uc3m | Universidad Carlos III de Madrid

Network Theory for the Social Sciences in Python

Methods Workshop: Social Sciences PhD Program (2025/2026)

4. Link prediction and network inference

https://github.com/blas-ko/uc3m_networks_workshop_2025

Blas Kolic
blas.kolic@uc3m.es

Outline

1. Introduction to the problem of link prediction
2. Graph heuristic predictors
3. Example
4. Other approaches
5. References

Shameless self-promotion

My band is playing a show in Malasaña **tonight**!

1 Introduction

Introduction

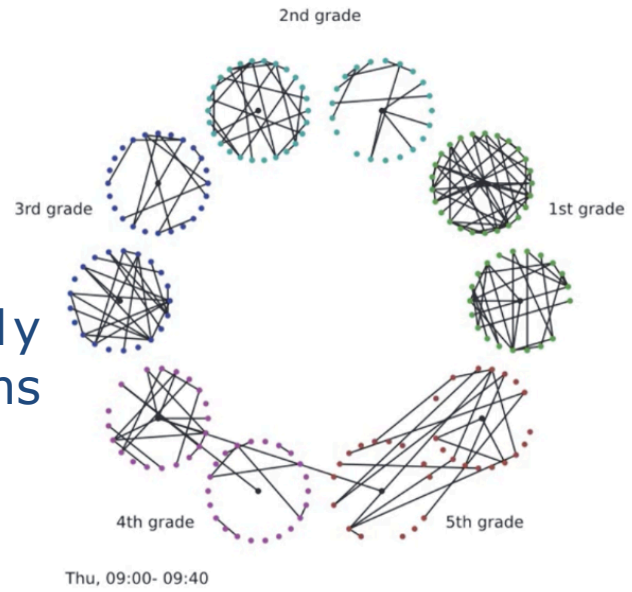
Networks are constantly changing, but many patterns prevail.

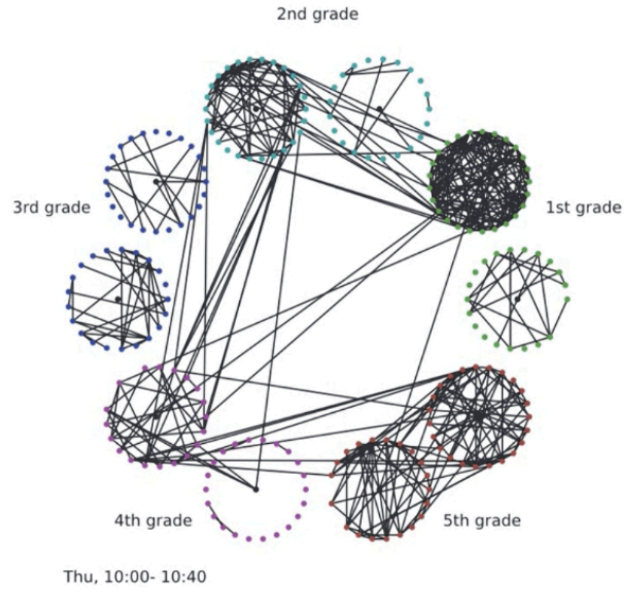
Introduction

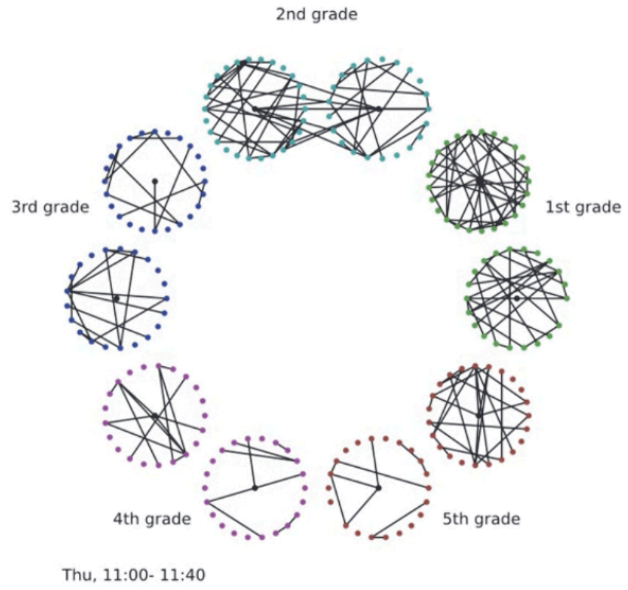
Networks are constantly changing, but many patterns prevail.

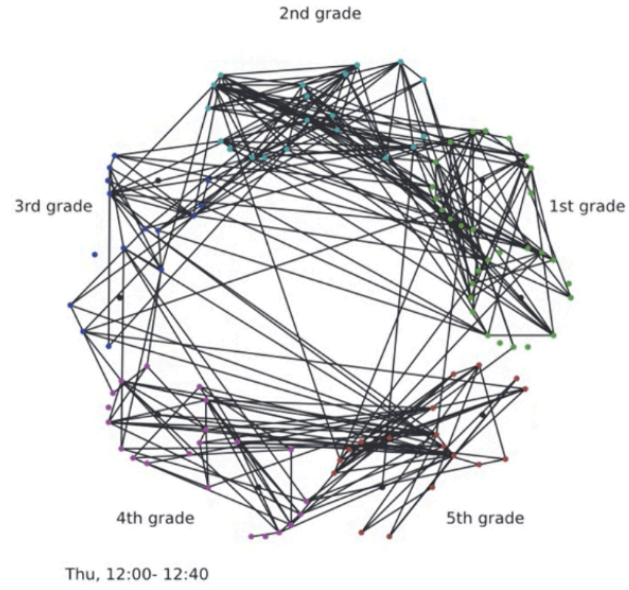
Example: Primary school face-to-face interactions

- Within classroom interactions during class
- Spread-out interactions outside class
- Patterns may carry on into next days









uc3m

Masuda, N., & Lambiotte, R. (2016). A guide to temporal networks.

Introduction

Rivera et al. (2010) discuss the **mechanisms** behind the dynamics of interactions in social networks:

1. **Assortative mechanisms**
    1. *Homophily* between individuals promotes attachment.
    2. *Heterophily* sometimes creates diversity.
2. **Relational mechanisms**
    1. *Triadic closure:* common contacts are exposed to interact.
    2. *Reciprocity*: if you receive a stimulus, you're likely to respond.
    3. *Repetition*: Repeated interactions make bonds strong.
    4. *Degree*: Connectivity attracts new connections.
3. **Proximity**
    1. *Geographical distance.* Likelihood of new connections and persistence of old ones.

Rivera, M. T., Soderstrom, S. B., & Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. annual Review of Sociology, 36(1), 91-115.

