# From social theories to models

Tutors: Lisette Espín-Noboa and Jan Bachmann

ECML PKDD 2024 VILNIUS

bit.ly/snma-2024

# Overview

Time: 09:15 - 11:00

| | |
|---|---|
| 09:15 - 09:30 | Social theories<br>• Popularity<br>• Similarity<br>• Friend-of-friend<br>• Activity |
| 09:30 - 09:40 | Network properties & structure<br>• Graphology<br>• Properties at varying scales<br>• From link formation to structure |
| 09:40 - 10:15 | Network Models<br>• Modeling networks by link formation<br>• *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)

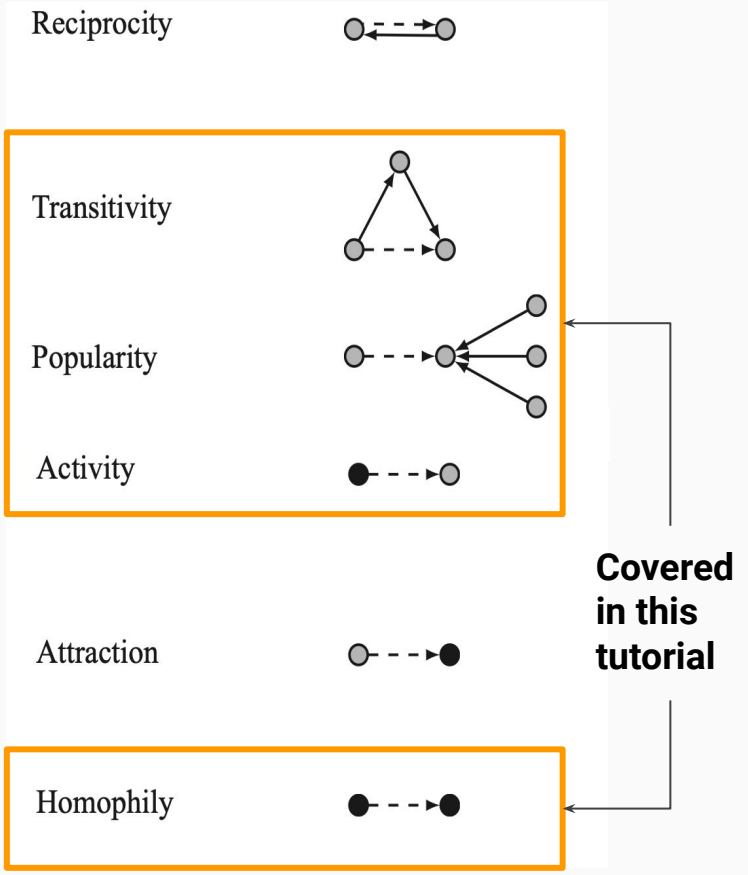
- Actor → person, company, county
  - It may also include entities produced or generated by individuals, e.g., email, retweet.
- Relation → friendship, collaboration, partnership, trade, war, hierarchy
  - When the actor is a resource, the relation may represent a spreading dynamic, access to opportunities,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

# How do networks form?

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

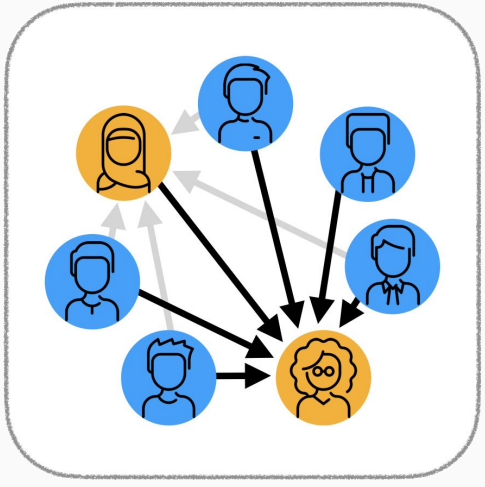
Prevalent **mechanisms** identified in social science literatures:

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




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

# Popularity (preferential attachment)

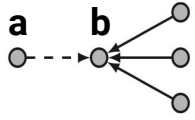


Popularity

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.

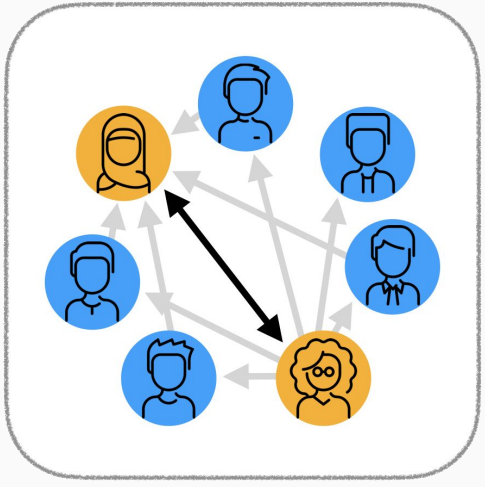
- **Matthew effect or rich-gets-richer effect** mechanism in science discussed by Merton (1968), Price (1976) and others. Increasing recognition of an actor's scientific work (e.g. number of ties in a citation network).
- **Preferential attachment** in the work of Barabási and Albert (1999) operationalised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




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

# Homophily (similarity)

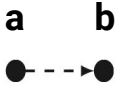


Homophily

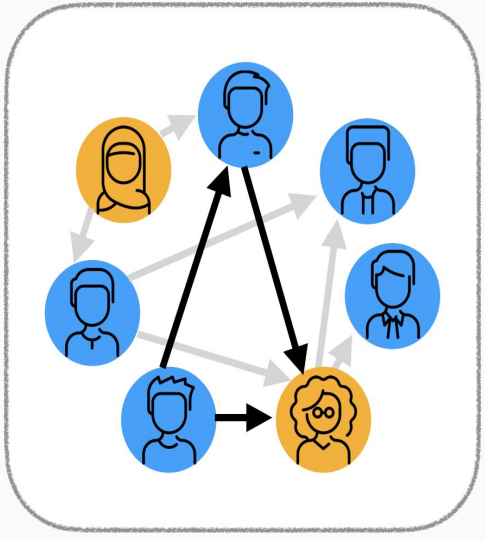
The tendency that similar actors are more likely to connect (Lazarsfeld, et al. 1954; McPherson et al. 2001). Its structural position is defined by **the attribute similarity of the sender and receiver** of the explained tie.

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.




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



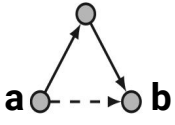
Transitivity

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.

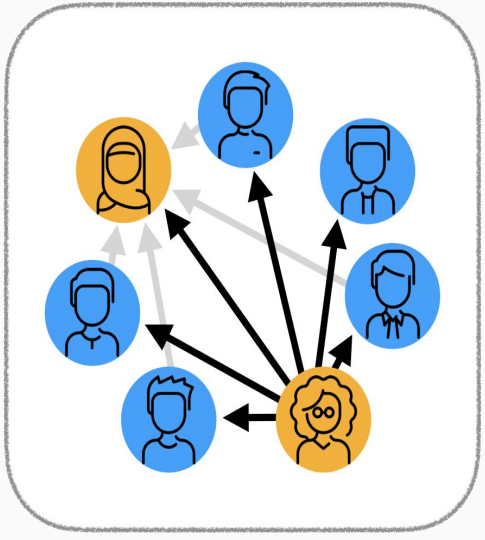
Causal explanations:

- 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).






