Overview of social network analysis

Emphasis:

- Properties of social networks
- Important analysis tasks
- Useful measures and solution principles

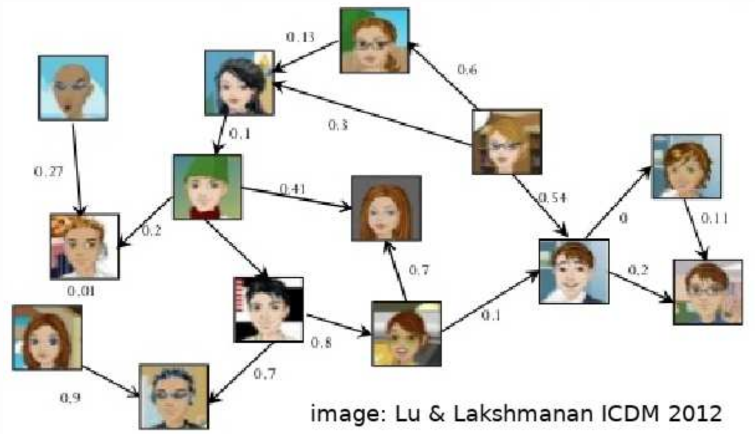


image: Lu & Lakshmanan ICDM 2012

More on course CS-E5740 **Complex Networks**

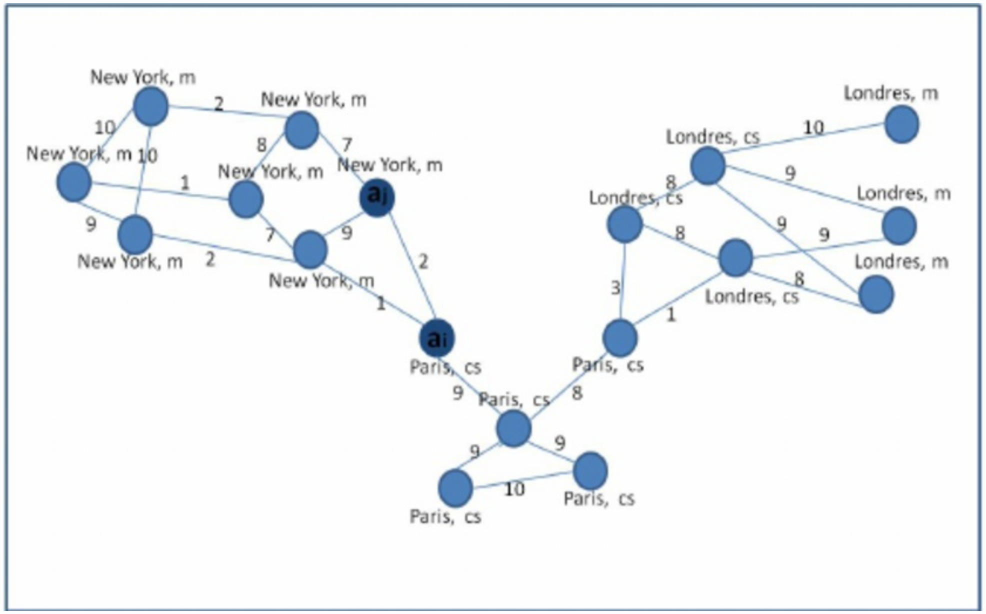
I Introduction: Types of social networks

- online networks (Twitter, LinkedIn, Facebook)
- indirect communication networks (telecommunications, email, chat messages)
- media sharing sites (Youtube, Instagram, Tiktok)
- interaction networks in professional communities (e.g., citation networks between researchers)
- networks recorded in observational studies (e.g., interactions in a class room, between animals)

