

From social theories to models

Tutors: Lisette Espín-Noboa and Jan Bachmann



bit.ly/snma-2024

Overview

Time: 09:15 - 11:00

- 09:15 - 09:30 Social theories
- Popularity
 - Similarity
 - Friend-of-friend
 - Activity

- 09:30 - 09:40 Network properties & structure
- Graphology
 - Properties at varying scales
 - From link formation to structure

- 09:40 - 10:15 Network Models
- Modeling networks by link formation
 - *Exercise 1: netin graph generation*

10:15 - 11:00 *Exercise 2: Auditing node rankings*

Social theories

Literature

Non-exhaustive list of material covered in this section.

1. Jackson, M. O. (2008). Social and economic networks (Vol. 3). Princeton: Princeton university press.
2. Albert-László Barabás (2016). Network Science. (available online as an interactive book)
3. Jackson, M. O. (2019). The human network: How your social position determines your power, beliefs, and behaviors. Vintage.
4. Stadfeld, C., & Amati, V. (2021). Network mechanisms and network models. In Research Handbook on Analytical Sociology (pp. 432-452). Edward Elgar Publishing.
5. Gamper, M. (2022). Social Network Theories: An Overview. *Social Networks and Health Inequalities*, 35.
6. Karimi, F., & Oliveira, M. (2022). On the inadequacy of nominal assortativity for assessing homophily in networks. arXiv preprint arXiv:2211.10245.

What are social networks?

Social networks consists of **actors** (nodes/agents) and **relations** (edges/links)

- Actor → person, company, country
 - It may also include entities produced or generated by individuals, e.g., email, retweet.
- Relation → friendship, collaboration, partnership, trade, war, hierarchy
 - When the actor is a resource, the relation may represent a spreading dynamic, access to opportunities, etc.



How do networks form?

Social mechanisms in micro-level govern how we organize our social lives and interact with others. Some of those social theories include: group identity theory, cultural evolutions, and broadly social psychology theories.

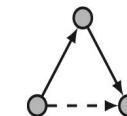
Prevalent **mechanisms** identified in social science literatures:

- Preferential attachment a.k.a. rich-get-richer or Matthew effect
- Homophily and assortative mixing
- Social balance and triadic closure

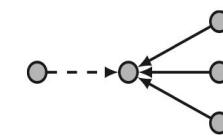
Reciprocity



Transitivity



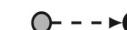
Popularity



Activity



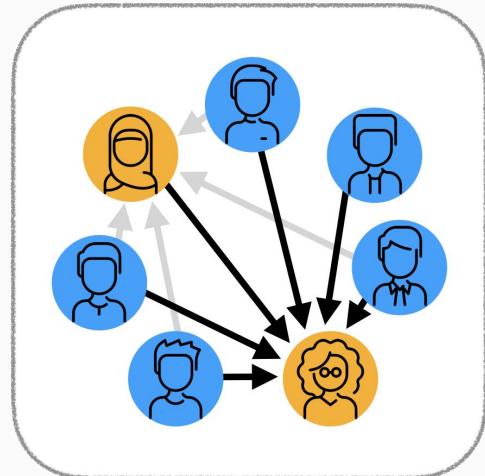
Attraction



Homophily



Covered
in this
tutorial



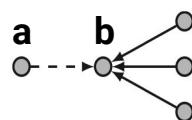
Popularity

The tendency of actors to connect to those who receive ties from many others.

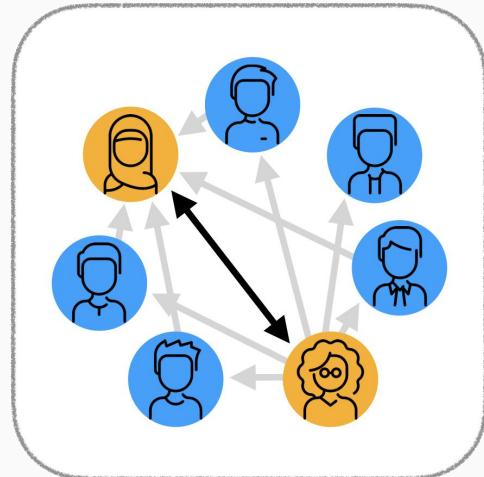
The structural position is defined by the **in-degree of the receiver** (target node) of an explained tie.

- **Matthew effect or rich-gets-richer effect** mechanism in science discussed by Merton (1968), Price (1976) and others. Increasing recognition of an actor's scientific work (e.g. number of ties in a citation network).
- **Preferential attachment** in the work of Barabási and Albert (1999) operationalised this mechanism in networks.

The presence of a popularity mechanism is not sufficient evidence that a Matthew effect exists in a network, e.g., other actor **attributes**, such as the career stage or gender, might send similar signals that affect the beliefs of others.



$$P(a \rightarrow b) = P(b|a) = p_{ab} = \frac{k_b}{\sum_c k_c}$$



Homophily

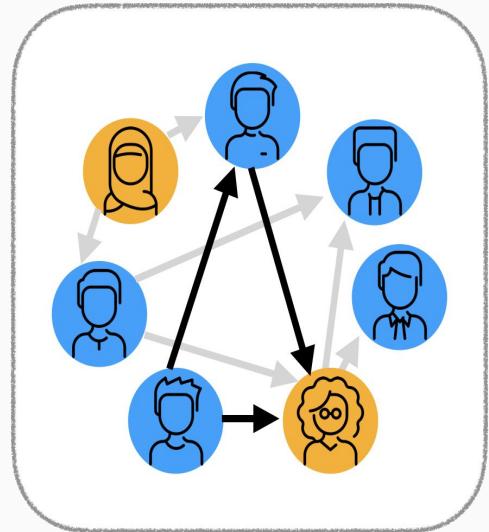
The tendency that similar actors are more likely to connect (Lazarsfeld, et al. 1954; McPherson et al. 2001). Its structural position is defined by **the attribute similarity of the sender and receiver** of the explained tie.

Causal explanations:

- These may be cognitive processes about similarity attraction (Huston and Levinger 1978)
- Structural processes that are affected by existing baseline segregation and social distances in the social setting under study.

$$\begin{matrix} a & b \\ \bullet & \cdots & \bullet \end{matrix}$$

$$P(a \rightarrow b) = P(b|a) = p_{ab} = \text{similarity}(a, b)$$

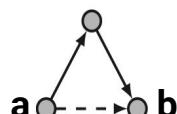


Transitivity

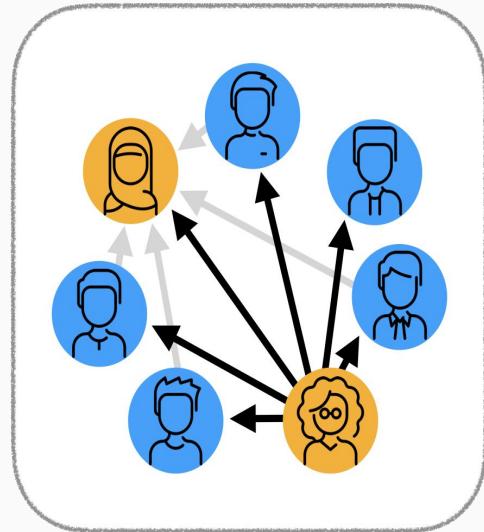
One of the central mechanisms that goes *beyond the dyad* and thus *involves more than two actors*: it expresses the tendency of actors to connect to others to whom they are indirectly tied to through a third actor. The structural position is defined by the presence of a “**two-path**” **between the sender and the receiver** of the explained tie.