Introduction

These mechanisms explain:

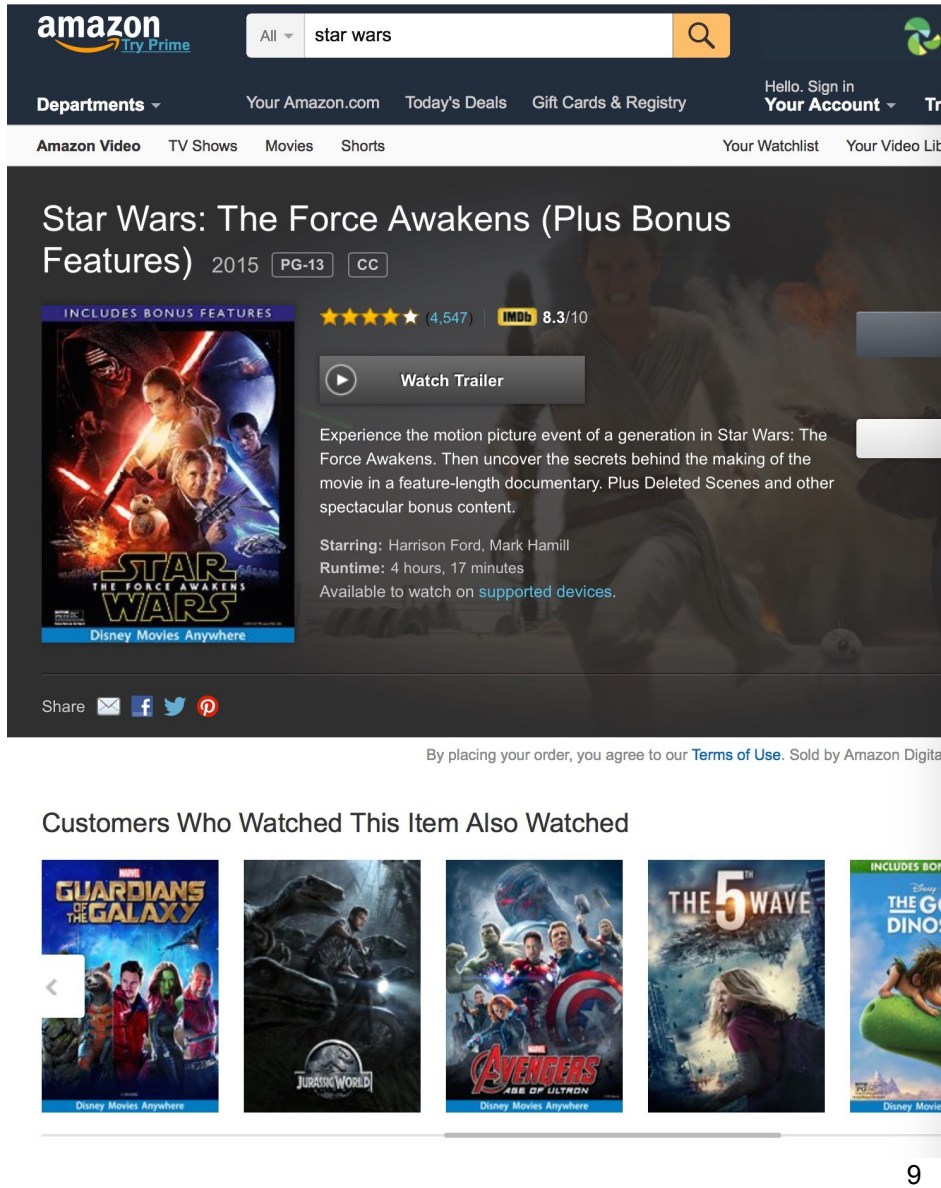
1. Why new connections form
2. Why old connections prevail
3. Which connections are likely happening even when we don't measure them

Today, we will quantify some of these mechanisms in order to **predict future or missing links** based on the connections we are able to observe.

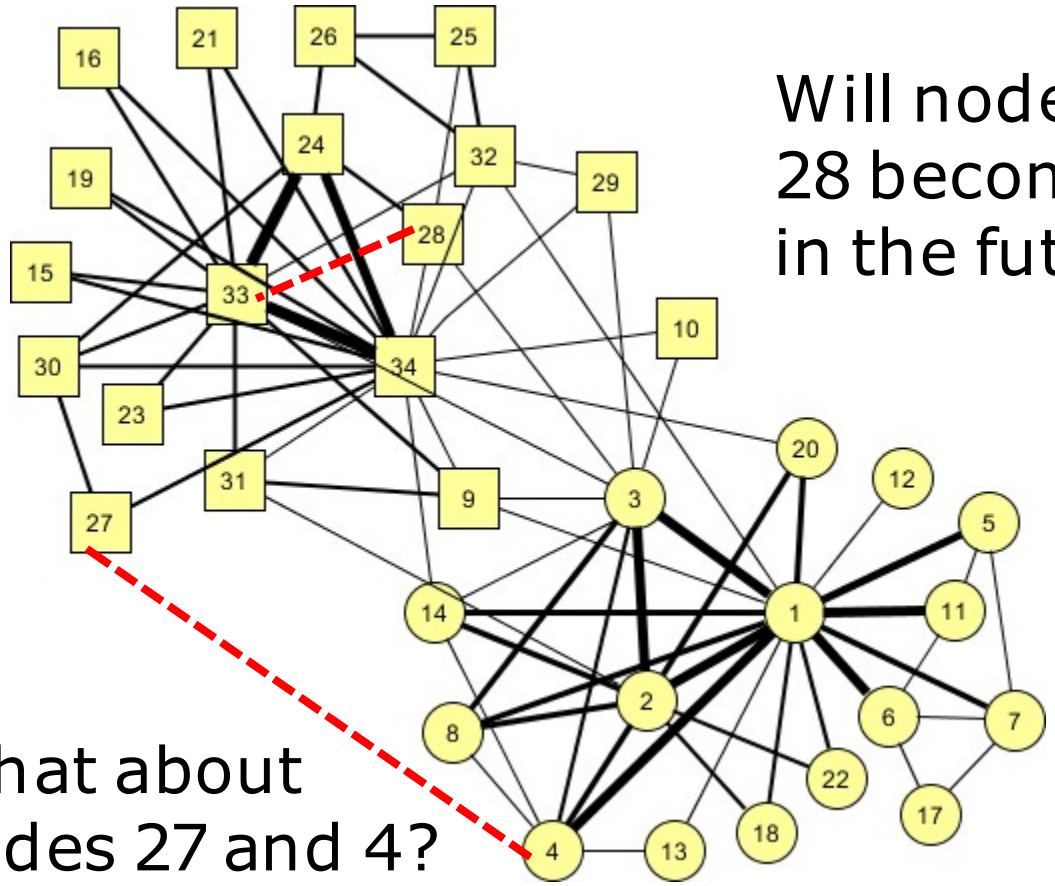
Rivera, M. T., Soderstrom, S. B., & Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. annual Review of Sociology, 36(1), 91-115.

Introduction

- Common applications:
  - Friend recommendation in social networks
  - Recommendation of movies/ products (Netflix, Amazon) "people who bought this also bought…"
  - Potential collaboration co-authorship networks
  - New protein interactions
  - Detect hidden groups of criminals



Introduction



Will nodes 33 and 28 become friends in the future?

Does network structure contain enough information to predict what new links will form in the future?

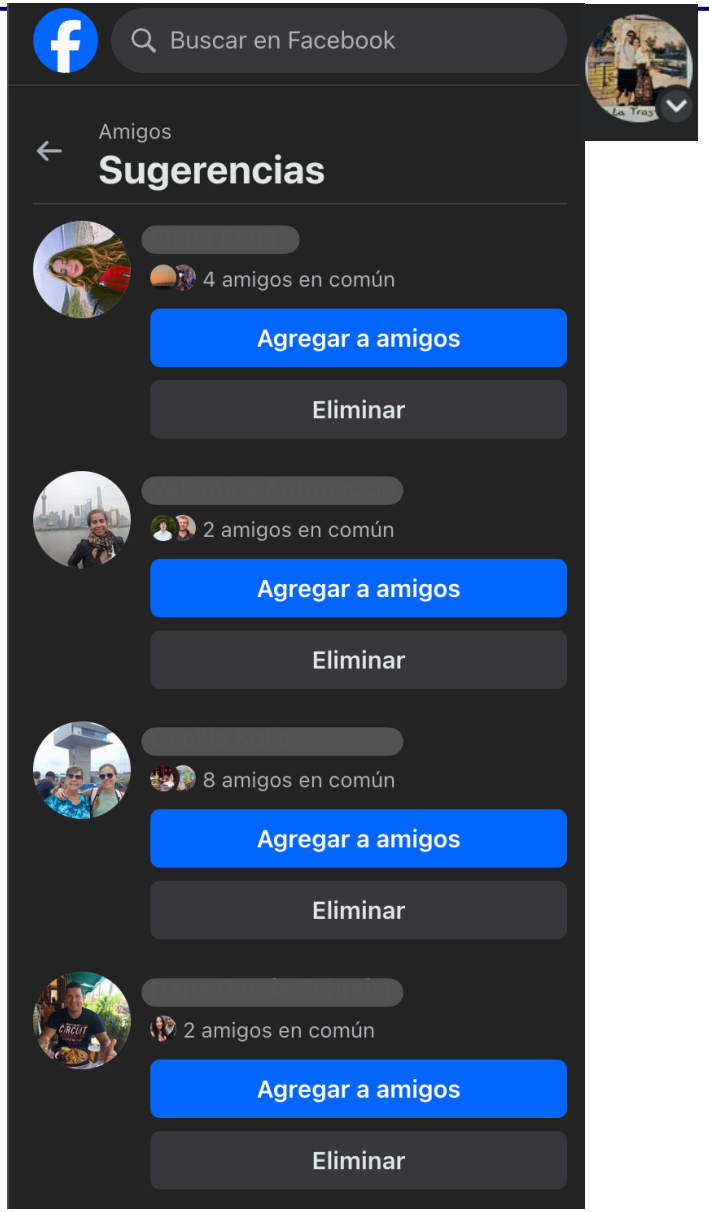
What about nodes 27 and 4?