+ many more! but not always data

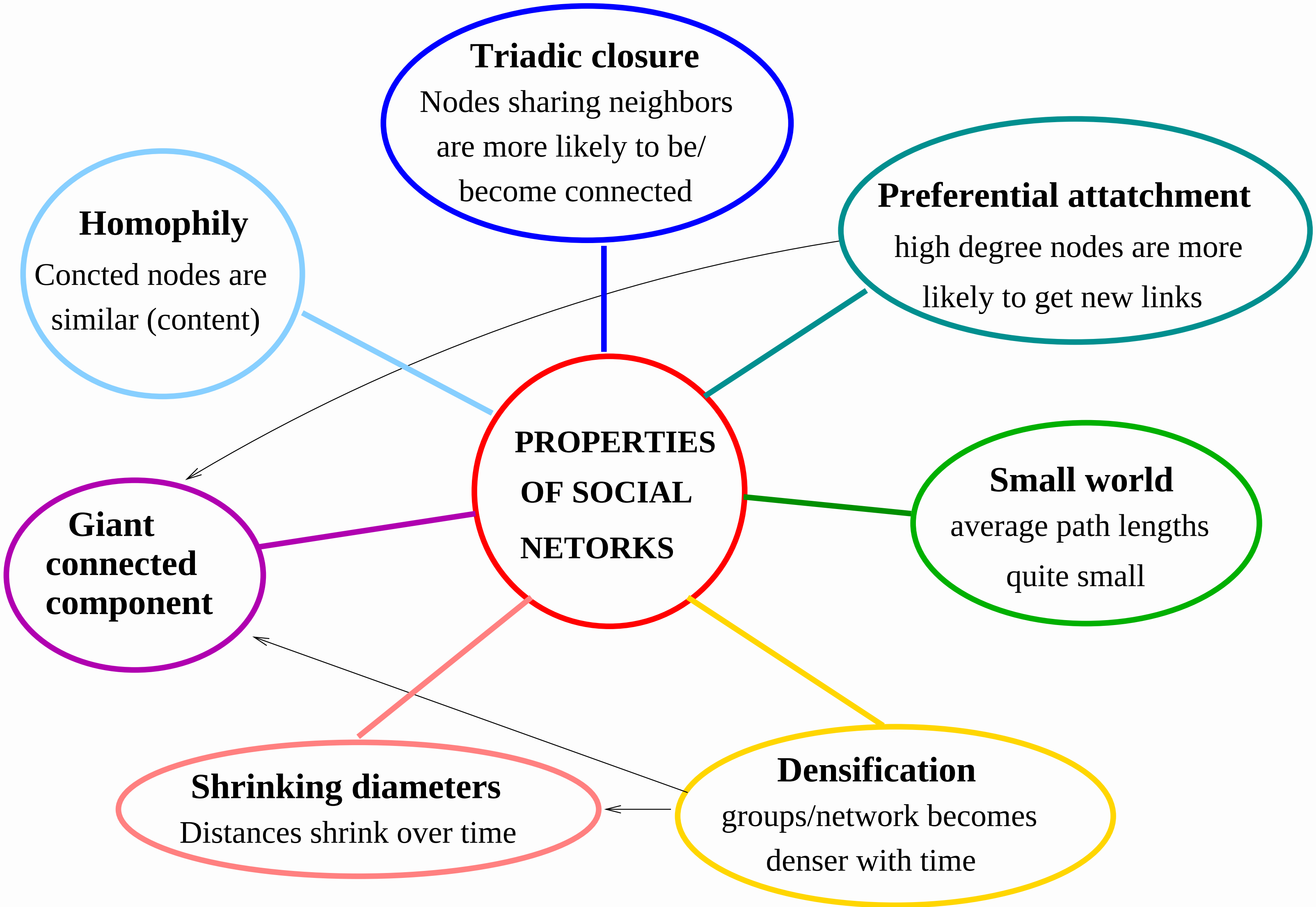
Presentation as a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$

- $\mathbf{V}$ set of nodes corresponding to **actors**
  - may have labels or content (attributes, documents)
- $\mathbf{E}$ set of edges corresponding to links
  - undirected (friendship) or directed (“following”)
  - may have weights w_{ij}



Example by Zardi et al. (2014)
node attributes: city and education
edge weight = number of exchanged messages

Basic properties



Analysis tasks

- Social influence analysis (influential nodes and influence spread)
- Community detection (graph clustering)
- Link prediction (predict future links between nodes)
- Collective classification (predict missing node labels)

II Social influence analysis

Which nodes have most influence? How influence (information, ideas, opinions) spreads?

