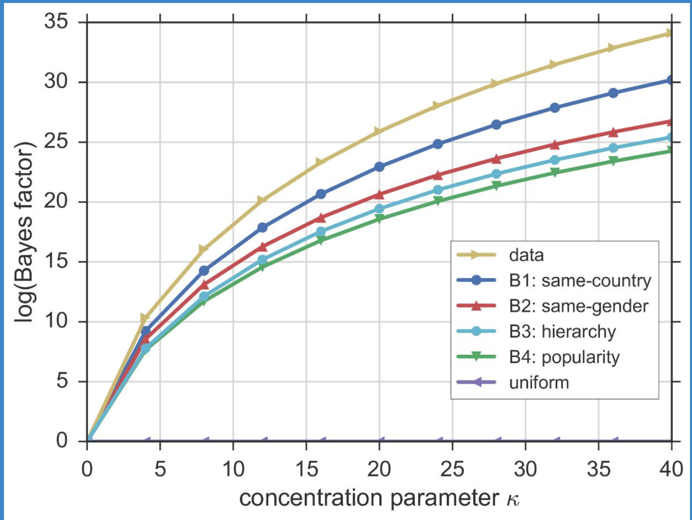
α_{ij} : prior (belief)
m_{ij}: number of actual edges in the graph

Espín-Noboa et al. 2017

Interpretation

Comparing Hypotheses

- B1: same country: **0.9, 0.1**
- B2: same gender: **0.9, 0.1**
- B3: hierarchy: **position_i * position_j**
- B4: popularity: **$\text{sum(articles+citations)}_{ij}$**
- uniform (baseline): random
- data: upperbound



Exercise #4

Model selection

Task:

1. Generate a synthetic graph of your choice
2. Generate the three baseline hypotheses: uniform, data, and self-loops
3. Generate hypothesis of your own using mechanisms of edge formation

(30 min)

Open **`4_exercise.ipynb`**

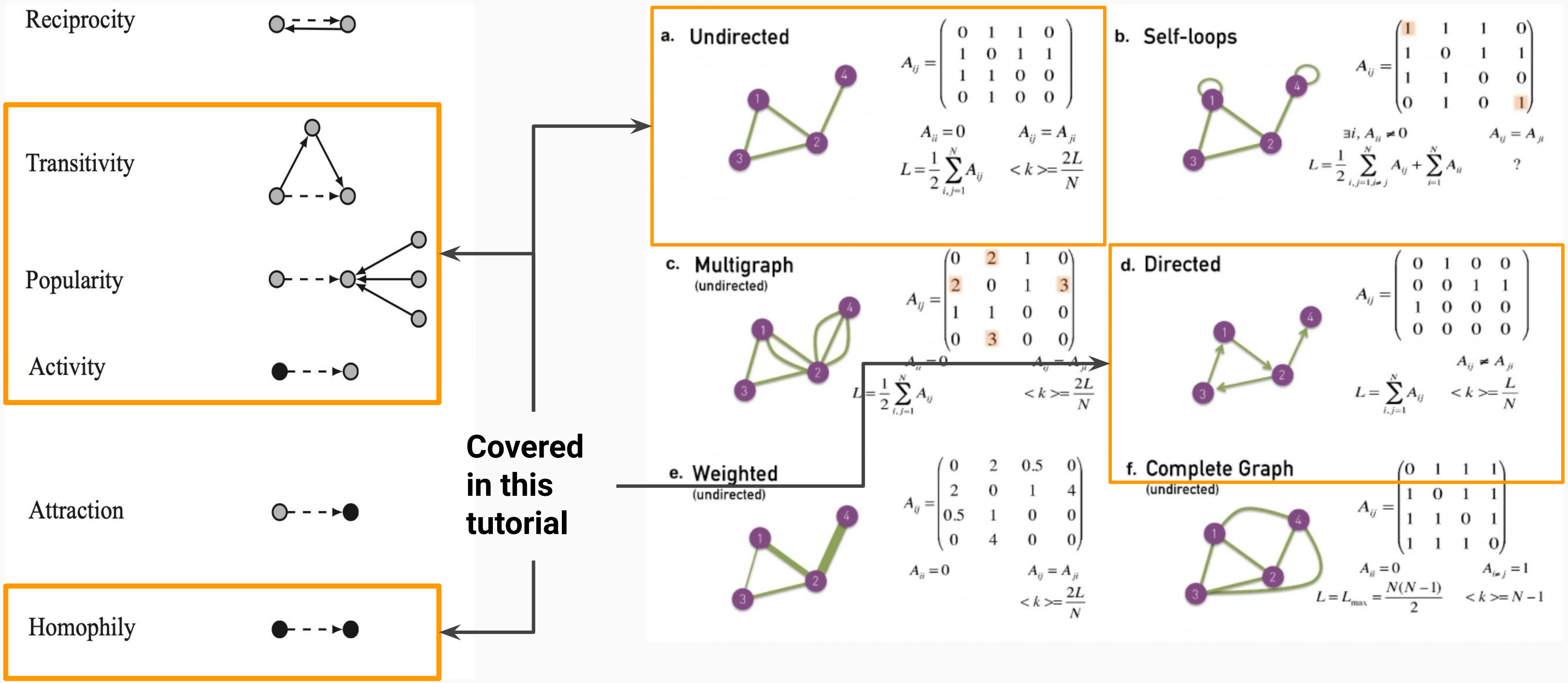
1. Alternatively, you can open the notebook from Google Colab (you need a Google account): [**`bit.ly/snma2024-notebooks`**](bit.ly/snma2024-notebooks)