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



Node Activity

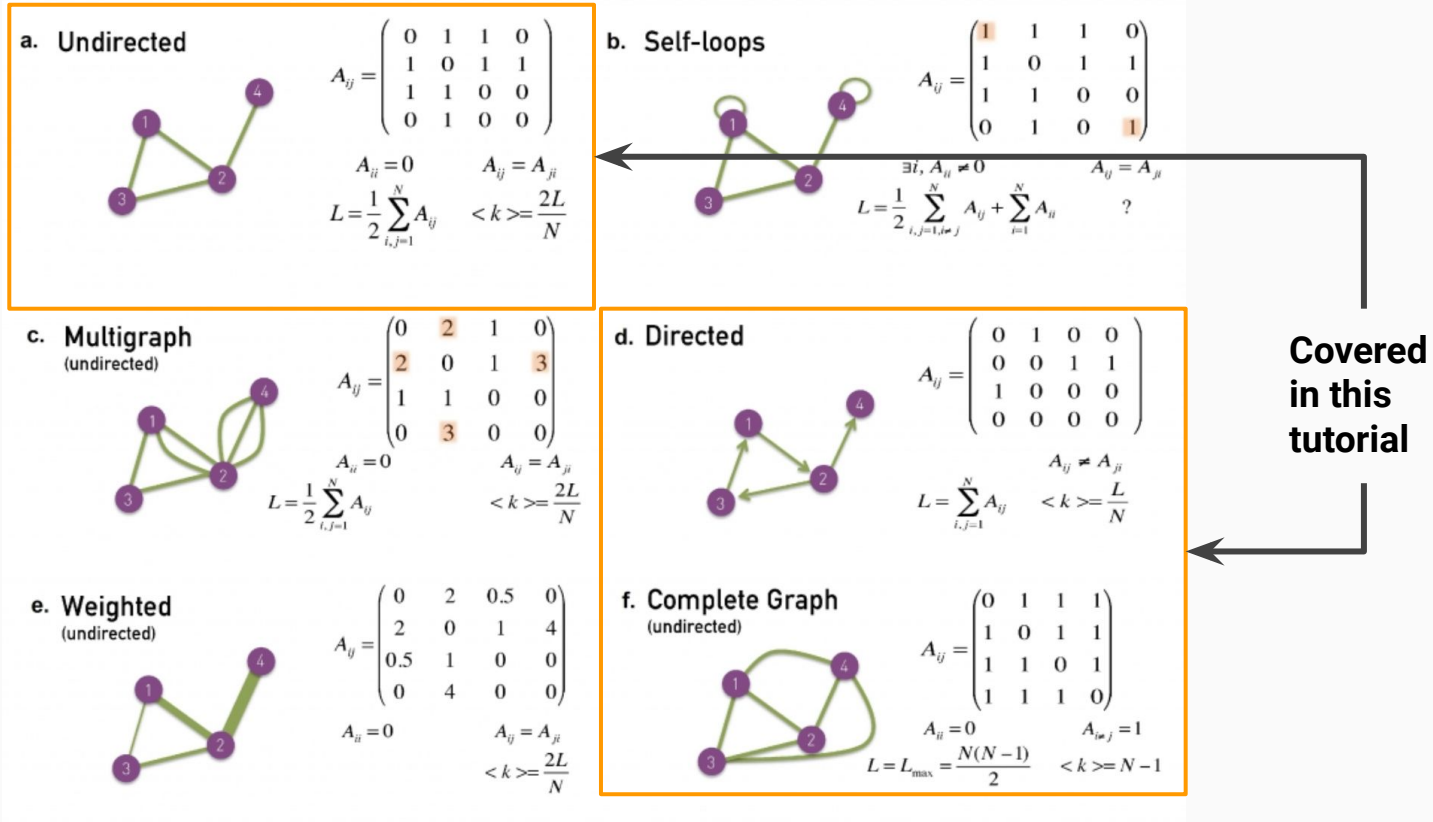
The tendency that actors with specific attributes will be more likely to send ties. The structural positions are defined by **attributes of the sender** of the explained tie. Whether these differences relate to differences in networking resources is not further specified by the mechanism.