A valuable advertising channel!

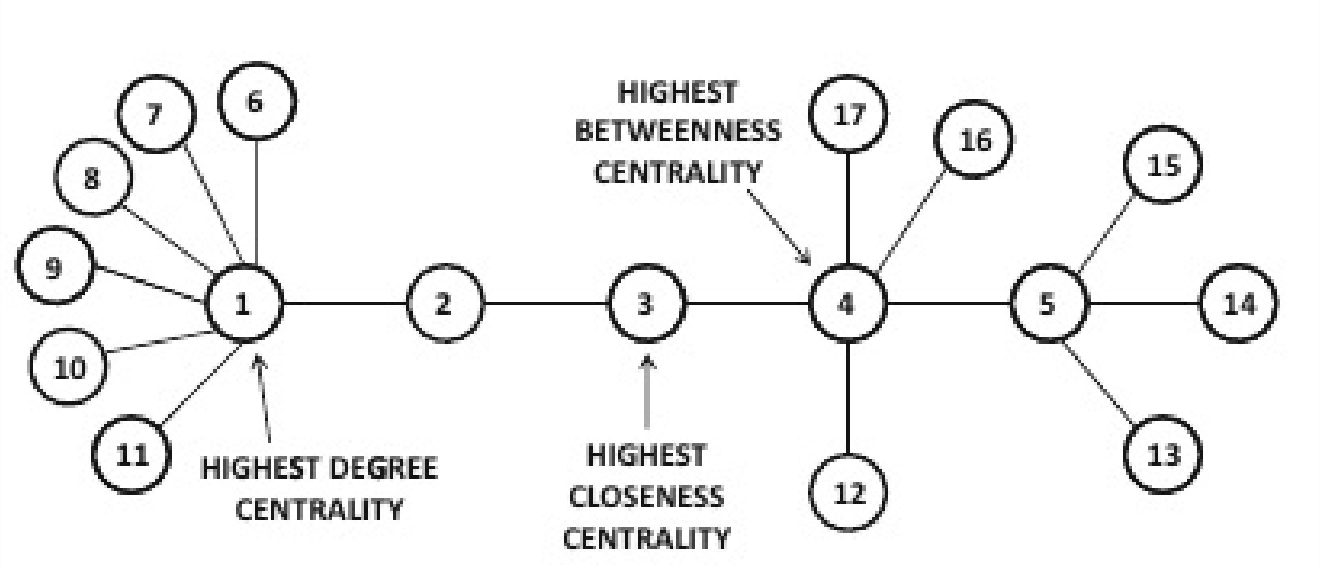
1. Measures for evaluating which nodes are influential:
   - **centrality** of a node in an undirected graph
   - **prestige** of a node in a directed graph
2. **Influence propagation** or **diffusion models**
   - given influence weights on edges and a model to evaluate total influence of a set of nodes
   - determine a set of **seed nodes** such that spread of influence is maximal

Measures for the centrality of node v

Degree centrality: $C_D(v) = \frac{Degree(v)}{n-1}$

Closeness centrality: $C_C(v) = \frac{1}{avg_{u \in V, u \neq v}\{Dist(v,u)\}} = \frac{n-1}{\sum_{u \in V, u \neq v} Dist(v,u)}$

Betweenness centrality: $C_B(v) = \frac{\sum_{u,w \in V, u \neq w} \frac{\#\{\text{shortest-paths(u,w) through v}\}}{\#\{\text{shortest-paths(u,w)}\}}}{\binom{n}{2}}$



Note: $C_c(v)$ may be calculated such $v \neq u$, $v \neq w$. Image: Aggarwal Fig. 19.1

III Community detection: cluster the graph

Given $\mathbf{G} = (\mathbf{V}, \mathbf{E})$. Each edge (v_i, v_j) has weight w_{ij}

- if cost c_{ij}, transform, e.g. by $w_{ij} = \frac{1}{c_{ij}}$ $(c_{ij} \neq 0)$

Common objective: Cluster $\mathbf{V}$ into groups $\mathbf{V}_1, \ldots, \mathbf{V}_K$ such that the edge-cut cost

$$cost(\mathbf{V}_1, \ldots, \mathbf{V}_k) = \sum_{(v_i,v_j)\in E, v_i\in\mathbf{V}_p, v_j\in\mathbf{V}_q, p\neq q} w_{ij}$$

is minimal.