Causal explanations:

- Differences in spatial or social distances and similarity attraction based on an actor attribute (Granovetter 1973).
- Actors may perceive cognitive dissonance and stress if they are not in a positive relationship with those they are indirectly positively tied to (Heider 1958).



$$P(a \rightarrow b) = P(b|a) = p_{ab} = p_{TC} \propto \text{clustering}(g)$$



The tendency that actors with specific attributes will be more likely to send ties. The structural positions are defined by **attributes of the sender** of the explained tie. Whether these differences relate to differences in networking resources is not further specified by the mechanism.

Node Activity

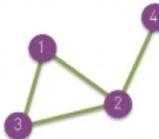
a •---• b

$$P(a \rightarrow b) = P(b|a) = p_{ab} \propto \text{activity}(a)$$

Network properties and structure

Graphology (elementary property of the underlying graph)

a. Undirected

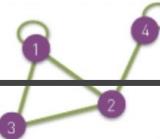


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

b. Self-loops

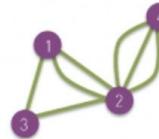


$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$\exists i, A_{ii} \neq 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii} \quad ?$$

c. Multigraph (undirected)

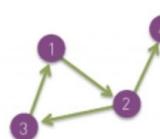


$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

d. Directed

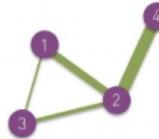


$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{ij} \neq A_{ji}$$

$$L = \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{L}{N}$$

e. Weighted (undirected)

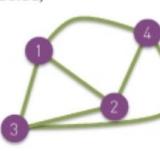


$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$\langle k \rangle = \frac{2L}{N}$$

f. Complete Graph (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = 1$$

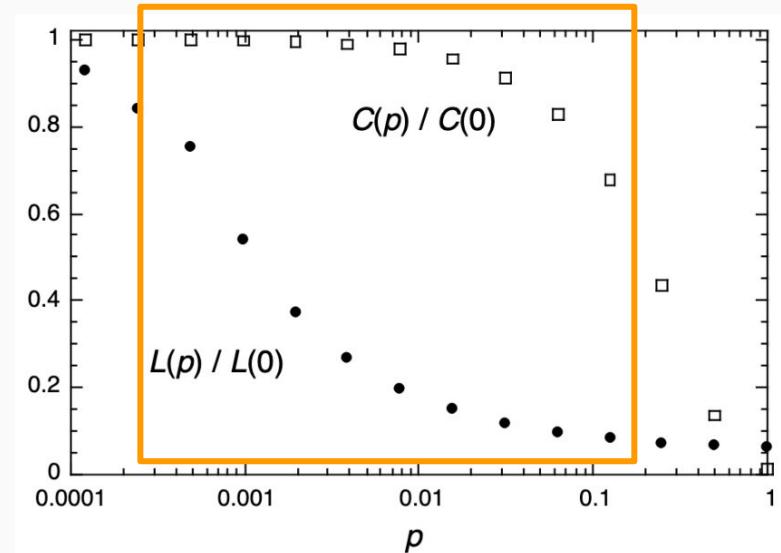
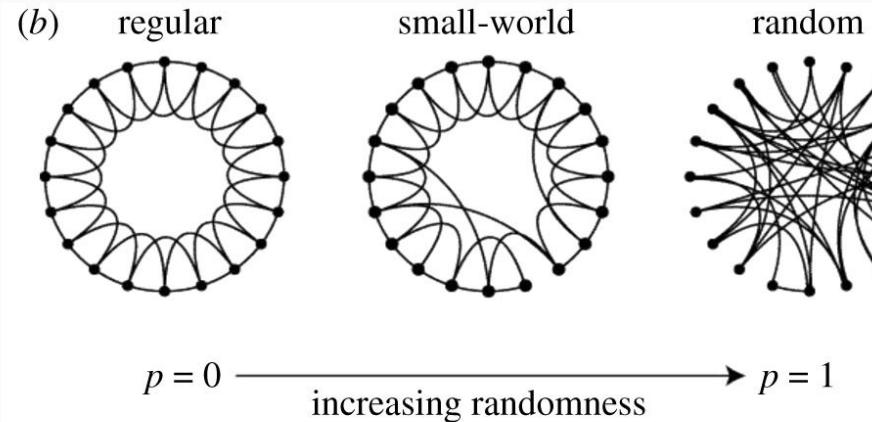
$$L = L_{\max} = \frac{N(N-1)}{2} \quad \langle k \rangle = N - 1$$

Covered
in this
tutorial

Important properties of social networks at varying scales

Graph	Node	Edge
Average and minimum degree (note: average degree does not make sense if you have power-law degree distribution)	Degree (in/out)	Weights
Degree distribution (the probability that a randomly selected node in the network has degree k)	Centrality (e.g., PageRank)	Shortest path
Adjacency matrix / Mixing matrix	Clustering coefficient	Homophilous type (MM, Mm, mm, mM)
Sparsity (out of the total possible number of edges, how many actually exist)	Activity (high activity = high outdegree)	
Diameter		
Average Path Length		
Connected components		

Common structure (topology) of social networks

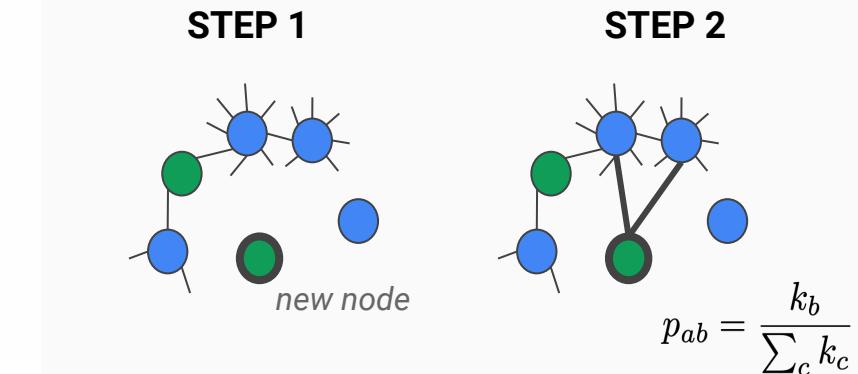
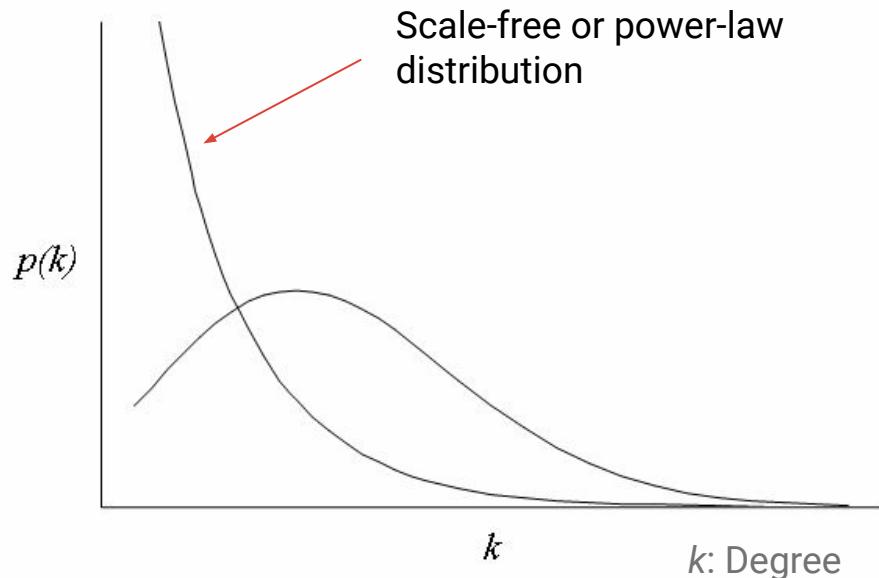