Closing remarks

Challenges & open questions

Tutor: Lisette Espín-Noboa

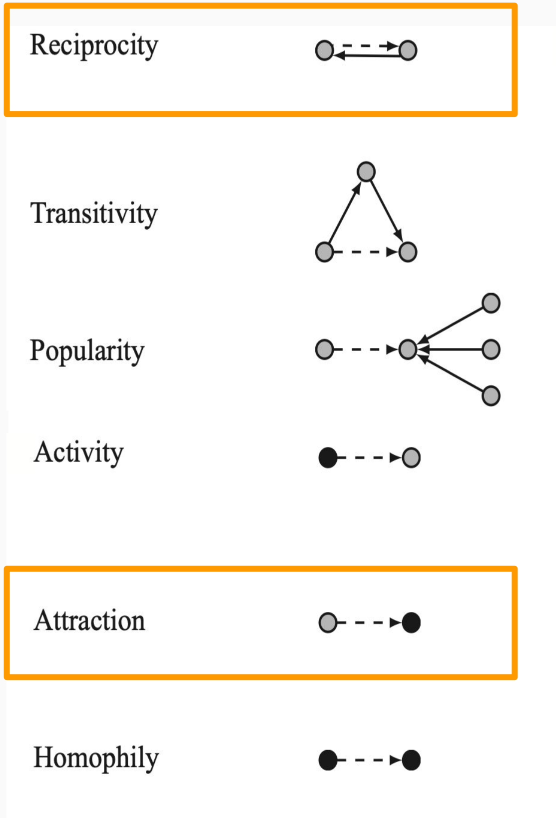
We need more realistic models!