- **many variants and extra constraints!**
- in general NP-hard problem, but polynomially solvable, if $\forall i, j : w_{ij} = 1$, $K = 2$ and no balancing requirements

Example

Clustering based on both structural and content-based features

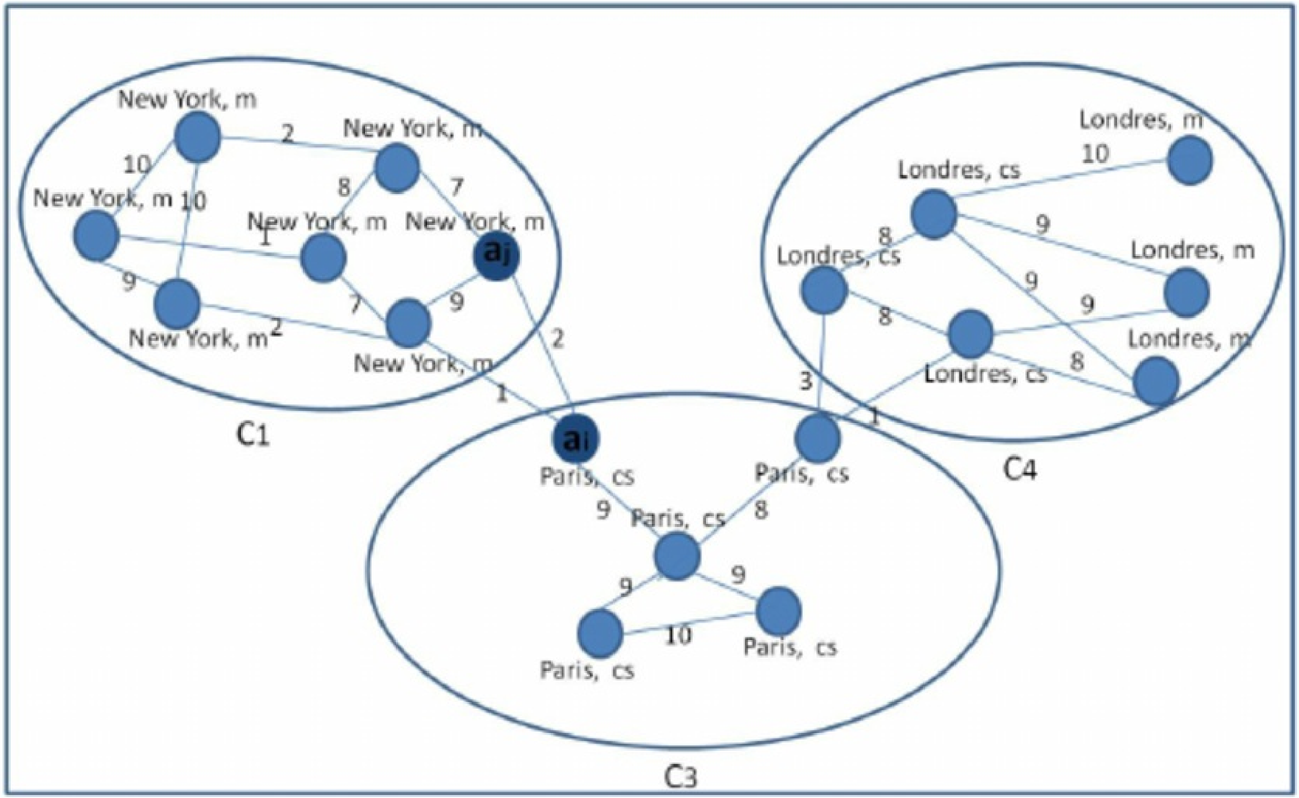
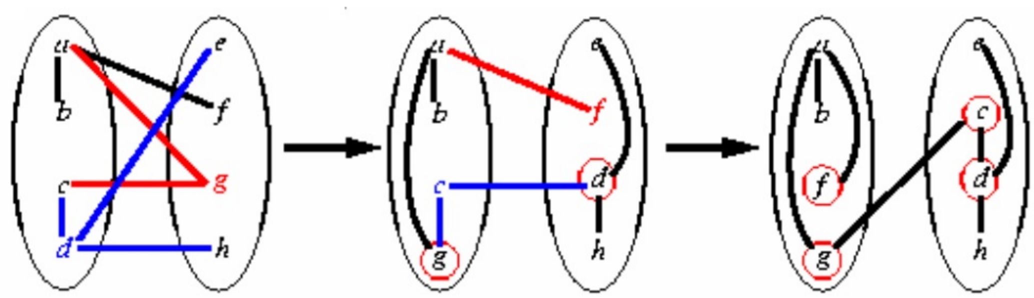