C: Clustering

L: Average shortest path length

Clustering and small-world property achievable (by random rewiring)

Common structure (topology) of social networks



Scale-free degree distributions (by preferential attachment)

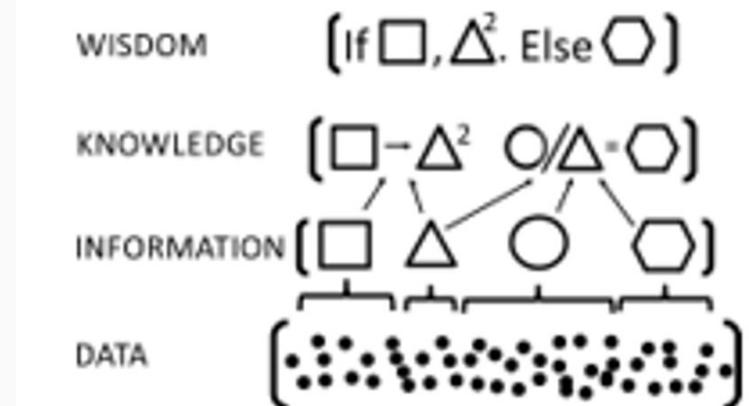
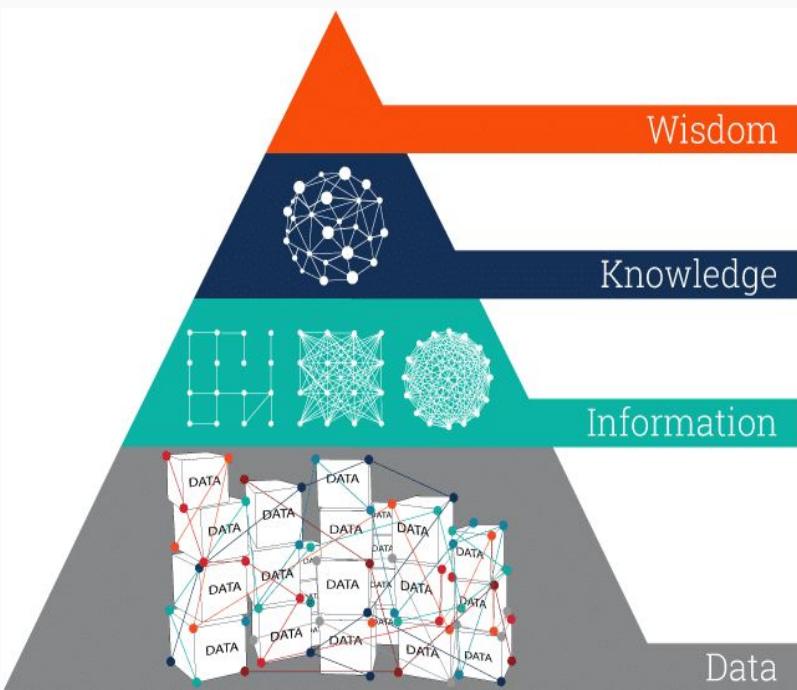
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509-512.
Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, 10(1), 1017.

Network models

... and where to find them

Why modeling social networks?

The wisdom hierarchy



The seven uses of Models (REDCAPE)

1. **Reason**: to identify conditions and deduce logical implications
2. **Explain**: To provide explanations for empirical phenomena
3. **Design**: to choose features of policies, institutions and rules
4. **Communicate**: to relate knowledge and understandings
5. **Act**: to guide policy choices and strategic actions
6. **Predict**: to make numerical and categorical predictions of future and unknown phenomena
7. **Explore**: to investigate possibilities and hypotheticals scenarios

Why modeling social networks?

Mechanistic models refers to the use of social mechanisms to design and develop network models with the goal of understanding their effects on our social world and algorithms.

Holme, P., & Liljeros, F. (2015). Mechanistic models in computational social science. *Frontiers in Physics*, 78.

Why network models in AI and ML?

Data can be only one realization of the social structure. Generating realistic synthetic social networks can help us scrutinize the robustness and fairness of ML algorithms when data is biased or not representative of the reality.

Steinbacher, M., Raddant, M., Karimi, F., Camacho Cuena, E., Alfarano, S., Iori, G., & Lux, T. (2021). Advances in the agent-based modeling of economic and social behavior. *SN Business & Economics*, 1(7), 99.

Class of models in Network Science

Covered in
this tutorial

Model Class	Examples	Characteristics
Static Models	Erdos–Rényi Watts-Strogatz	<ul style="list-style-type: none">• N fixed• p_k exponentially bounded• Static, time independent topologies
Generative Models	Configuration Model Hidden Parameter Model	<ul style="list-style-type: none">• Arbitrary pre-defined p_k• Static, time independent topologies
Evolving Network Models	Stochastic Block Model Barabási–Albert Model Bianconi-Barabási Model Initial Attractiveness Model Internal Links Model Node Deletion Model Accelerated Growth Model Aging Model	<ul style="list-style-type: none">• p_k is determined by the processes that contribute to the network's evolution.• Time-varying network topologies

Table 6.1

Classes of Models in Network Science

The table summarizes the three main modeling frameworks used in network science, together with their distinguishing features.