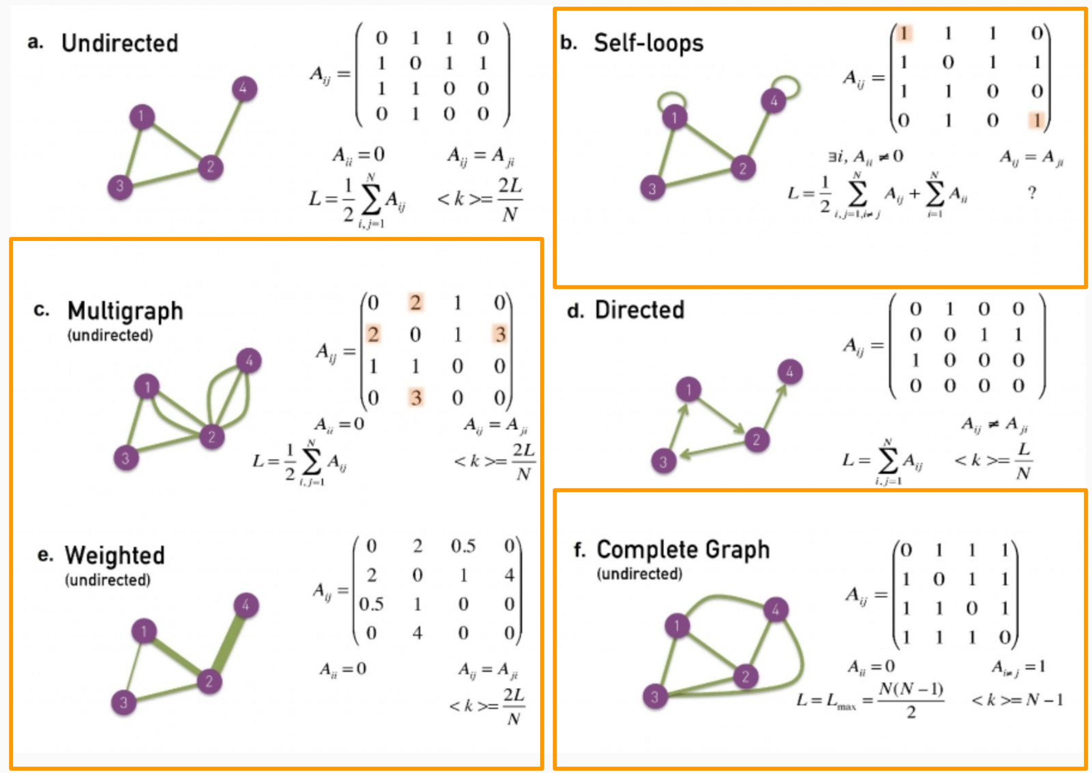
Stadfeld, Christoph, and Viviana Amati. "Network mechanisms and network models." Research Handbook on Analytical Sociology. Edward Elgar Publishing, 2021

http://networksciencebook.com/chapter/2#summary2

We need more realistic models!



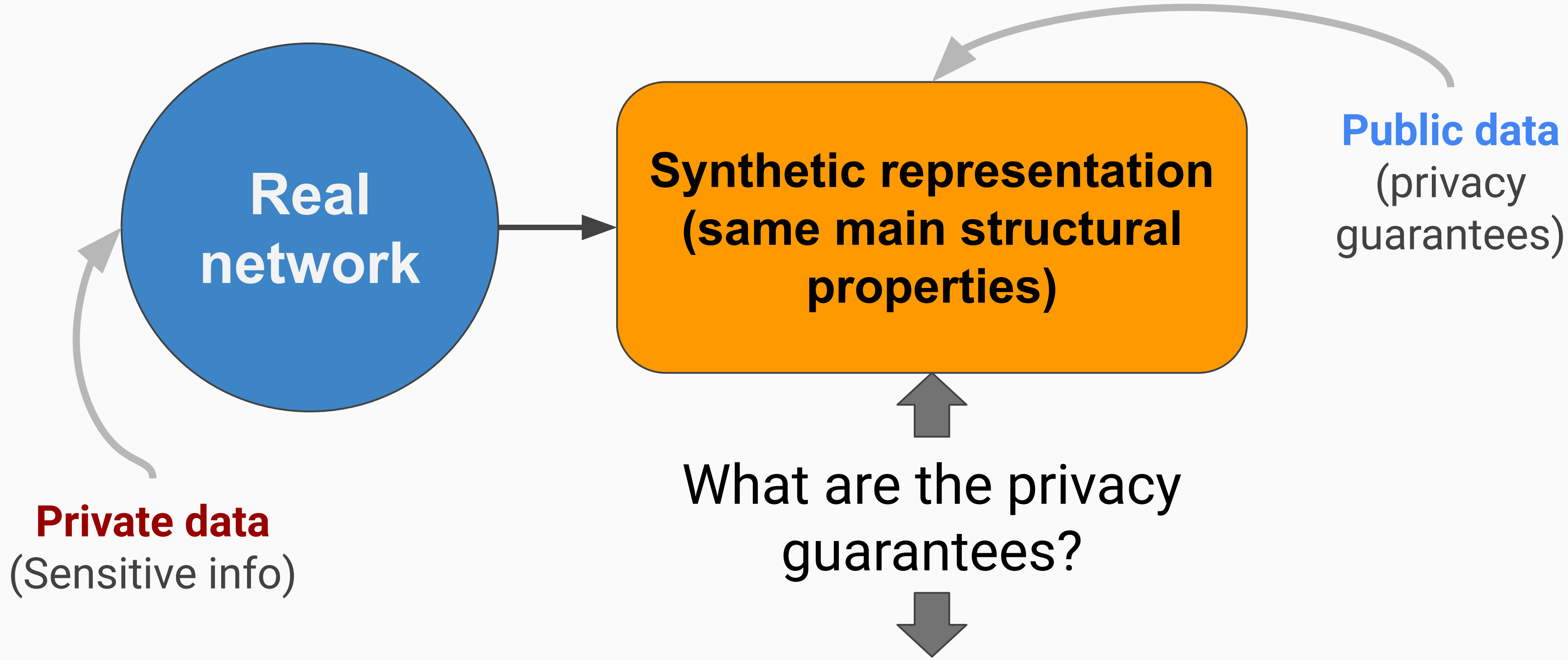
Stadfeld, Christoph, and Viviana Amati. "Network mechanisms and network models." Research Handbook on Analytical Sociology. Edward Elgar Publishing, 2021