a ●- - -▸○ b

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

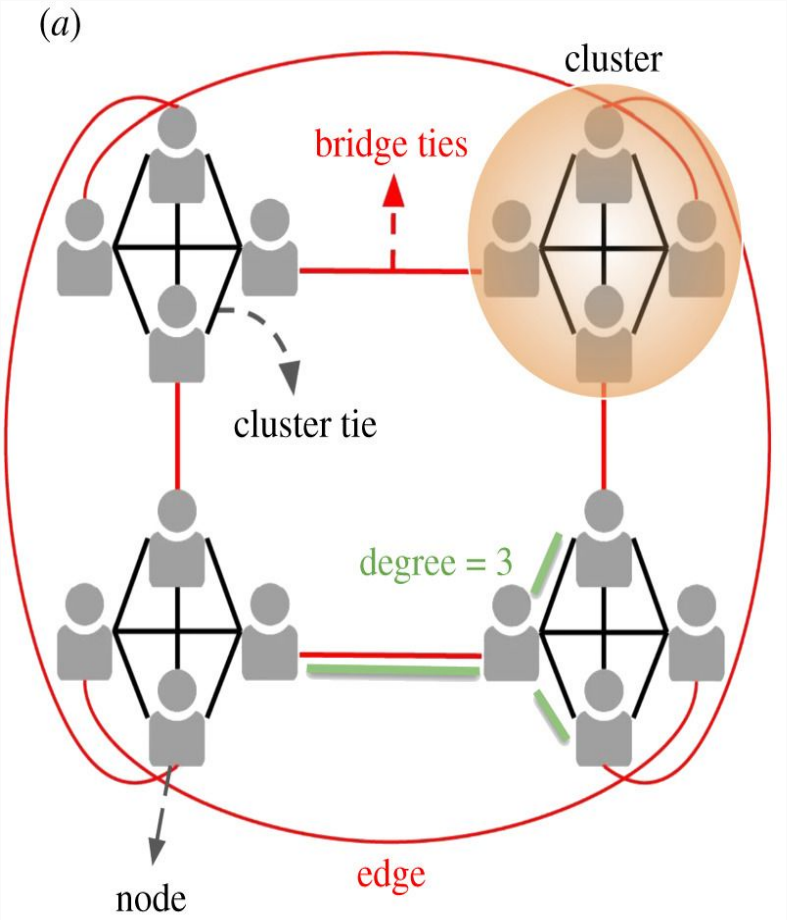
# Network properties and structure

# Graphology (elementary property of the underlying graph)

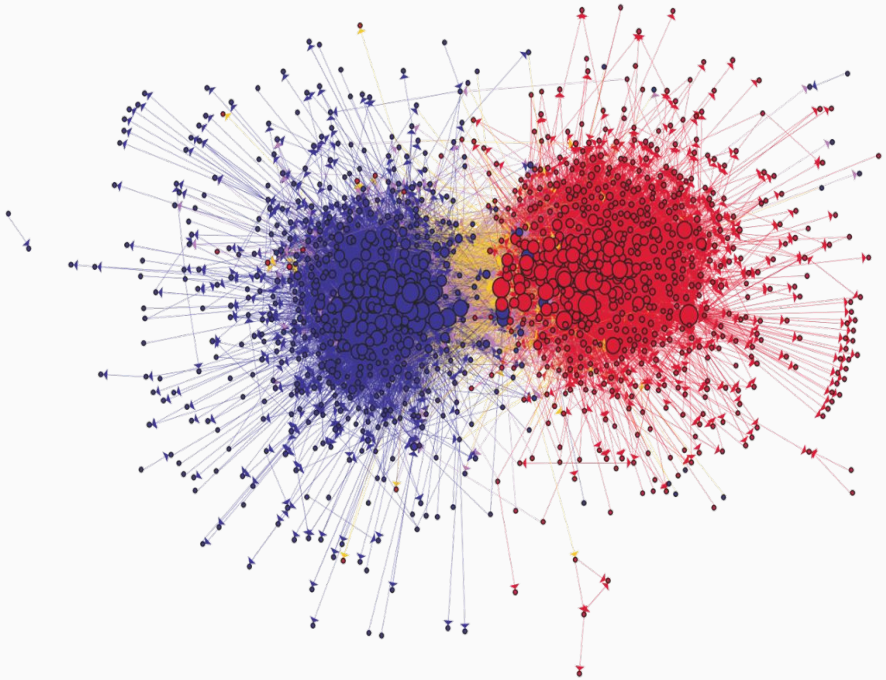



http://networksciencebook.com/chapter/2#summary2

# Common structure (topology) of social networks




https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0315



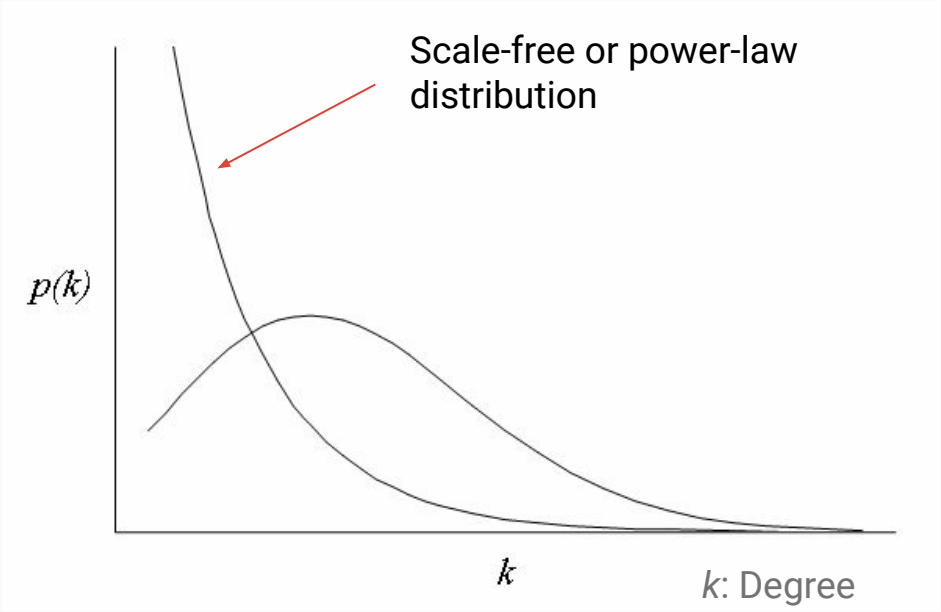
Adamic et al., “The political blogosphere and the 2004 U.S. election: divided they blog” LinkKDD ‘05, 36 – 43

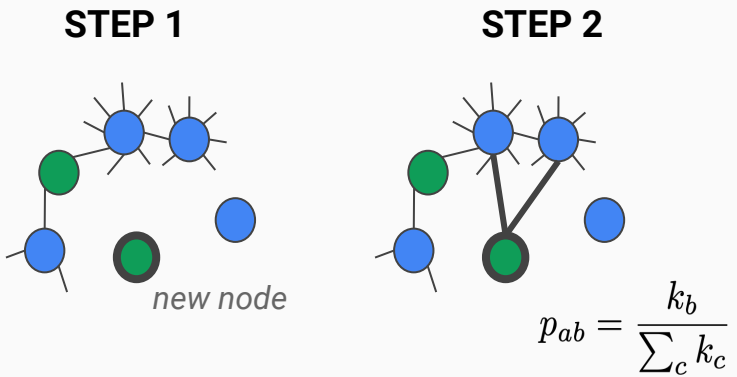
# Important properties of social networks at varying scales

| Graph | Node | Edge |
| --- | --- | --- |
| Community structure | Degree (in/out) | Weights |
| Scale-free degree distributions<br>(the probability that a randomly selected node in the network has degree k) | Other centrality (e.g,. PageRank) | Shortest path |
| Segregation | Clustering coefficient | Homophilous type<br>(MM, Mm, mm, mM) |
| Sparsity<br>(out of the total possible number of edges, how many actually exist) | Activity<br>(high activity = high outdegree) | Betweenness |
| Small world property | | |
| Size of largest connected component | | |

**How do these properties emerge from micro-scale interactions?**

# 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?

**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<br>Watts-Strogatz | • N fixed<br>• pk exponentially bounded<br>• Static, time independent topologies |
| Generative Models | Configuration Model<br>Hidden Parameter Model<br>Stochastic Block Model | • Arbitrary pre-defined $p_k$<br>• Static, time independent topologies |
| Evolving Network Models | Barabási–Albert Model<br>Bianconi-Barabási Model<br>Initial Attractiveness Model<br>Internal Links Model<br>Node Deletion Model<br>Accelerated Growth Model<br>Aging Model | • $p_k$ is determined by the processes that contribute to the network's evolution.<br>• 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.