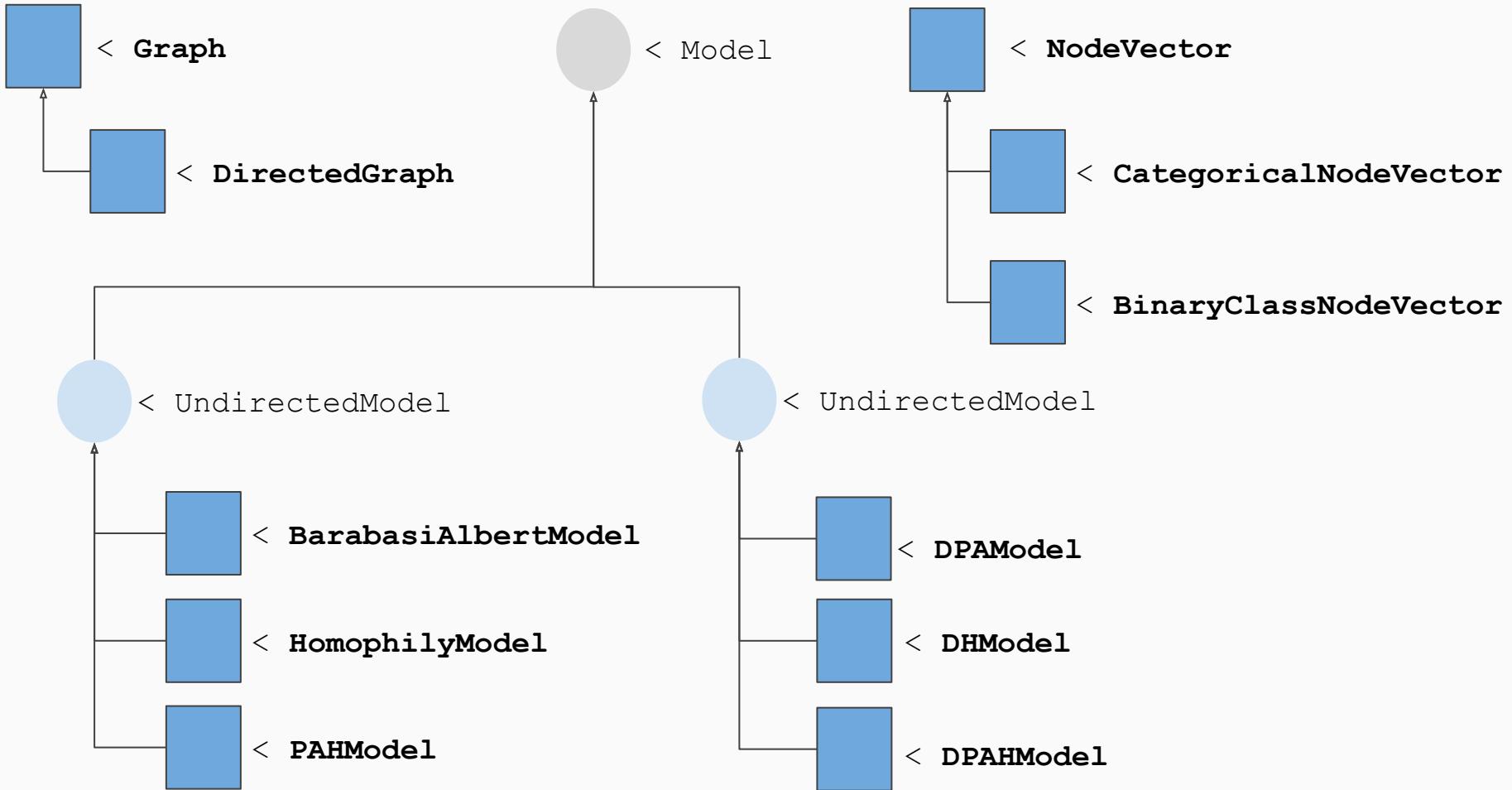
Pre-requisites

Follow these instructions in case you are using your own environment to run the exercises.

Alternatively, you can run all the exercises in Google Colab (more details later).

1. Download and install conda
[condo.io/projects/conda/en/stable/user-guide/install/download.html](https://conda.io/projects/conda/en/stable/user-guide/install/download.html)
2. Create an environment with python 3.9
`conda create -n "ecmlpkdd" python=3.9`
3. Activate your newly created conda environment
`conda activate ecmlpkdd`
4. Clone the tutorial in your computer
`git clone`
<https://github.com/snma-tutorial/ecmlpkdd2024.git>
5. Install the additional dependencies
`conda install pip`
`pip install -r requirements.txt`

The `netin` python package (alpha)



Undirected networks

BarabasiAlbertModel

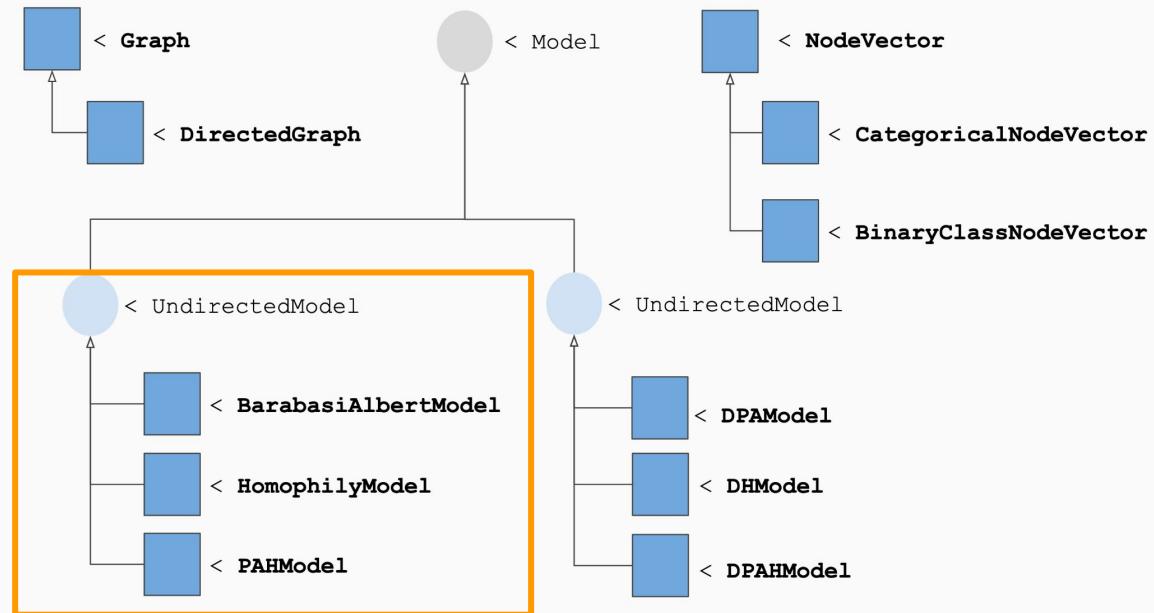
preferential attachment

PAHModel

PA + homophily

PATCHModel

PAH + triadic closure



`class BarabasiAlbertModel(...)`

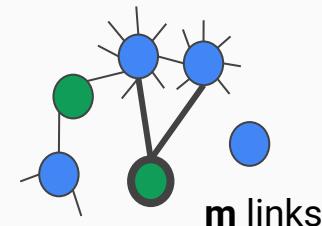
Parameters

`N (int)`

- number of nodes to be added to the network.

`m (int)`

- number links per new node

STEP 1:
 $m = 2$ **STEP 2:***targets: existing nodes**new node***STEP 3 (until N):**

$$p_{ab} = \frac{k_b}{\sum_c k_c}$$

```
class PAHModel(N, m, f_m, h_m, h_M)
```

Parameters

f_m (float)

- Fraction of minority nodes in the network

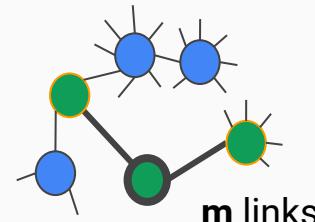
h_m (float)

- Homophily of the minority group

h_M (float)

- Homophily of the majority group

STEP 3 (until N):



$$p_{ab} = \frac{h_{ab}k_b}{\sum_c h_{ac}k_c}$$

class PATCHModel(...)