Image source: Zardi et al.: A Multi-agent homophily-based approach for community detection in social networks, ICTAI 2014

Some community detection methods

1. **Spectral clustering**
2. **Kerninghan-Lin**: balanced 2-way partitioning

- at each iteration, test a set of possible swap sequences and choose the one with greatest improvement



Step #	Vertex pair	Cost reduction	Cut cost
0	–	0	5
1	{d, g}	3	2
2	{c, f}	1	1
3	{b, h}	-2	3
4	{a, e}	-2	5

Image source: Chang 2004

3. Girwan-Newman algorithm

- remove “bridge edges” until K connected components remain
- edges with high **betweenness**: large proportion of shortest paths go through them

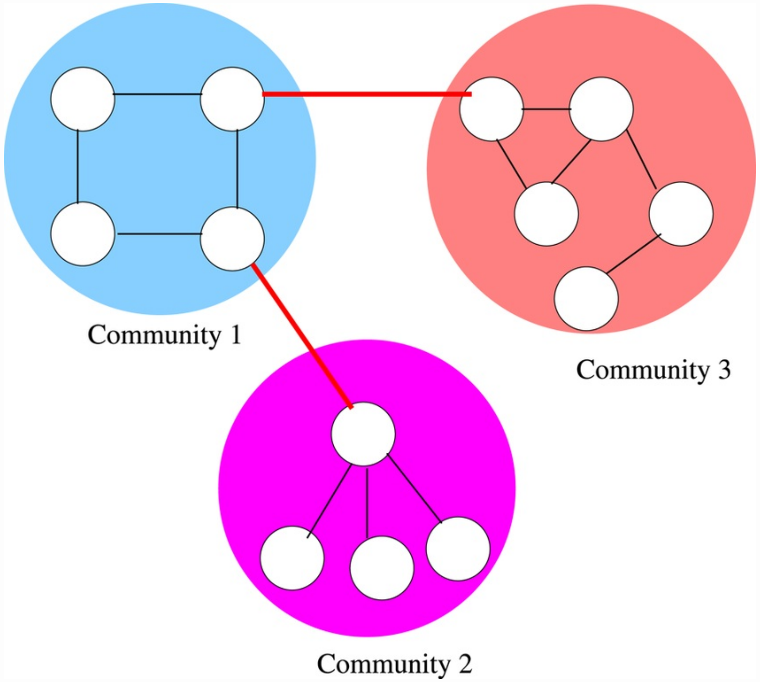


Image source: Namtirtha et al. 2023

4. METIS algorithm

1. Coarsen the graph by combining tightly interconnected nodes and parallel edges
2. Partition the coarsened representation (easier)
3. Refine partitioning when expanding graphs back