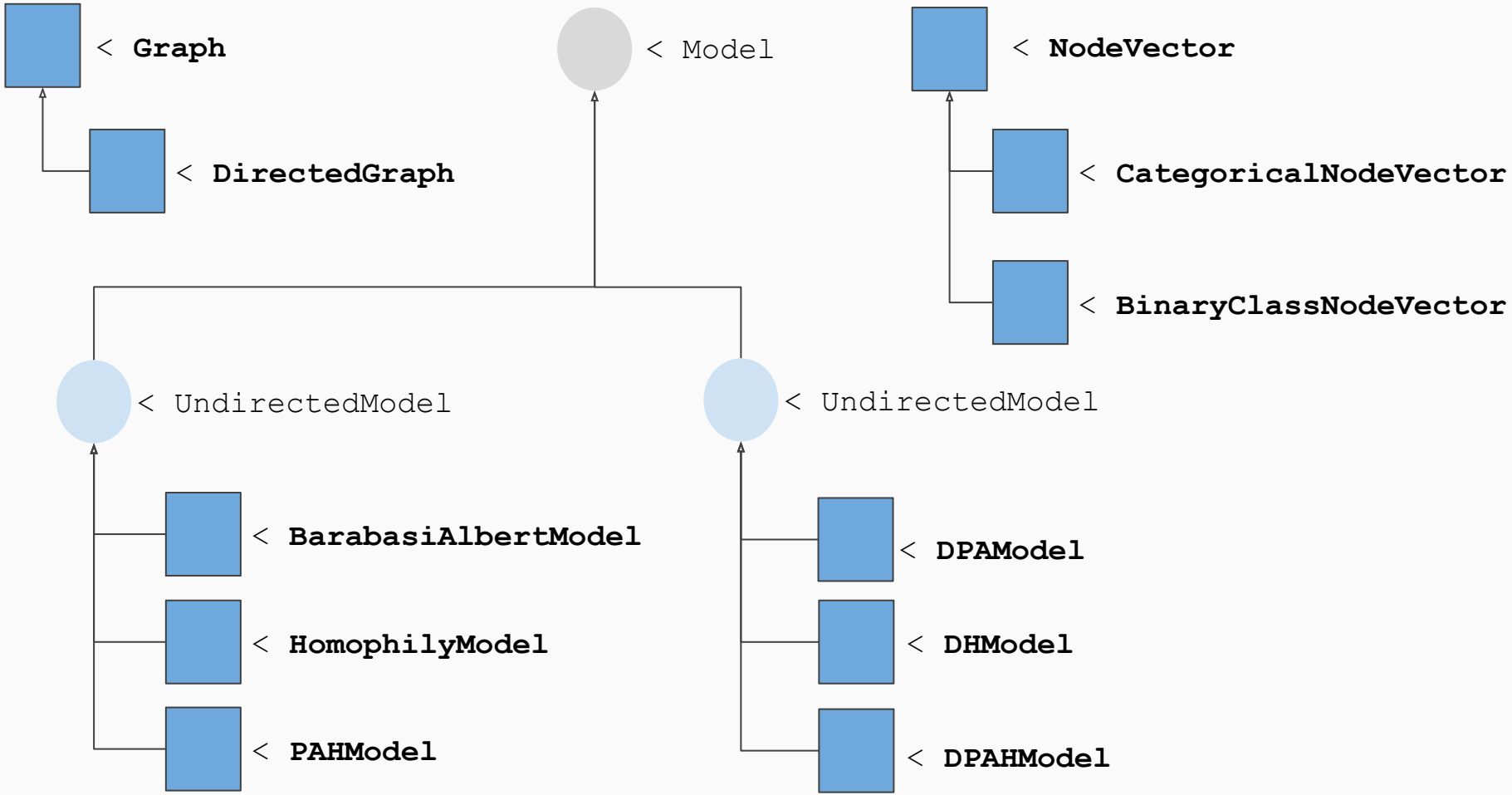
# Pre-requisites

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
   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)

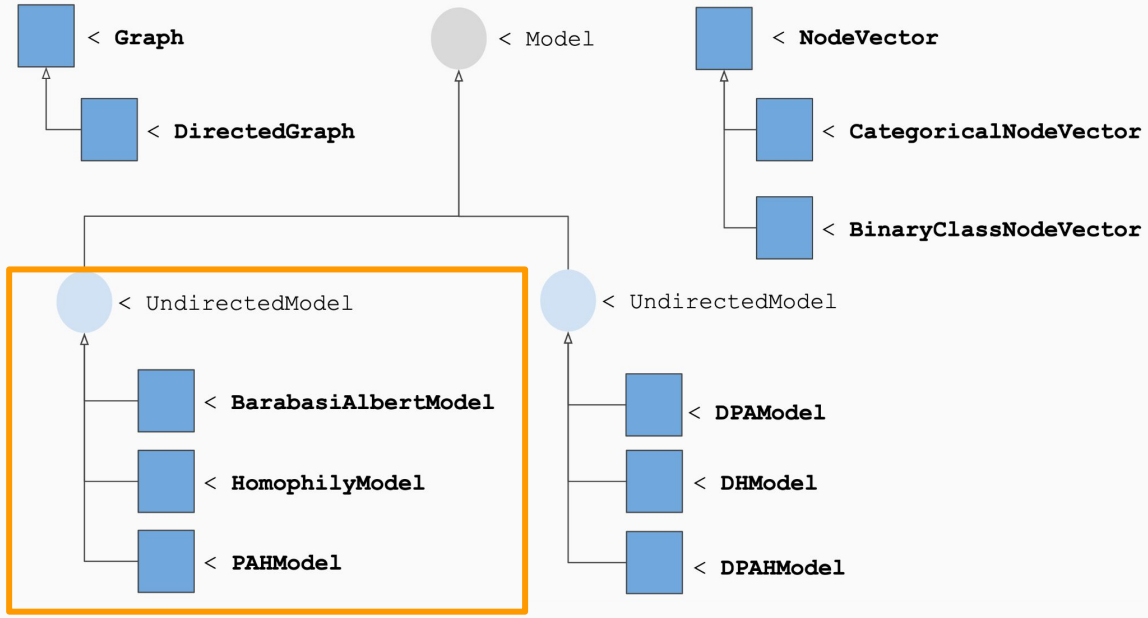


# Undirected networks

**BarabasiAlbertModel**
preferential attachment

**PAHModel**
PA + homophily

**PATCHModel**
PAH + triadic closure



## *class* **BarabasiAlbertModel**(...)

Parameters

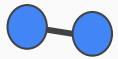
`N (int)`

- number of nodes to be added to the network.