Parameters

`p_tc` (float)

- Probability to perform a triadic closure link

`lfm_local` (“Uniform”, “Homophily”, “PAH”)

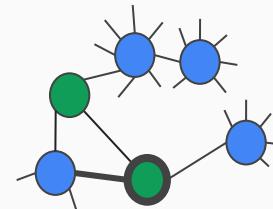
- How to choose target nodes locally

`lfm_global` (“Uniform”, “Homophily”, “PAH”)

- How to choose target nodes globally

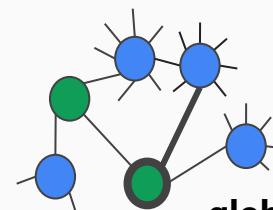
STEP 3 (until N):
m times

Case 1: p_{TC}



local link (restricted to friends of friends)

Case 2: $(1 - p_{TC})$



global link (any node available)

Directed networks

DPAModel

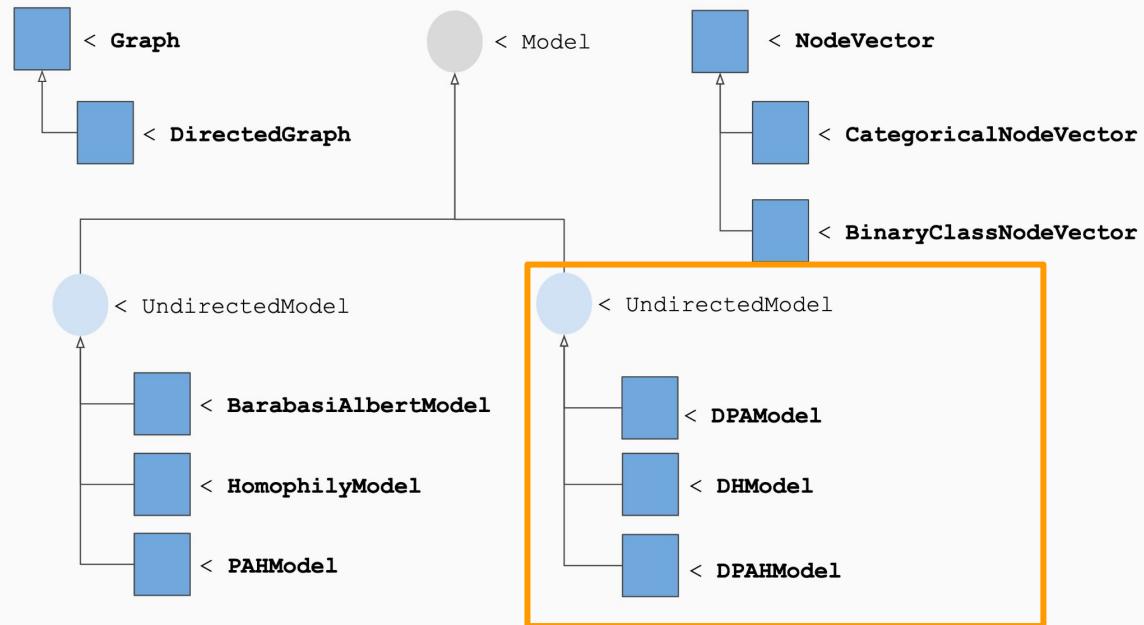
preferential attachment

DHModel

homophily

DPAHModel

PA+H



```
class DPAModel(N, d, f_m, plo_m, plo_M)
```

Parameters

- d (float): Target network density
(number of links out of all possible links)
- plo_m (float): Out-degree activity of minority group (power law exponent)
- plo_M (float): Out-degree activity of majority group (power law exponent)

Do while density < d:

Step 1: choose a **source** node
(using the activity of all nodes)

Step 2: choose a **target** node *
(using preferential attachment → in_degree)

* target nodes must have out_degree > 0
(should have been picked as source at least once)

class DHModel(...)

Parameters

h_m (float)

- Homophily of the minority group

h_M (float)

- Homophily of the majority group

Do while density < d:**Step 1:** choose a **source** node
(using the activity of all nodes)**Step 2:** choose a **target** node *
(using mixing matrix → homophily within/across groups)* target nodes must have out_degree > 0
(should have been picked as source at least once)*class DPAHModel(...):***Step 2:** choose a **target** node *
(using mixing matrix and preferential attachment)

Exercise #1

Graph generation

Task:

1. Simulate existing network models.
2. Retrieve and analyze node attributes
3. Study the effect of homophily on network segregation
4. Inject your own code to track the growth of in- and out-group links

Repeat for other graphs and check the effects of other mechanisms.

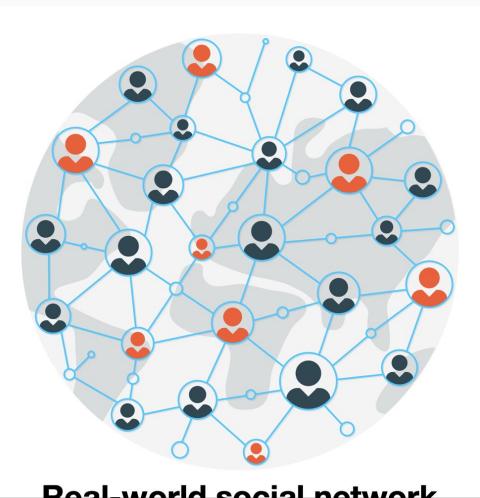
(until 10:15)

Open **1_exercise.ipynb**

1. Alternatively, you can open the notebook from Google Colab (you need a Google account):

bit.ly/snma2024-notebooks

Ranking inequalities



1. Identify network structure

Fraction min.
 $f_m=0.3$

Node activity
 $\gamma_M = \gamma_m = 3$

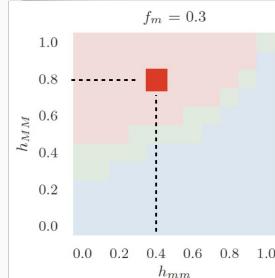
Density
 $d=0.0015$

Homophily Maj.
 $H_{MM}=0.8$

Homophily min.
 $H_{mm}=0.4$

(Inequity is driven by homophily and fraction of minorities)

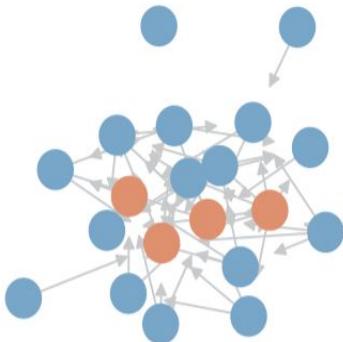
2. Identify inequality and inequity in ranking



On average minorities are under-represented in top-k's
(Interventions needed)

Given a network,

Heterophilic
 $h_{MM} = 0.2$
 $h_{mm} = 0.2$



$n = 20$
 $\min(m) = 20\% \text{ (fm)}$
 $\text{maj}(M) = 80\%$

... and a ranking of its nodes

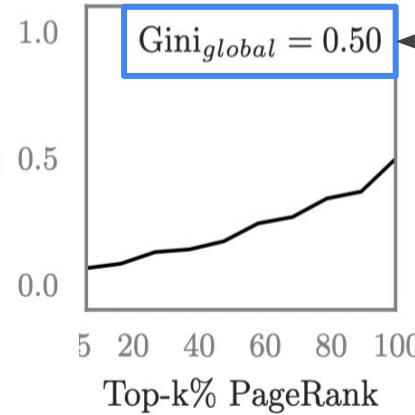
Ranked nodes
 (PageRank)