uc3m

Introduction

In social networks:

- Can we predict future connections based on the network so far?
- The basic idea is to define the **"similarity" of node pairs in a network:**
  - Nodes which have large similarity are likely to be connected.



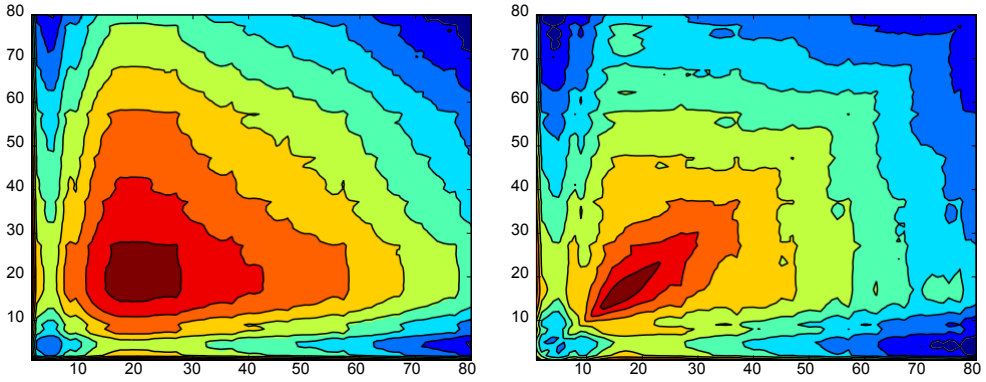
Mechanisms

Homophily: The greater the similarity between individuals, the more likely they are to establish a connection

Attribute	Random	Communicate
Age	-0.0001	0.297
Gender	0.0001	-0.032
ZIP	-0.0003	0.557
County	0.0005	0.704
Language	-0.0001	0.694

Correlation coefficient

Number of pairs of people at different ages



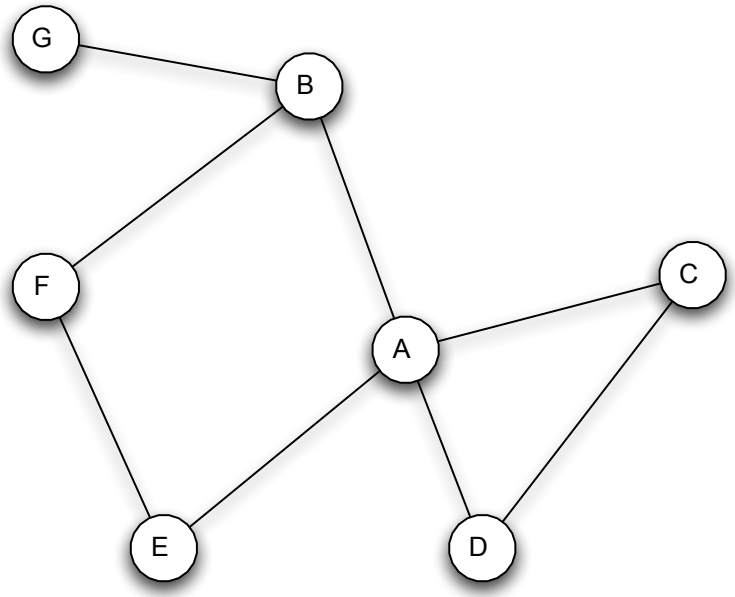
(a) Random (b) Communicate

Leskovec, J. & Horvitz, E., 2008. Planetary-scale views on a large instant-messaging network. pp.915–924.

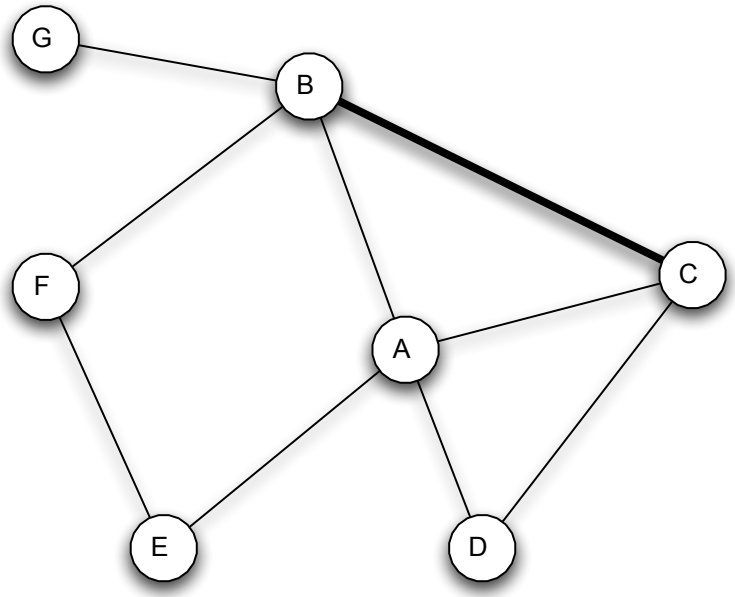
Proximity =similarity in age, language, geolocation, etc.

Mechanisms

Triadic closure: If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future



(a) *Before* B-C *edge forms.*

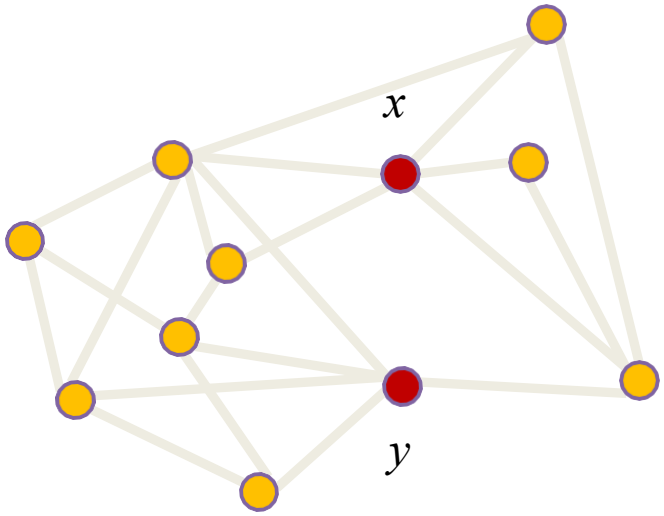


(b) *After* B-C *edge forms.*

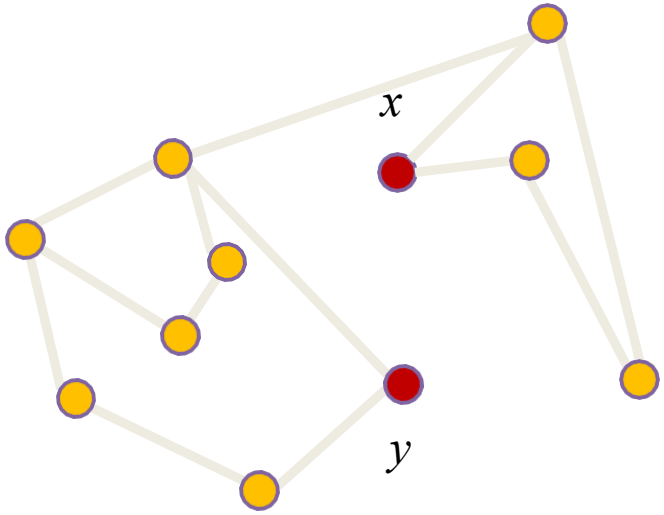
uc3m

Mechanisms

Embededness: If two people have mutual acquaintances, they are more likely to encounter one another and become linked themselves