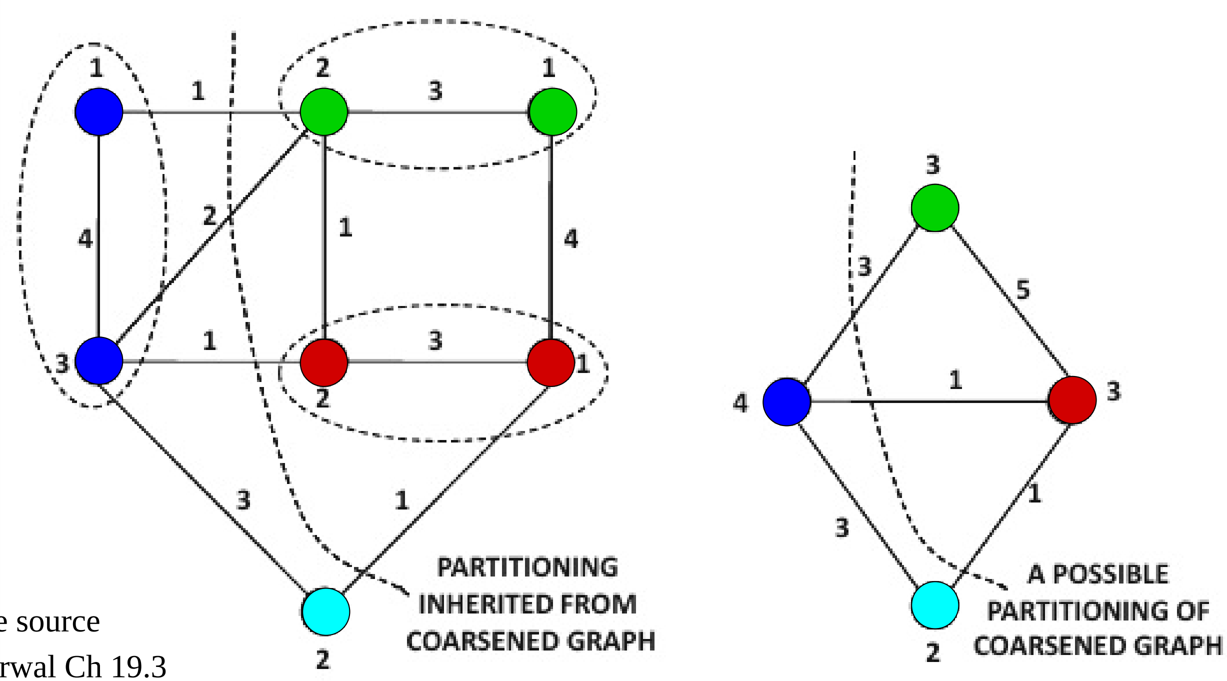


image source
Aggarwal Ch 19.3

Image source: Aggarwal Fig. 19.6

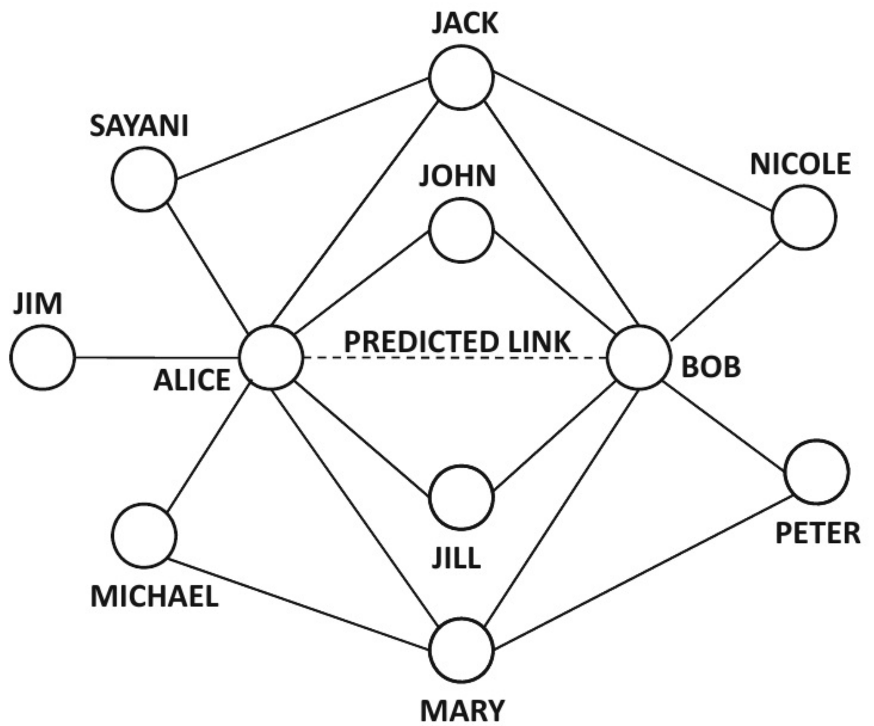
IV Link prediction and node similarity

Utilize especially **structural** features!

Approaches:

1. Evaluate potential connections with **node similarity measures**
   - \+ easy and fast to compute
2. Learn a classifier for predicting links or their absence
   - \+ more accurate
   - − computationally more expensive
3. Use missing value estimation methods (like matrix factorization)

Neighbourhood-based node similarity measures



(a) Many common neighbors between Alice and Bob

(normalized) number of common neighbours

- not good, if number of common neighbours small

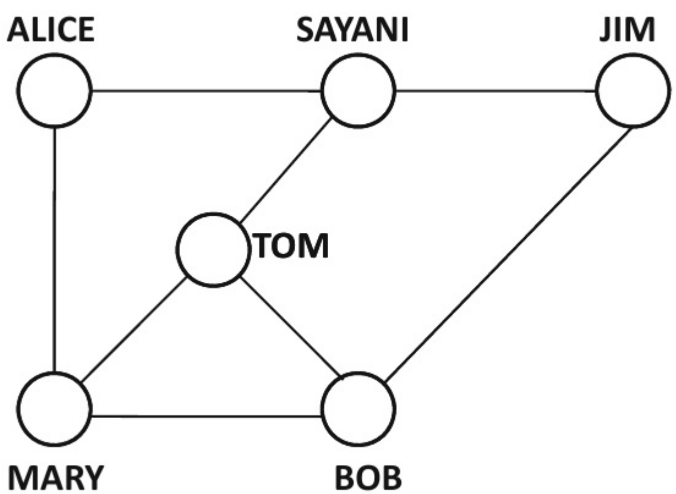
- $Jaccard(v_i, v_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$
- $AdamicAdar(v_i, v_j) = \sum_{v_k \in S_i \cap S_j} \frac{1}{\log(|S_k|)}$

$S_i = \{v_k \mid v_k \text{ neighbour of } v_i\}$

Image source: Aggarwal Fig. 19.12

Walk-based node similarity measures

Is Alice more similar to Bob or Jim?



(b) Many indirect connections between Alice and Bob

- Personalized PageRank with teleportation to v_i
- SimRank
- Katz measure

$$Katz(v_i, v_j) = \sum_{t=0}^{\infty} \beta^t \cdot n_{ij}^{(t)}$$

$n_{ij}^{(t)}$ = number of walks of length t between v_i and v_j

$\beta < 1$ discount factor (punishes long walks)

Image source: Aggarwal Fig. 19.12

Image sources

- Chang (2004): Unit 4: Circuit partitioning (lecture slides). EDA course, National Taiwan University. `http://cc.ee.ntu.edu.tw/~ywchang/Courses/EDA04/lec4.pdf`
- Namtirtha et al. (2023): Placement Strategies for Water Quality Sensors Using Complex Network Theory for Continuous and Intermittent Water Distribution Systems. Water Resources Research 59(7), doi:10.1029/2022WR033112.
- Zardi et al. (2014): A Multi-agent homophily-based approach for community detection in social networks, IEEE 26th Int. Conf. Tools with Artificial Intelligence, doi: 10.1109/ICTAI.2014.81.