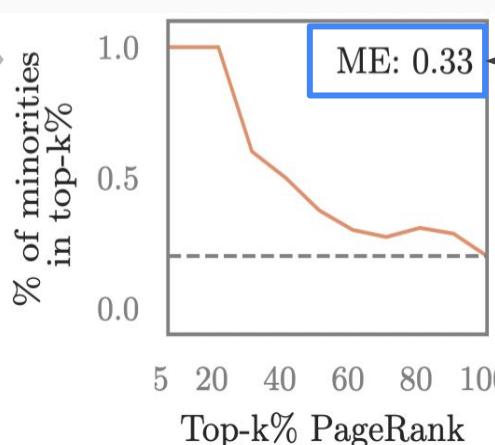
+

-

Measure:



Inequality

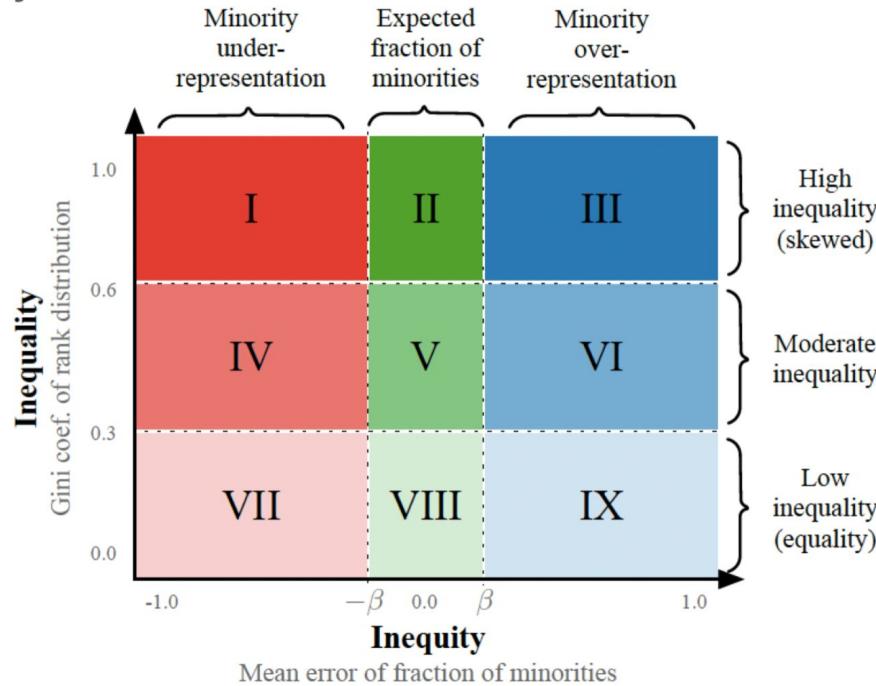
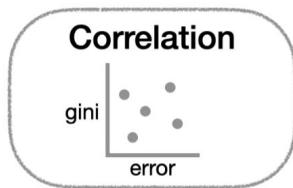


Disparity
 ME vs $\text{Gini}_{\text{global}}$

Inequity

Inequality vs. Inequity

Regions of disparity



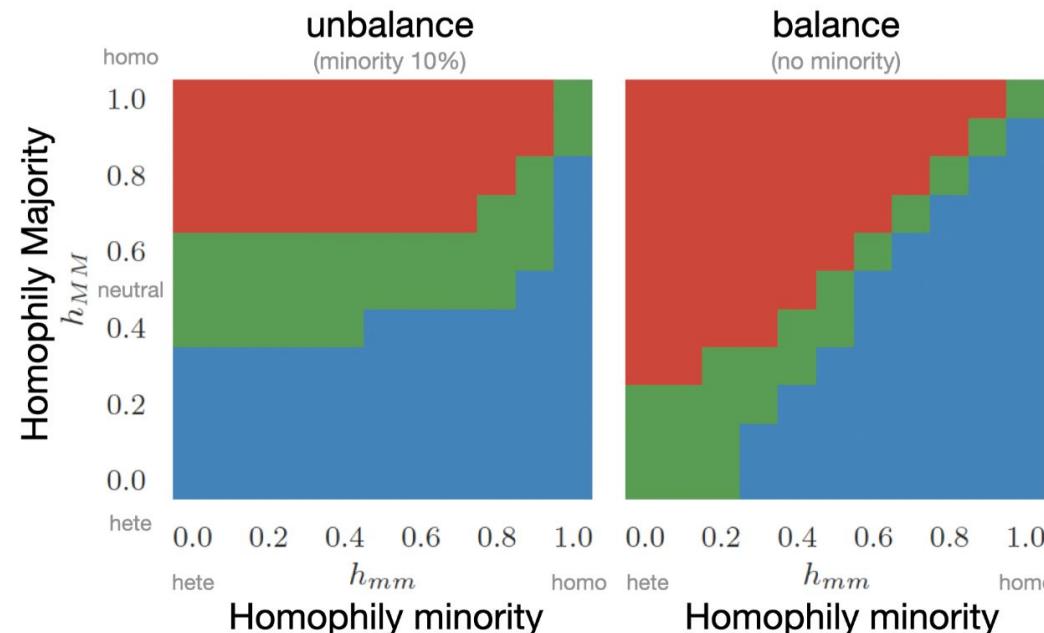
Inequality vs. Inequity in PageRank

As a function of Homophily and Fraction of minorities

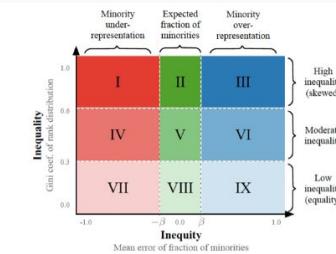
1. In balanced networks, both groups are well represented if $h_{mm} = h_{MM}$

2. In unbalanced networks, minorities are well represented when majority is neutral and the minority is not too homophilic.