Red nodes are close to each other

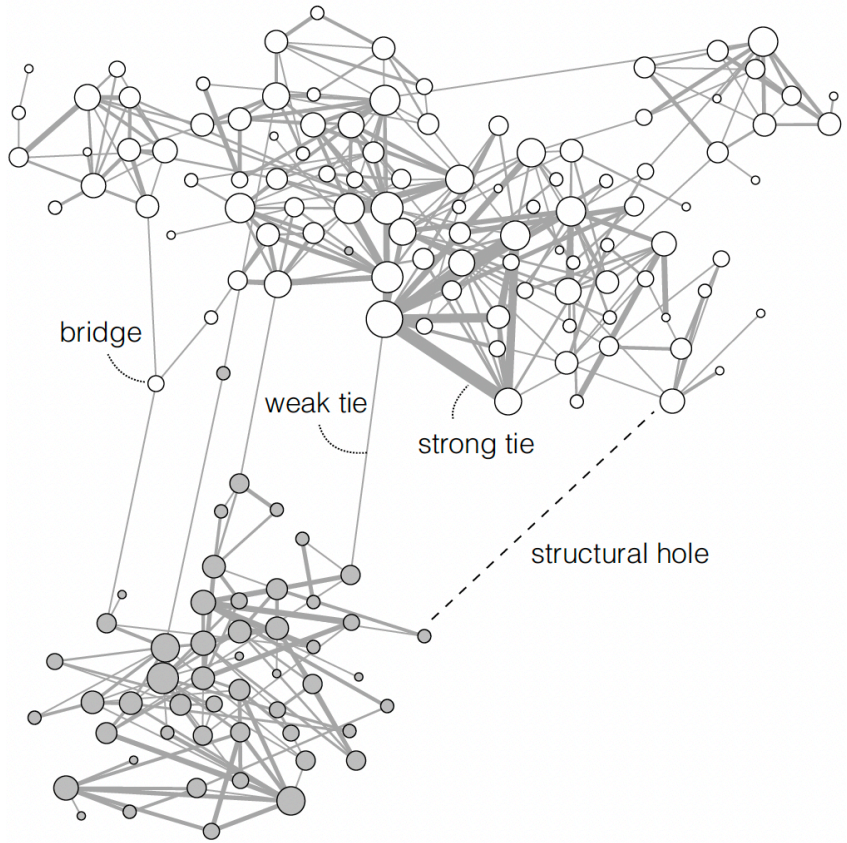


Red nodes are distant

Mechanisms

More on Embededness:

1. ***Embeddedness, clustering:*** People who spend time with a third are likely to encounter each other (triadic closure). Minimizes conflict, maximizes trusts,…
2. ***Bridges, structural holes*** (Burt): Bridges have structural advantages since they have access to non- redundant information
3. ***Weak ties*** **(Granovetter):** weak ties tend to connect different areas of the network (they are more likely to be sources of novel information)

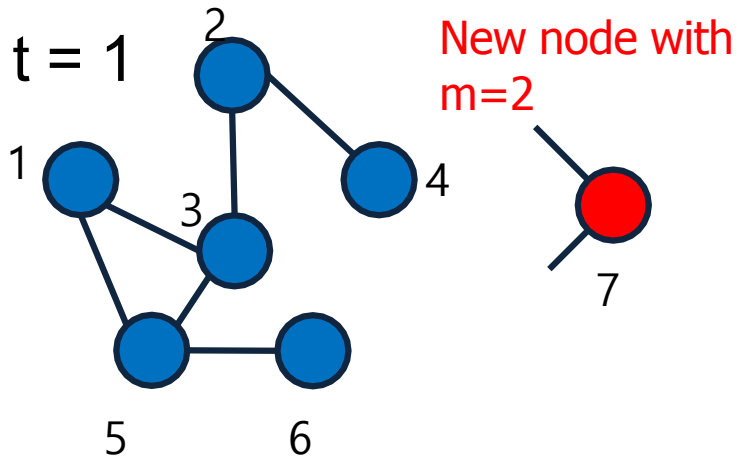


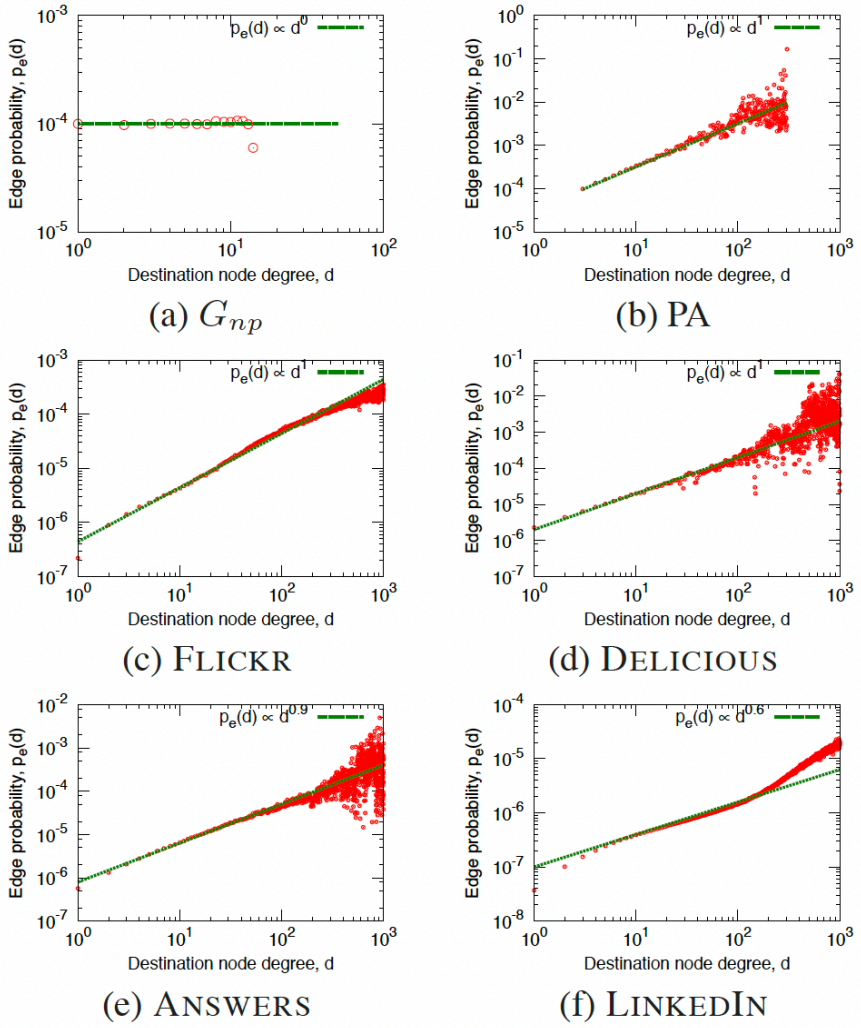
Rivera, M.T., Soderstrom, S.B. & Uzzi, B., 2010. Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annual Review of Sociology*, 36(1), pp.91–115.