What about these other types?

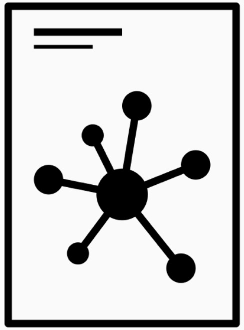
http://networksciencebook.com/chapter/2#summary2

Do synthetic networks solve privacy issues for data sharing?



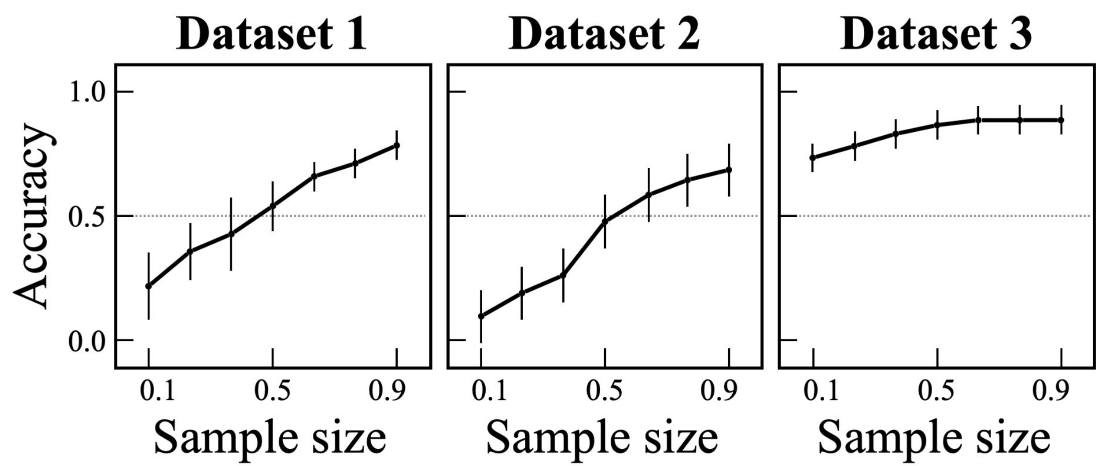
Can a particular (maybe rare) node attribute reveal the identity of an individual in the real network?

Machine learning on network data must be audited thoroughly with synthetic graphs!



Smith et al.
Novel **node classification algorithm** outperforms state-of-the-art algorithm X. Top-tier Venue (2024).

5.3. Results



Evaluating your algorithms on benchmark datasets is NOT enough if we want to understand the WHY of their outcomes!

1. The larger the training sample, the better the accuracy
2. Accuracy “seems” to correlate w/ net. structure
3. It “seems” to work best for assortative & directed net.
4. What about other types of networks?

We appreciate your feedback.
Thank you very much!

bitly

bit.ly/snma2024-survey