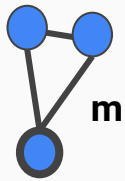
`m (int)`

- number links per new node

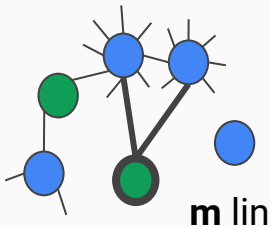
**STEP 1:**
m = 2



**STEP 2:**




**STEP 3 (until N):**




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

A. L. Barabási & R. Albert “Emergence of scaling in random networks” Science 1999.

## *class* **PAHModel**(*N, m, f_m, h_m, h_M*)

Parameters

`f_m (float)`

- Fraction of minority nodes in the network

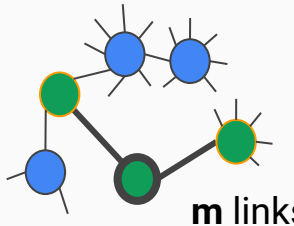
`h_m (float)`

- Homophily of the minority group

`h_M (float)`

- Homophily of the majority group

**STEP 3 (until N):**



**m** links

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

## *class* PATCHModel(...)

Parameters

`p_tc (float)`

- Probability to perform a triadic closure link

`lfm_local ("Uniform", "Homophily", "PAH")`

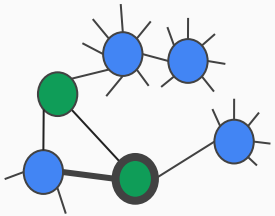
- How to choose target nodes locally

`lfm_global ("Uniform", "Homophily", "PAH")`

- How to choose target nodes globally

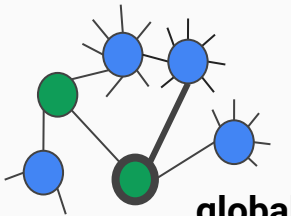
**STEP 3 (until N):**
**m** times

Case 1: $p_{TC}$



**local** link (restricted to friends of friends)

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



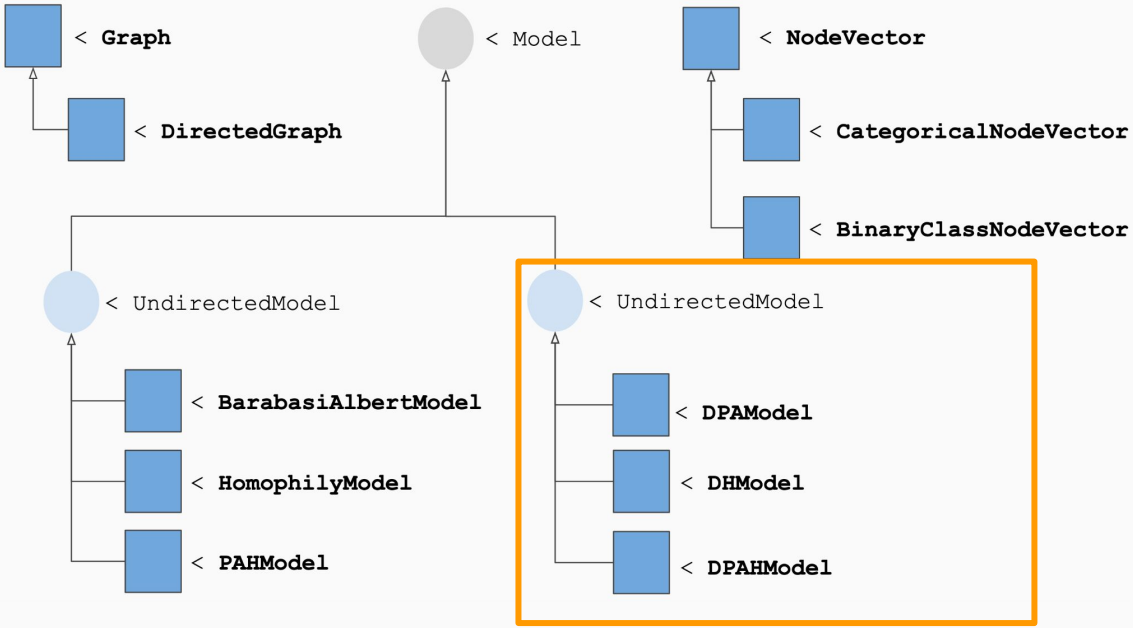
**global** link (any node available)

# 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)

L. Espín-Noboa, C. Wagner, M. Strohmaier, & F. Karimi “Inequality and inequity in network-based ranking and recommendation algorithms” Scientific reports 2022

## *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)

L. Espín-Noboa, C. Wagner, M. Strohmaier, & F. Karimi “Inequality and inequity in network-based ranking and recommendation algorithms” Scientific reports 2022

# 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

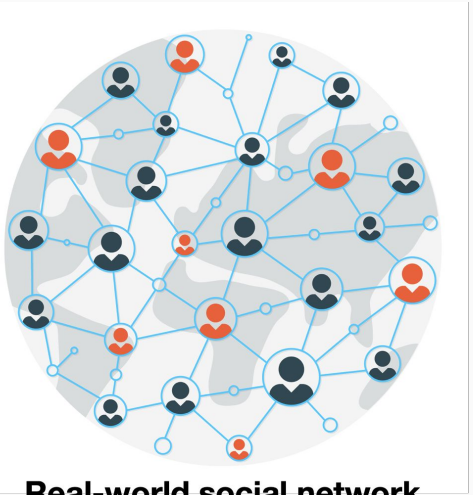
Repeat for other graphs and check the effects of other mechanisms.

(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



**Real-world social network**
(people with attributes)

**Ranking / RecSys.**
(PageRank or WTF)

**1. Identify network structure**

**Fraction min.**
**fm=0.3**

**Node activity**
**γM = γm = 3**

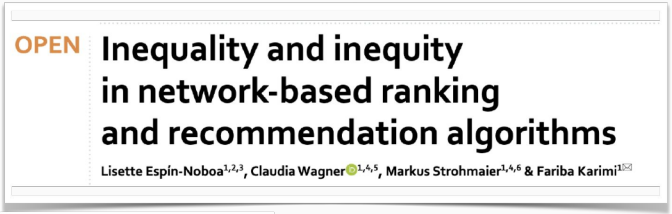
**Density**
**d=0.0015**

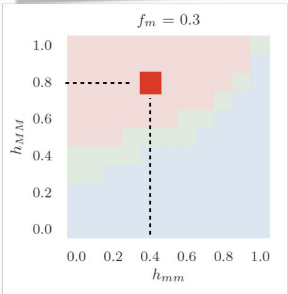
**Homophily Maj.**
**$H_{MM}$=0.8**

**Homophily min.**
**$H_{mm}$=0.4**

(Inequity is driven by homophily and fraction of minorities)

**2. Identify inequality and inequity in ranking**

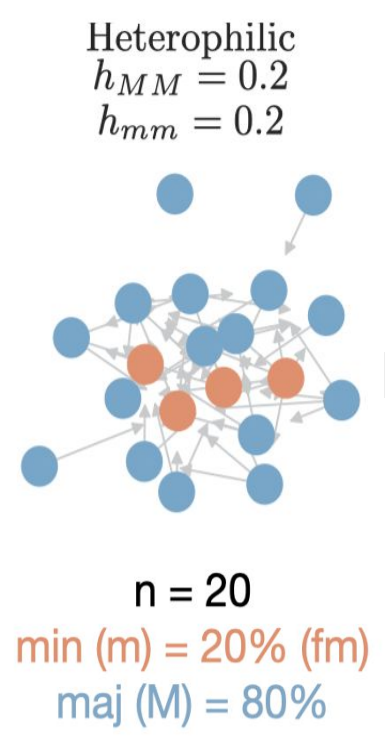


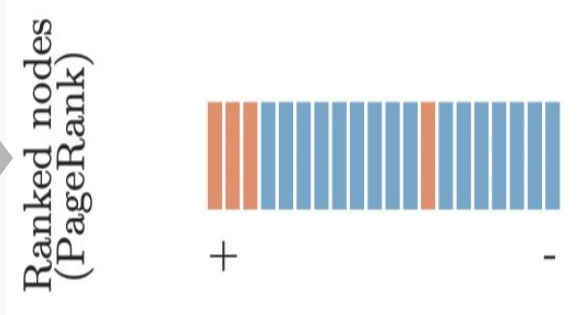

On average minorities are under-represented in top-k's
(Interventions needed)

Espín-Noboa, L., Wagner, C., Strohmaier, M., & Karimi, F. (2022). Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific reports*, *12*(1), 1-14.