Mechanisms

Preferential attachment: Nodes with larger degree are more likely to receive new connections

$$p(\text{connect to } i) \propto k_i$$



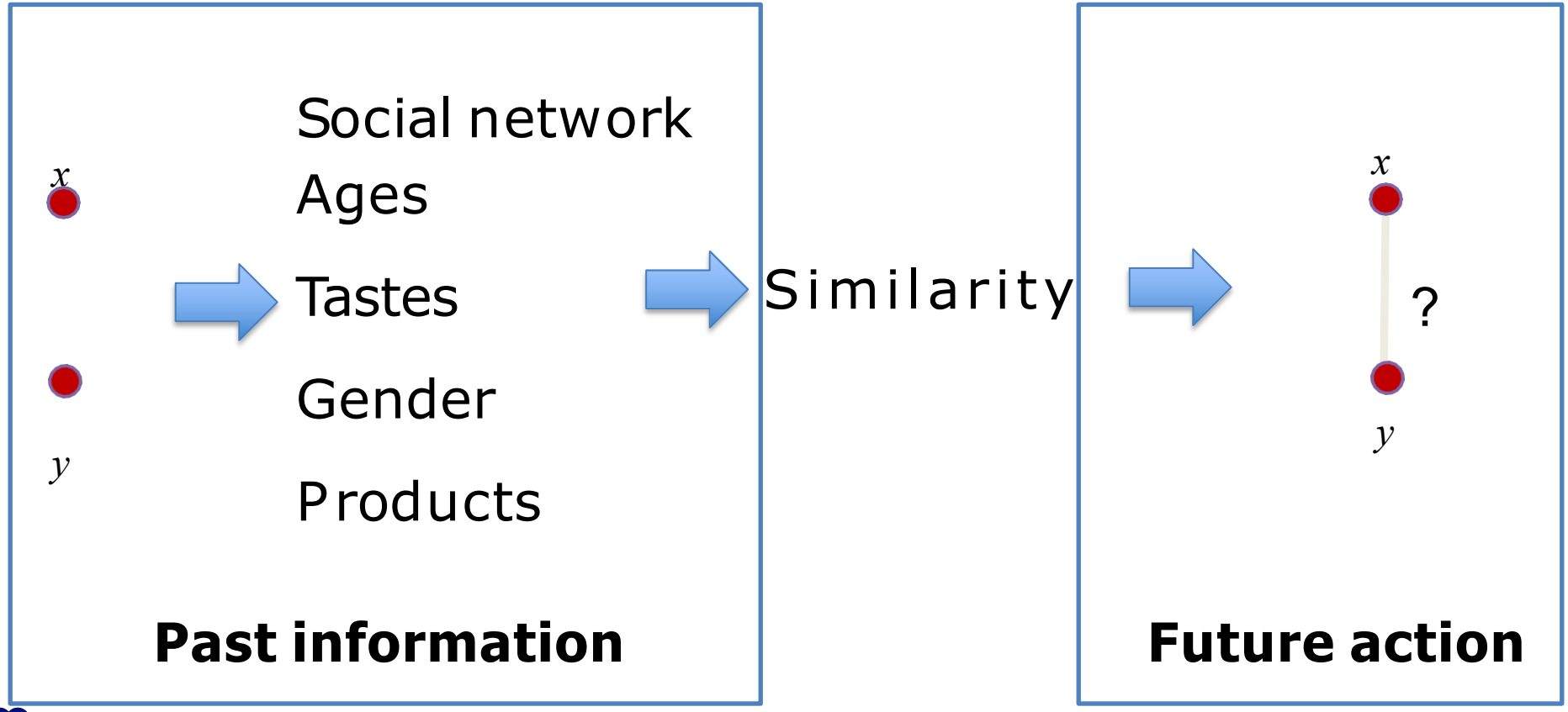


(a) G_{np} (b) PA (c) FLICKR (d) DELICIOUS (e) ANSWERS (f) LINKEDIN

Leskovec, J., Backstrom, L., Kumar, R., & Tomkins, A. (n.d.). Microscopic Evolution of Social Networks. *Cs.Cmu.Edu*.

Mechanisms

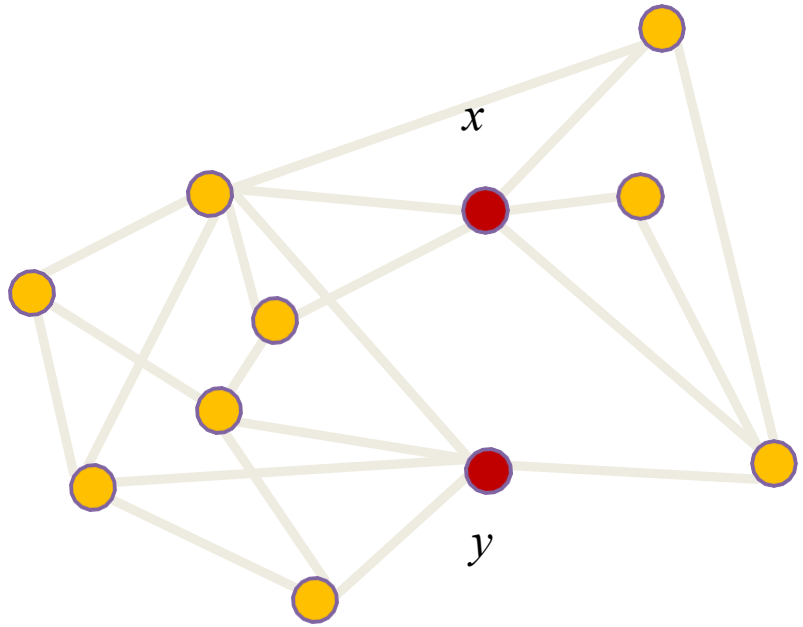
Link prediction problem: given two nodes, their characteristics, social connections, etc. in the past estimate the probability that they will establish a social connection in the future



2 | Link prediction using graph structure

Link prediction

Link prediction problem: in most settings only the social network (connections) is known



How can we define **similarity** in a social network only using the social structure?

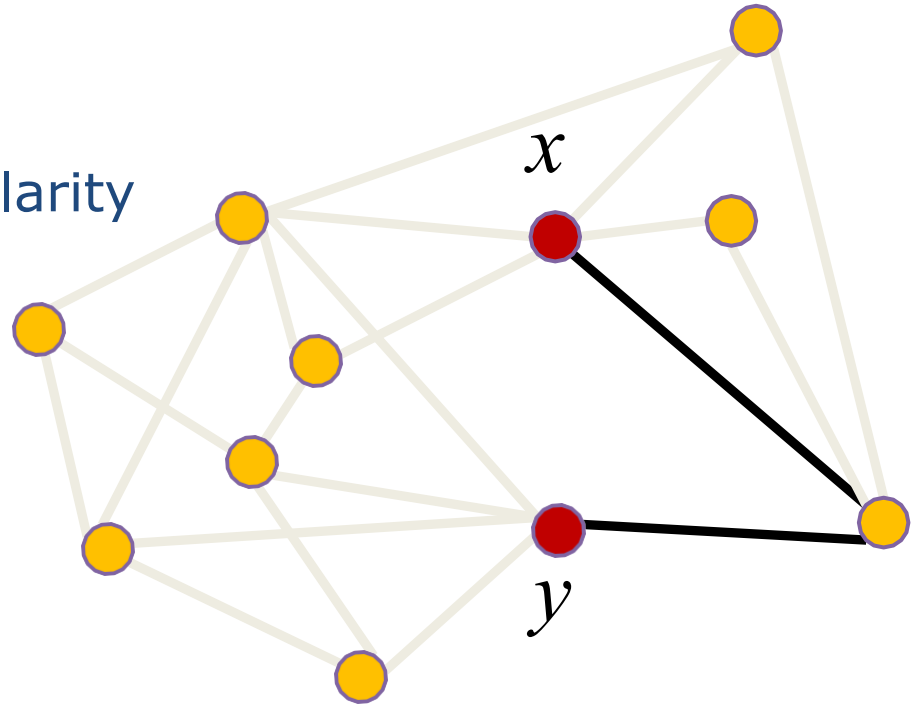
Link prediction

Graph distance: probably the simplest one, defined as the topological distance of two nodes in the graph