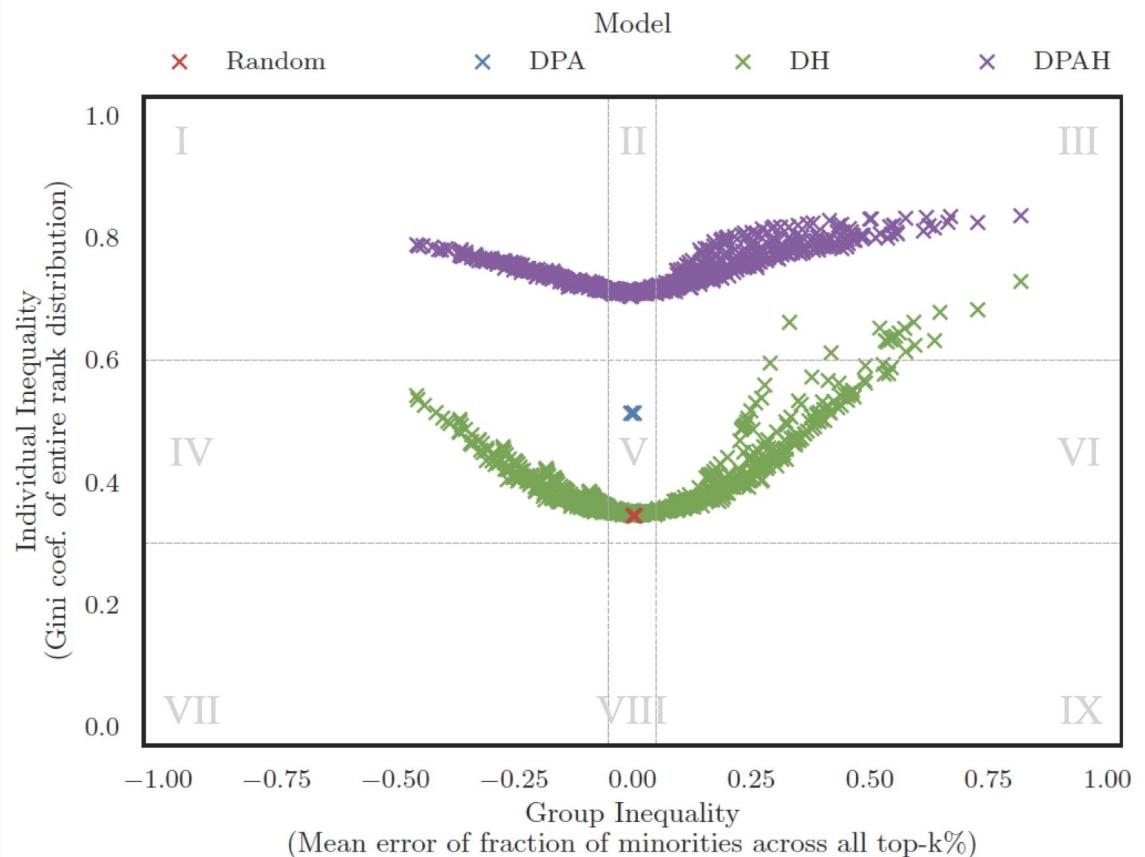
3. In unbalanced and homophilic networks, minorities are well represented when $h_{mm} > h_{MM}$.



Espín-Noboa et al. 2022

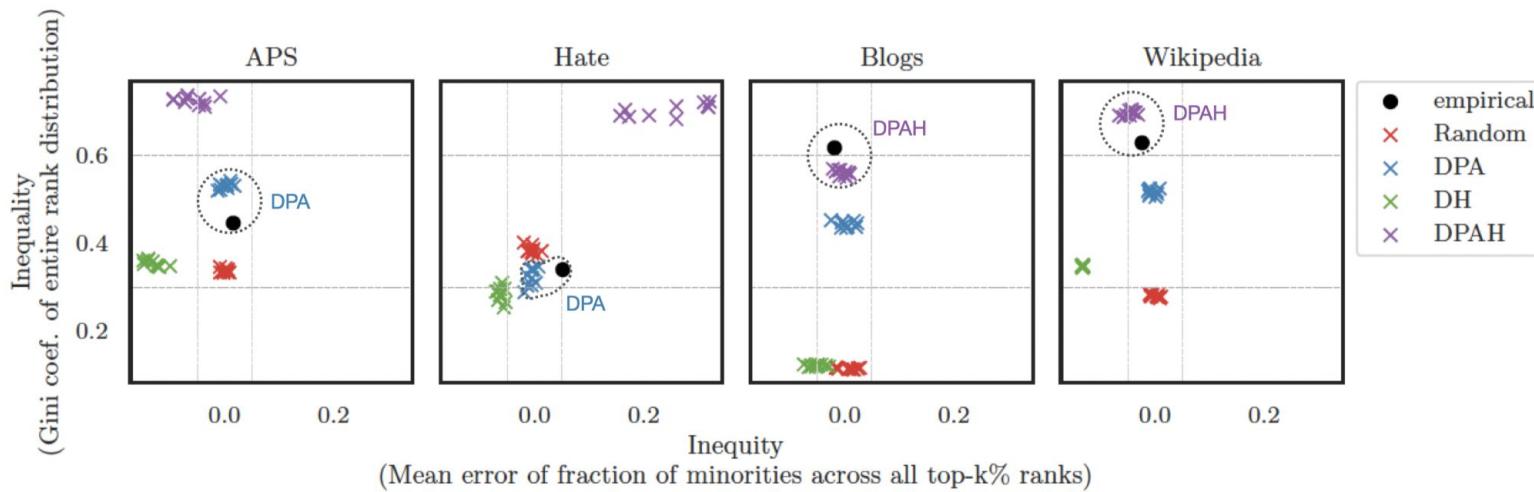


What mechanism of edge formation contributes to ranking inequality and inequity?



Empirical Networks

Model selection (best fit)



Exercise #2

Auditing node rankings

Task:

1. Generate multiple directed graphs.
2. Get to know them (attributes, visualize them, etc.)
3. Plot and compare their type of edges (MM, Mm, mm, mM).
4. Plot and compare the pdf of their degree distributions.
5. Compute and plot their disparity scores (inequality and inequity)

Bonus: Check the effects of homophily, preferential attachment and directed links.

(45 min)

Open `2_exercise.ipynb`

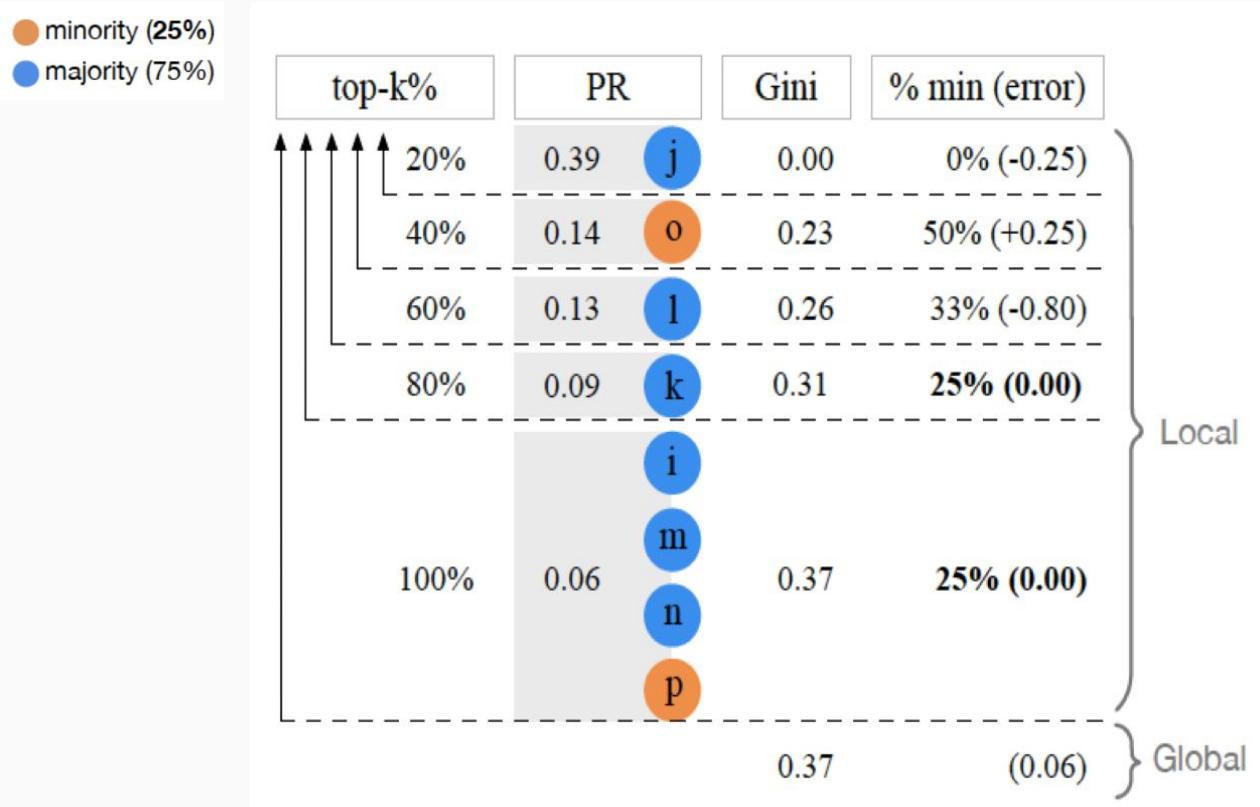
1. Alternatively, you can open the notebook from Google Colab (you need a Google account):