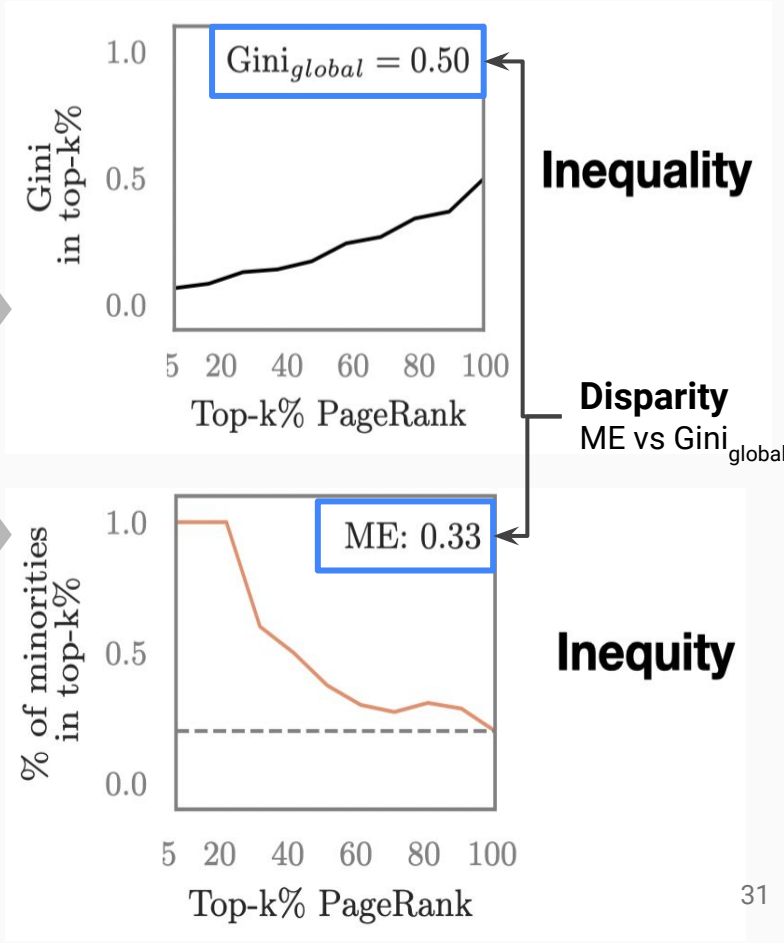
**Given a network,**




**… and a ranking of its nodes**



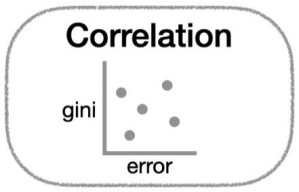

**Measure:**

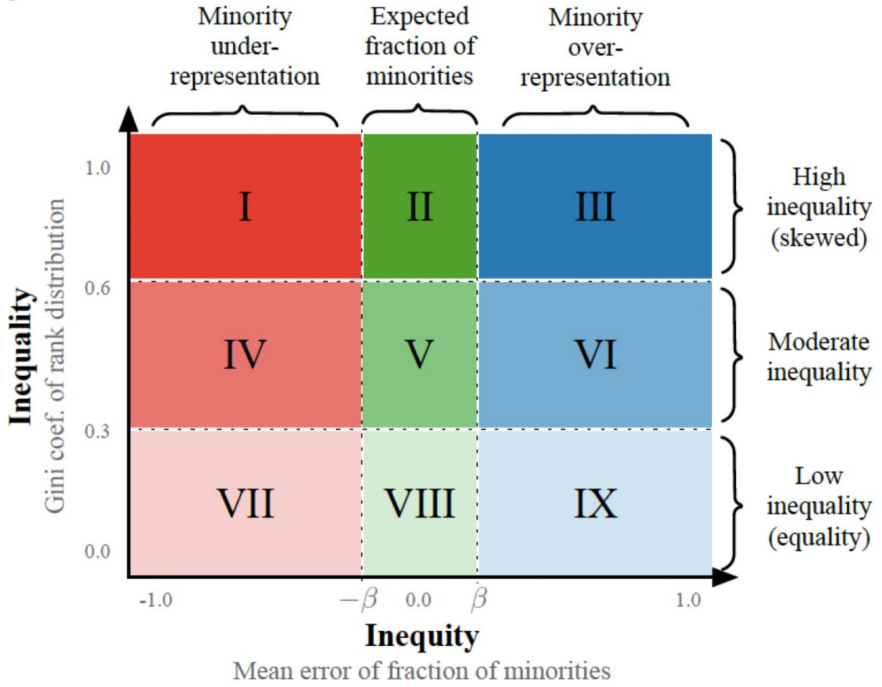



Espín-Noboa, L., Wagner, C., Strohmaier, M., & Karimi, F. (2022). Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific reports*, *12*(1), 1-14.

# Inequality vs. Inequity

## Regions of disparity

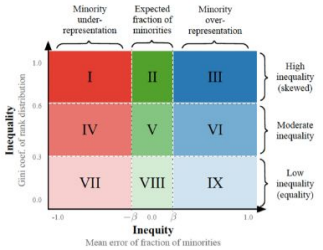



Espín-Noboa et al. 2022

# Inequality vs. Inequity in PageRank

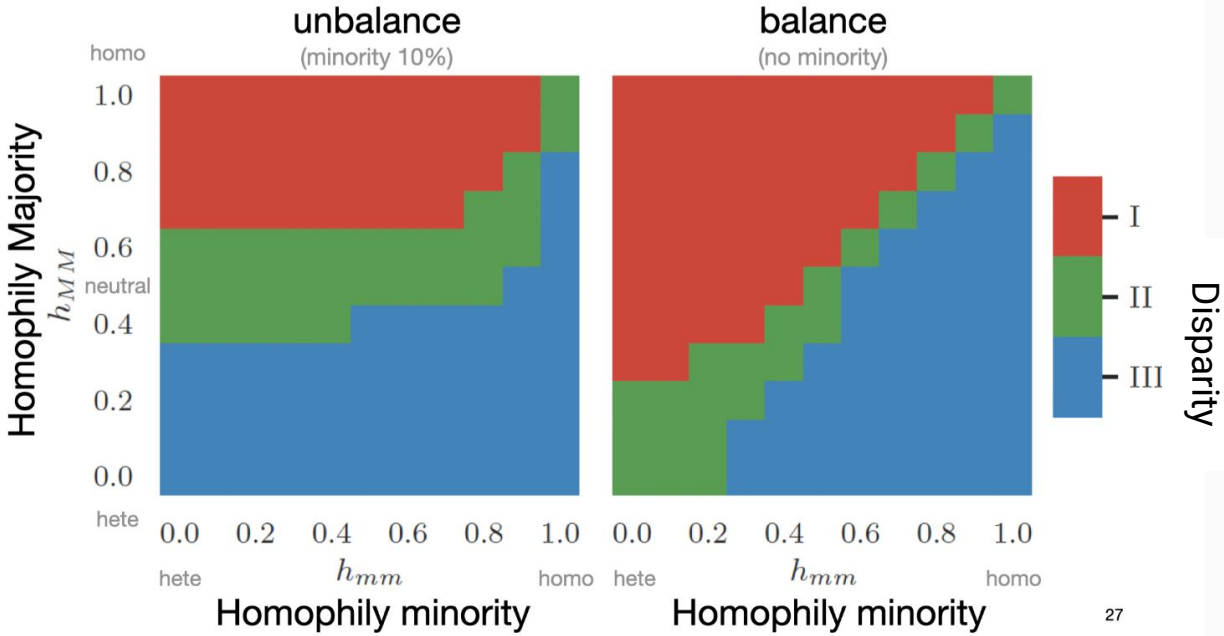
## As a function of **H**omophily and **F**raction of **m**inorities