$$S(x,y) = \frac{1}{d_{xy}}$$

Larger distance means smaller similarity

$$S(x,y) = \frac{1}{2}$$



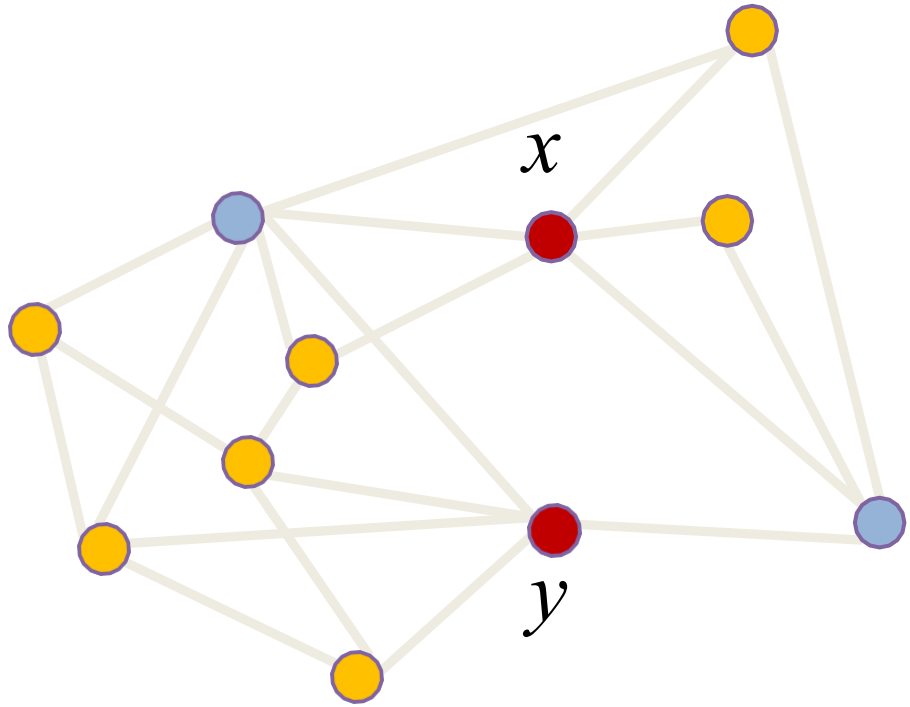
Link prediction

Common neighbors: two nodes that share a lot of neighbors might be introduced by a common friend

$$S(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

with $\Gamma(x)$: neighbors of x.

$$S(x, y) = 2$$



Link prediction

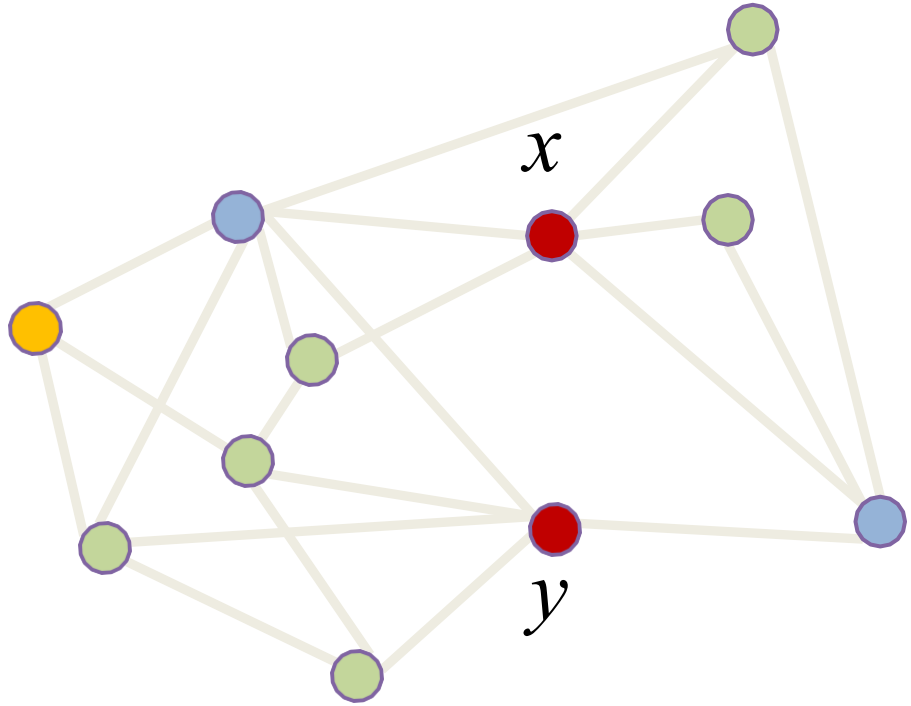
Jaccard's Coefficient: if nodes have very large connectivity, then they might probably share some neighbors. Thus we might consider instead the **fraction** of neighbors shared between them.

$$S(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \in [0,1]$$

common neighbors

total neighbors

$$S(x,y) = \frac{2}{8} = 0.25$$



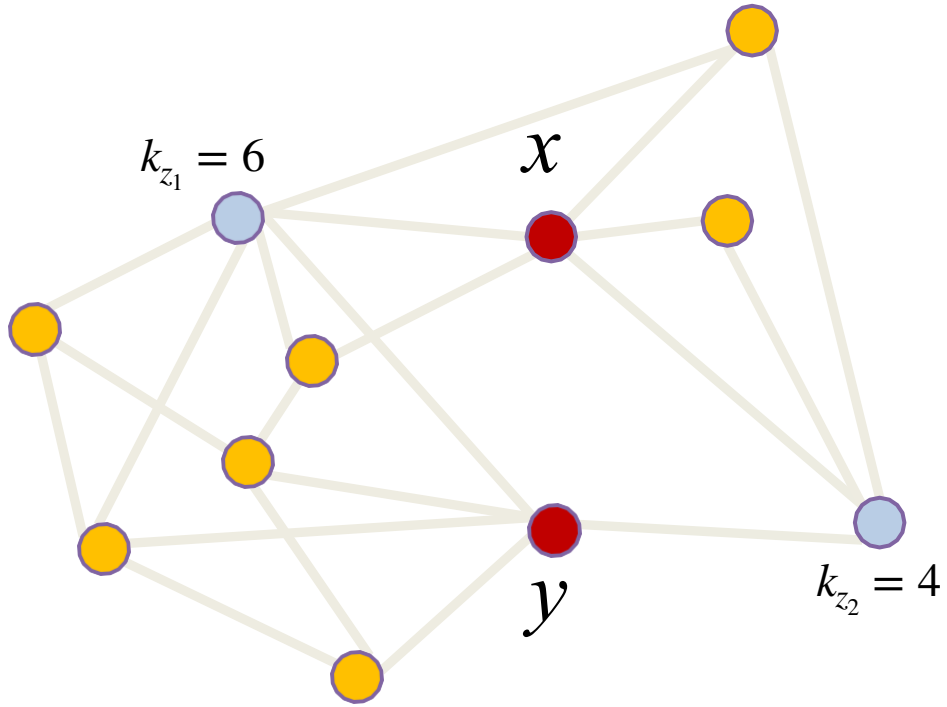
Link prediction

Adamic-Adar: *Common neighbors*, but weighted *inversely proportional to the degree* of each neighbor. Nodes with only x and y as neighbors (low k) count more than nodes with many contacts (high k) that happen to have x and y as neighbors.