bit.ly/snma2024-notebooks

Coffee break

see you back at 11:20

Extra materials

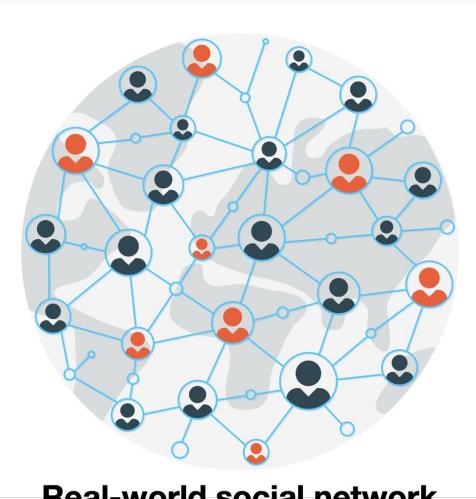


Classification

(not covered today)

Algorithmic bias

Biases in relational classification



Real-world social network

(people with attributes)

What you get

(people with no attributes)

1. Identify network structure

Inference in OSNs via Lightweight Partial Crawls

Konstantin Avrachenkov
INRIA
Sophia Antipolis, France
k.avrachenkov@inria.fr

Bruno Ribeiro
Dept. of Computer Science
Purdue University
West Lafayette, IN, USA
ribeiro@cs.purdue.edu

Jithin K. Sreedharan
INRIA
Sophia Antipolis, France
jithin.sreedharan@inria.fr

Class balance
 $B=0.3$

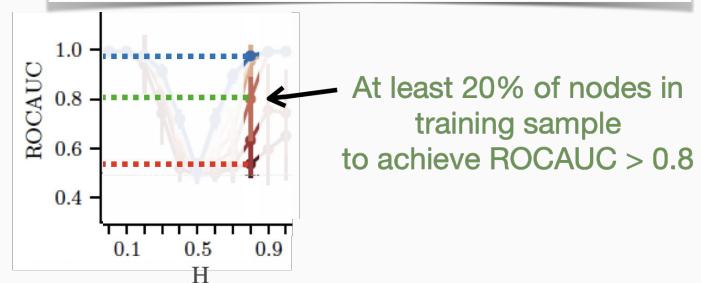
Homophily
 $H=0.8$

2. Identify ROCAUC range for that network

RESEARCH

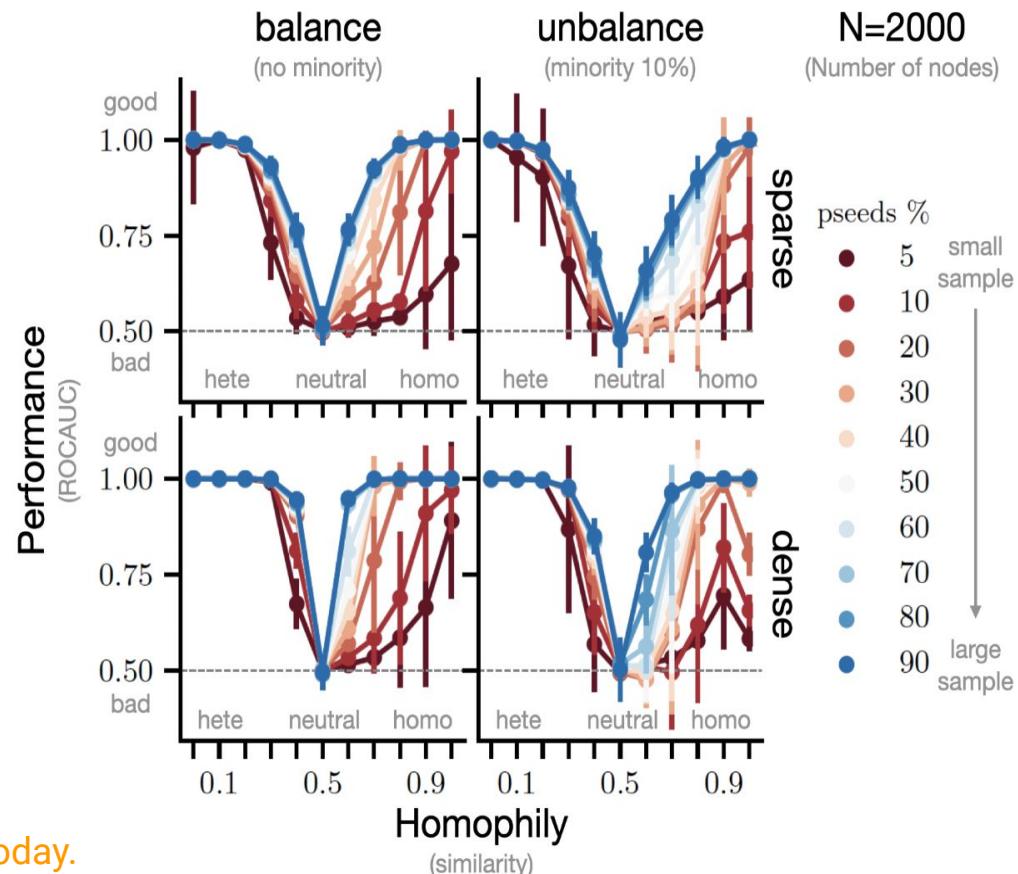
Explaining Classification Performance and Bias via Network Structure and Sampling Technique

Lisette Espin-Noboa, Fariba Karimi, Bruno Ribeiro, Kristina Lerman and Claudia Wagner



Network structure vs. Classification performance

1. Neutral networks ($H=0.5$) cannot be classified better than a random classifier.
2. Homophilic networks ($H>0.5$) achieve lower performance than heterophilic networks when samples are small.
3. Denser networks achieve higher performance compared to sparse networks.
4. Network size mainly affects ROCAUC variance. Larger networks produce more stable results. (not shown here)



Real-world (empirical) networks

Model fitting