1. In **balanced networks**, both groups are well represented if hmm = hMM

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

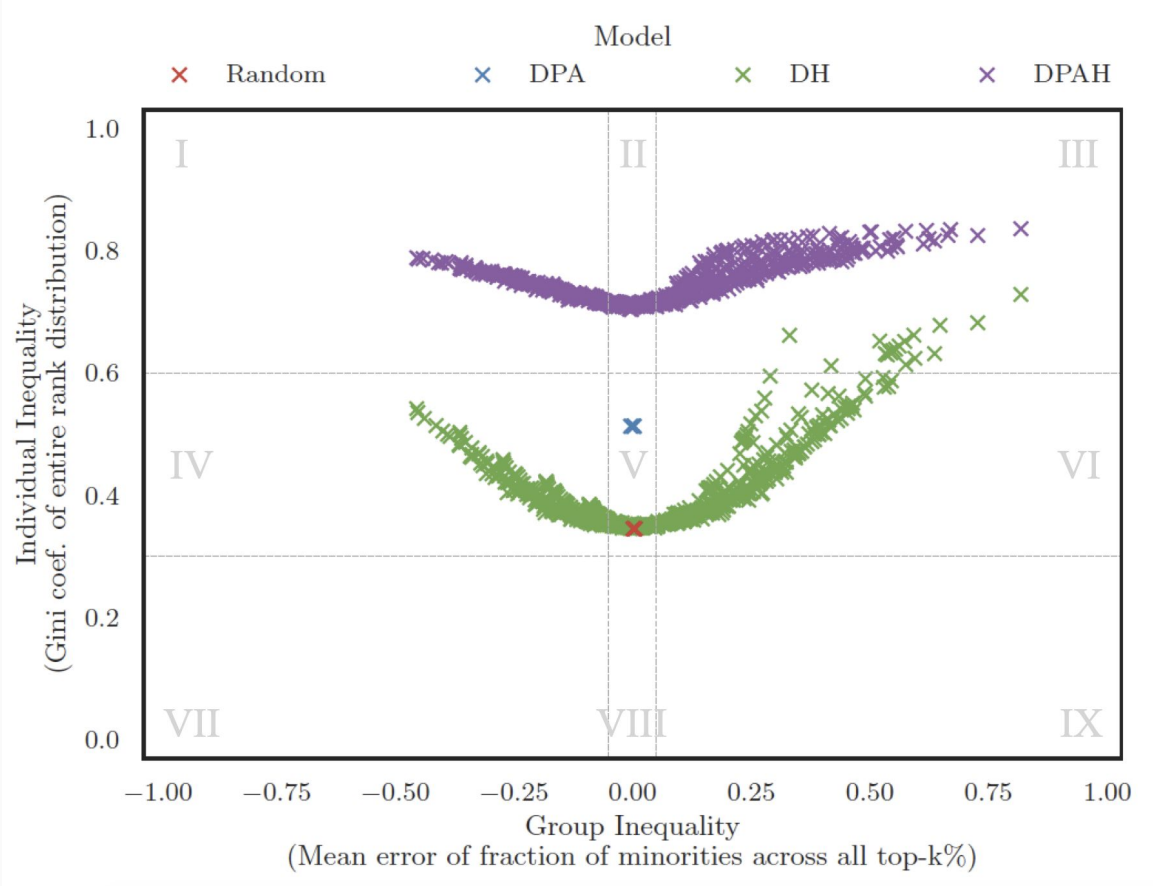
3. In **unbalanced and homophilic networks**, minorities are well represented when hmm > hMM.




Espín-Noboa et al. 2022

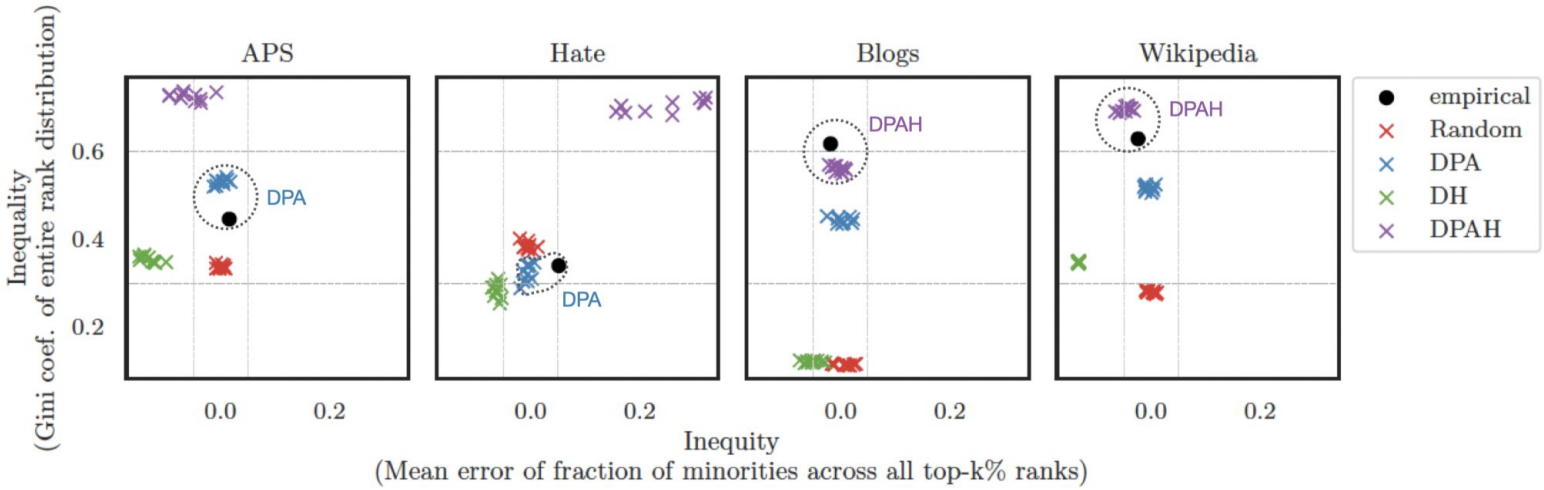
27

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



# Empirical Networks

## Model selection (best fit)




Espín-Noboa, L., Wagner, C., Strohmaier, M., & Karimi, F. (2022). Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific reports*, *12*(1), 1-14.