$$S(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}$$

common neighbors

$$S(x,y) = \frac{1}{\log 6} + \frac{1}{\log 4} \approx 1.28$$

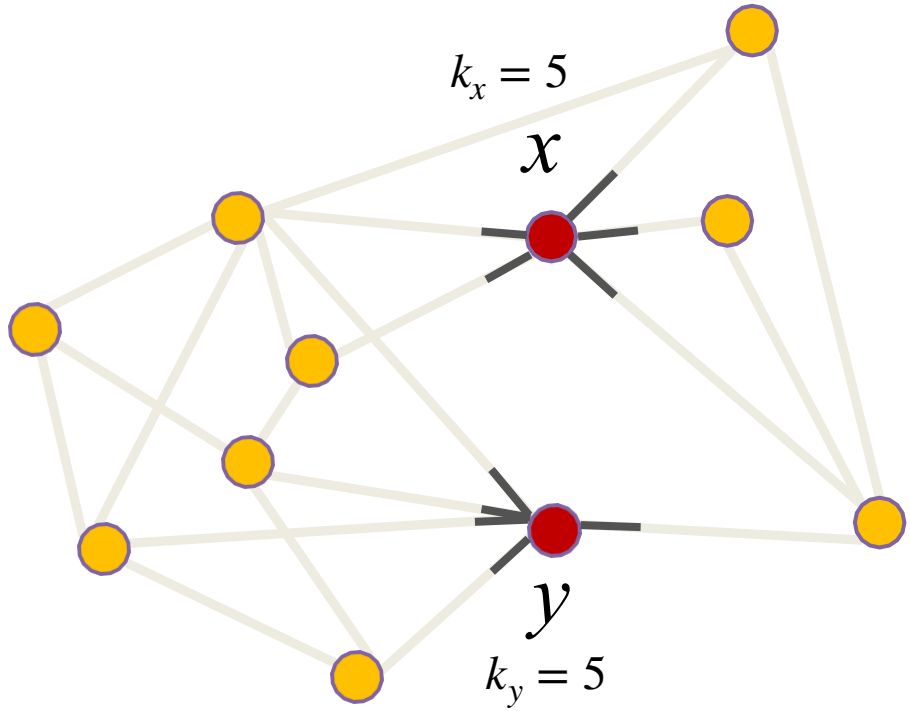


Link prediction

Preferential attachment: (rich get richer) if the probability that x and y get a new neighbor is proportional to their degree, the probability that a new link is formed between x and y grows with the *product of their degrees.*

$$S(x,y) = k_x \cdot k_y$$

$$S(x,y) = 5 \cdot 5 = 25$$



uc3m

Link prediction

Katz score: measures the *number of paths* between two nodes, attenuated by their length

$$S(x,y) = \sum_{l=1}^{d_{max}} \beta^l n_l(x,y)$$

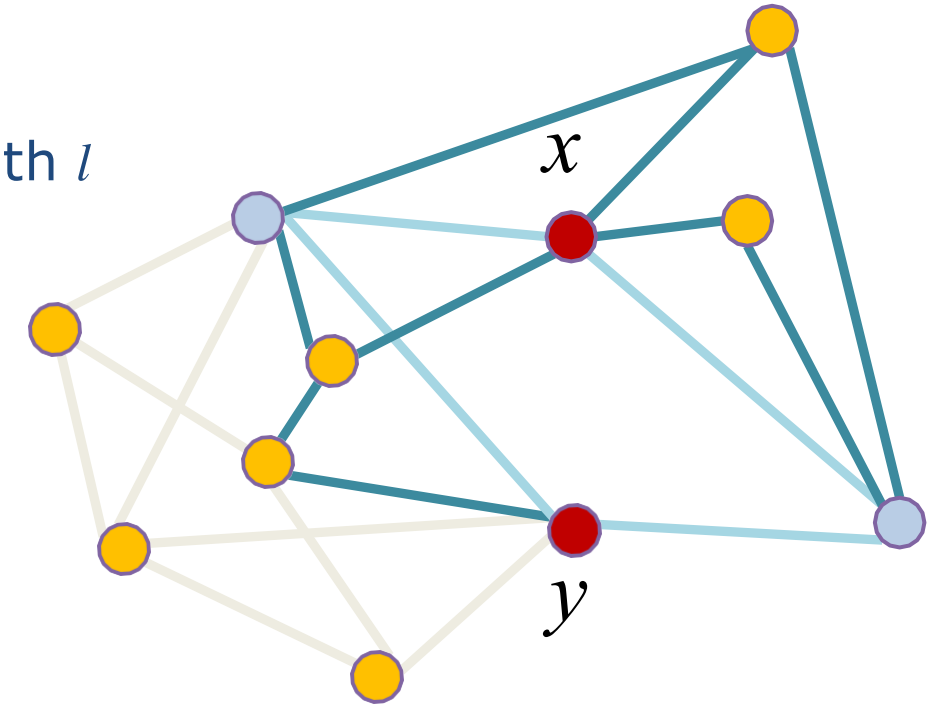
$n_l(x,y)$: number of paths of length l between x and y.

$\beta \leq 1$: attenuating factor.

$n_2(x,y) = 2 \quad n_3(x,y) = 5$

$$S(x,y) = \frac{1}{2} \cdot 2 + \frac{1}{2^2} \cdot 5 + \cdots$$

$(\beta = 1/2)$



3 | Example

Example

Prediction new scientific collaborations:

- Co-authorship network (G) from *author list* in papers
- **Training data**: G [1994, 1996]
- **Test data**: G' [1997, 1999]
- Core nodes: set of authors who have at least 3 papers during both training and test

	training period			Core		
	authors	papers	edges	authors	$\|E_{old}\|$	$\|E_{new}\|$
astro-ph	5343	5816	41852	1561	6178	5751
cond-mat	5469	6700	19881	1253	1899	1150
gr-qc	2122	3287	5724	486	519	400
hep-ph	5414	10254	17806	1790	6654	3294
hep-th	5241	9498	15842	1438	2311	1576

Liben-Nowell, D., & Kleinberg, J. (2003, November). The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management* (pp. 556-559).

Example

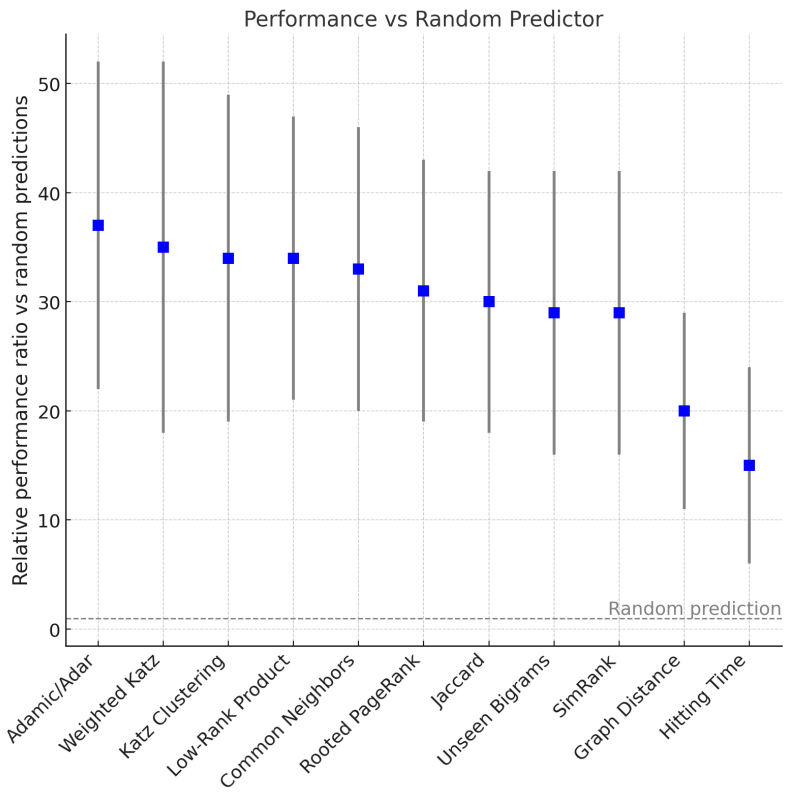
Link prediction algorithm

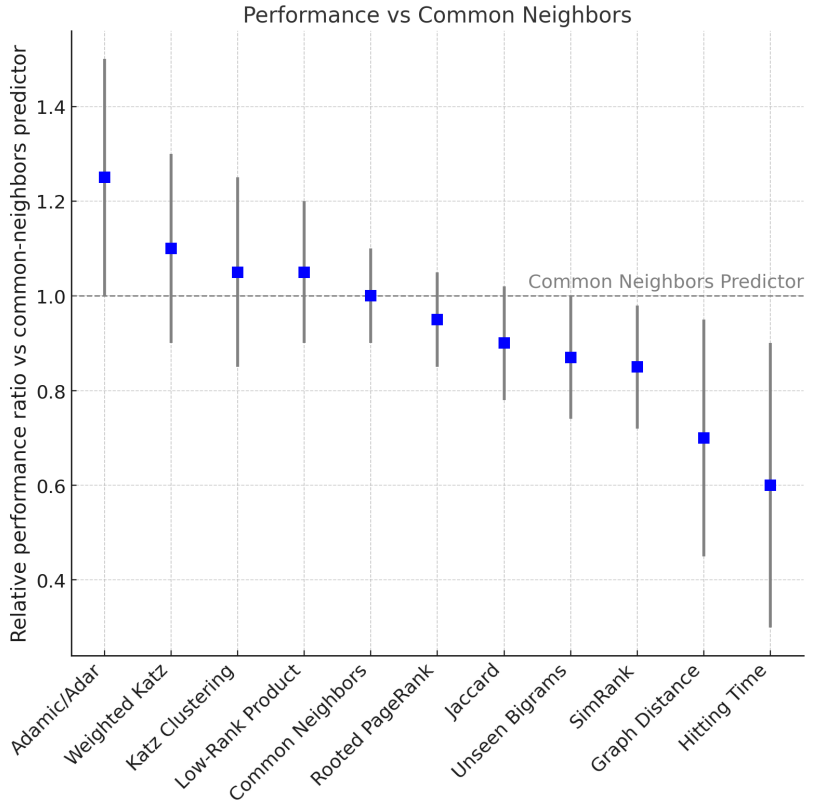
1. The input graph during training period
2. Assign each pair of nodes a similarity score $S(x, y)$
3. Predict links for the pairs with the **highest similarity score**
4. Compare the predicted links with the actual new links (*rank distance, AUC, accuracy, etc.*)
5. Compare the performance of the algorithm with
    1. Random predictor (each pair of nodes has a probability to have a new link)
    2. Graph distance predictor
    3. Common neighbors

Example

Results

- All algorithms are way better than random prediction. **Adamic-Adar** is 37 times better!
- But simple things like common neighbors work almost as well as A-A.



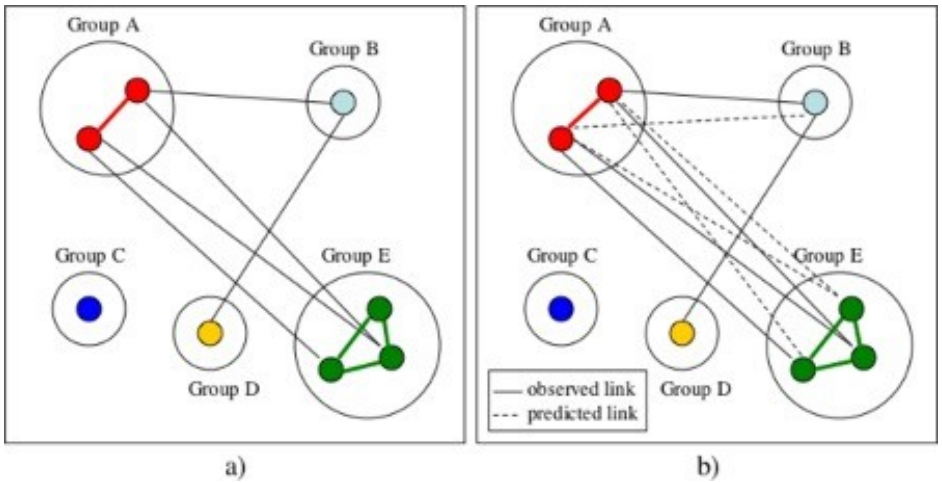


4 | Other approaches

Other approaches

Communities

- Nodes that belong to the same community will probably be linked



Dispersion

- Nodes that have many well-connected neighbors are more likely to have a strong link

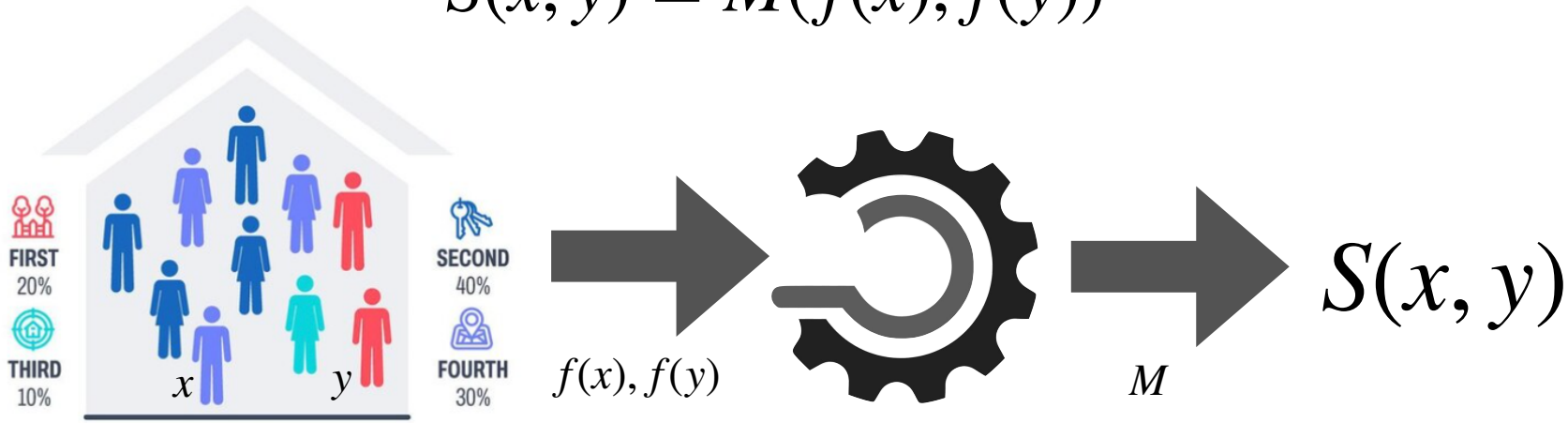
$$S(x,y) = \sum_{r,s \in \Gamma(x) \cap \Gamma(y)} - d_{rs}$$

Other approaches

Non-network data (e.g., homophily)

- Nodes that share sociodemographic characteristics might probably be linked

$$S(x, y) = M(f(x), f(y))$$



features of x and y

machine learning model

5 | Summary

Summary

1. Social mechanisms shape dynamics and persistence of networks
   - **i. Assortative:** homophily (attachment), heterophily (diversity)
   - **ii. Relational:** Network structure; common patterns
   - **iii. Proximity:** Geography increases connections and duration
2. Quantifying mechanisms → **similarity scores**
3. Similarity scores → **features for machine learning**
4. Link prediction models have been successful in several empirical networks from different domains