# 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)

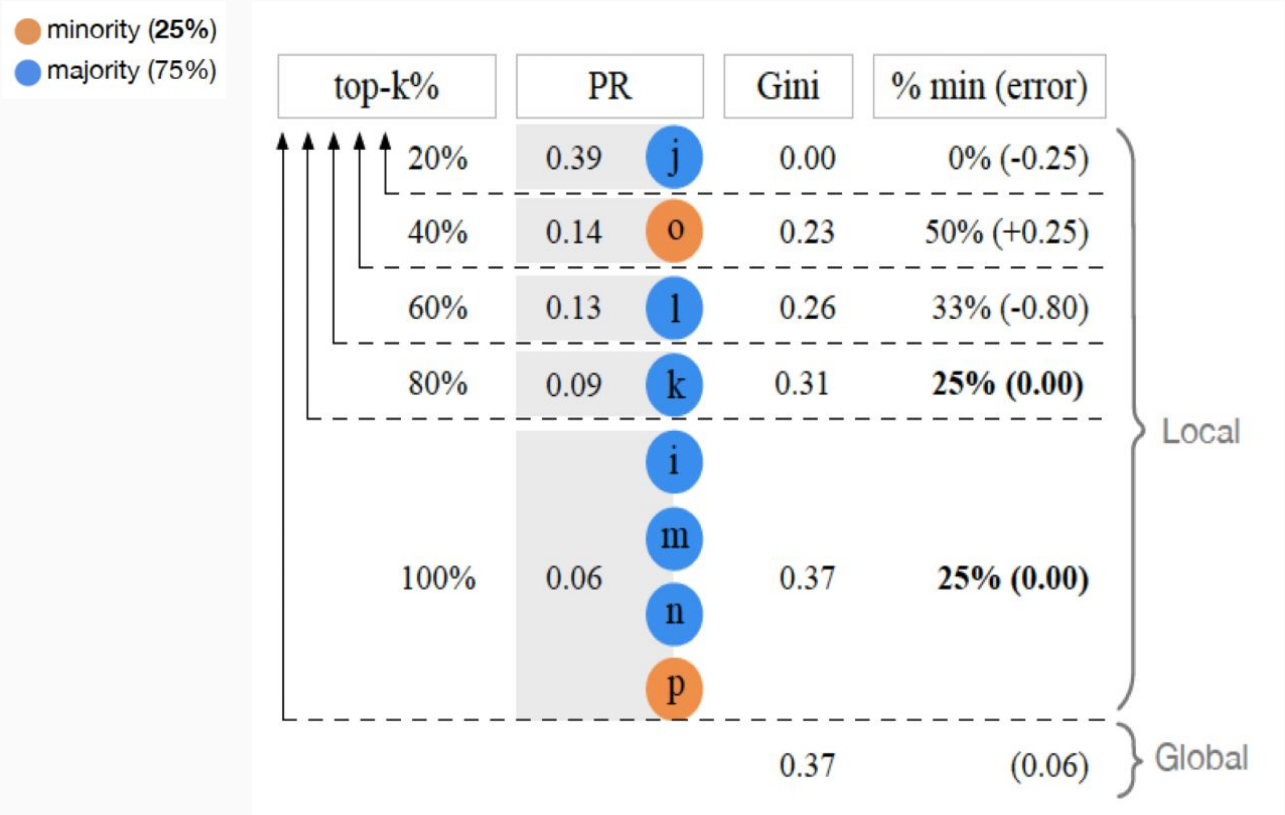
Open **`2_exercise.ipynb`**

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

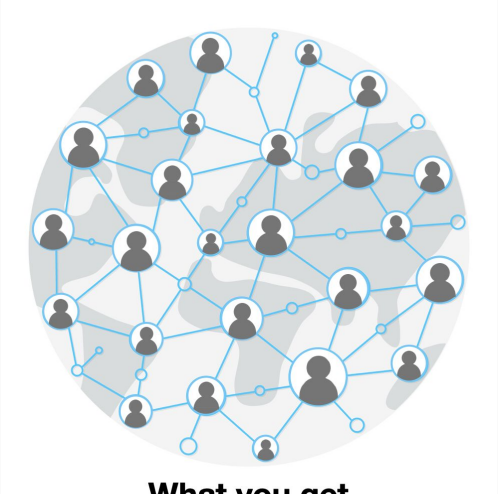



Espín-Noboa, L., Wagner, C., Strohmaier, M., & Karimi, F. (2022). Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific reports*, *12*(1), 1-14.

# Classification

(not covered today)

**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

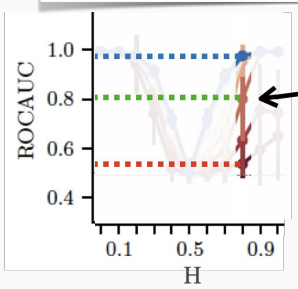
Homophily
H=0.8

**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




At least 20% of nodes in training sample to achieve ROCAUC > 0.8

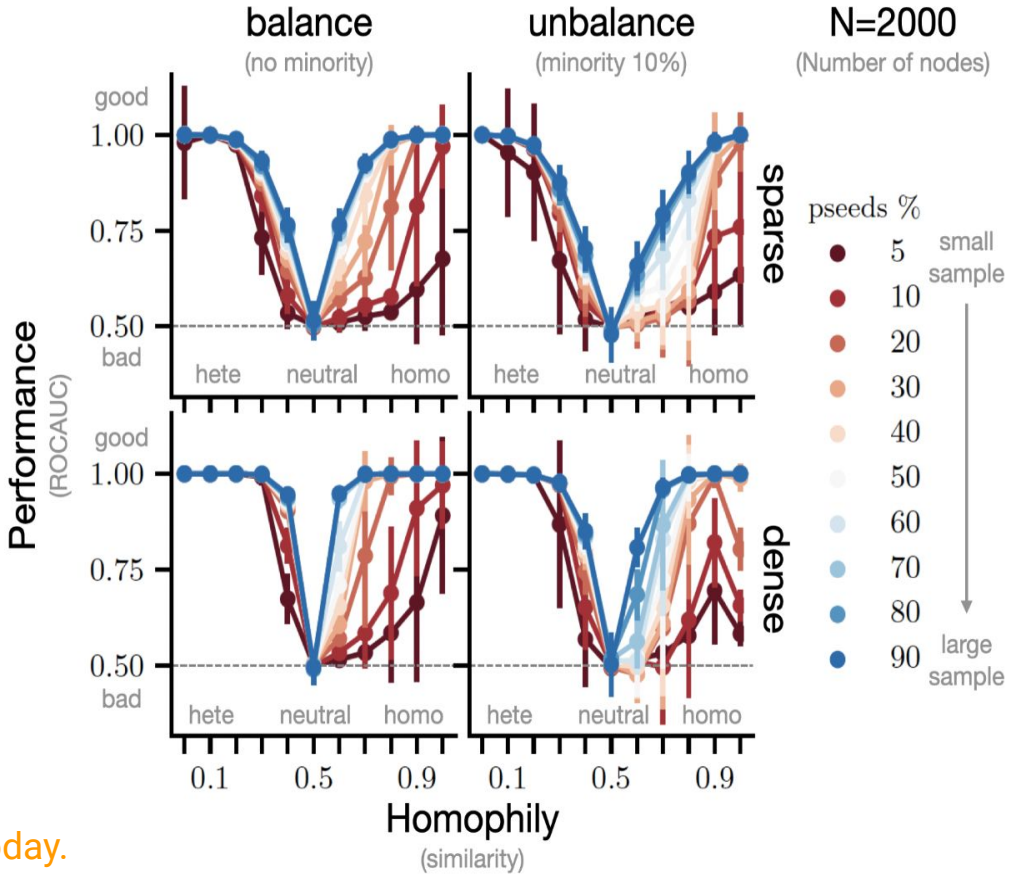
# Network structure vs. Classification performance

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)



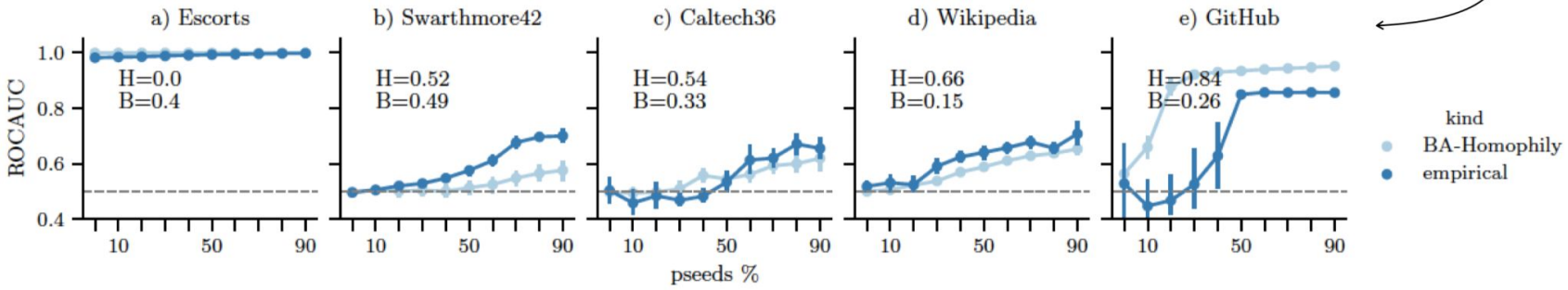

Espín-Noboa et al. 2021

Not covered today.

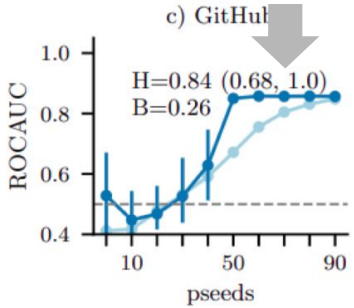
# Real-world (empirical) networks

## Model fitting

With symmetric homophily
$H_{mm} = H_{MM}$




With asymmetric homophily
**A more realistic representation**
$H_{mm} \neq H_{MM}$




76