# Chapter 5

# Social Network Analysis

# Introduction

In a social network we consider **social entities** such as a person or organization, which are represented by nodes. The edges in the social network represent social relationships between incident nodes.

**Social Network Analysis** tries to explain social phenomena based on the structure of the relationships between the social entities.

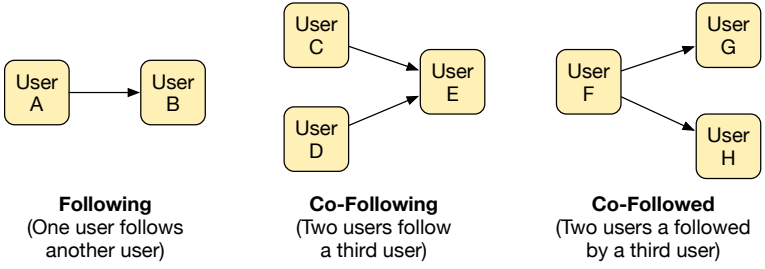
# Tools for graph processing and network analysis

Example graph processing and network analysis tools:

- http://www.neo4j.org
- http://orientdb.com
- https://gephi.org
- http://www.netminer.com
- http://graphexploration.cond.org/
- http://www.orgnet.com
- http://thinkaurelius.github.io/titan
- https://networkx.github.io/
- http://www.netvis.org
- http://www.analytictech.com
- http://visone.info
- http://www.caida.org/tools/
- http://www.stanford.edu/group/sonia
- http://www.touchgraph.com
- http://linkurio.us/
- Boost Graph Library
- http://pajek.imfm.si

# Dyadic measures

When analyzing the relationships between pairs (**dyads**) of social entities, we can derive secondary relationships.
Example: "follow" relationship in a micro blogging social network:




Questions:

- How many users are following both G and H? (also known as **cocitation**)
- How many users are followed by both C and D? (also known as **bibliographic coupling**)

# Cocitation

The **cocitation** $C_{ij}$ of two vertices $i$ and $j$ in a directed network is the number of vertices that have outgoing edges pointing to both $i$ and $j$:

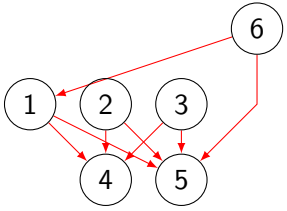
$$C_{ij} = \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{A}_{jk} = \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{A}^T_{kj}$$

Where $\mathbf{A}^T_{kj}$ is an element ot the transpose of $\mathbf{A}$.
We can define the **cocitation matrix** $\mathbf{C}$ to be the matrix with elements $C_{ij}$, which is thus given by

$$\mathbf{C} = \mathbf{A}\mathbf{A}^T$$

Note that $\mathbf{C}$ is symmetric, since $\mathbf{C}^T = (\mathbf{A}\mathbf{A}^T)^T = \mathbf{A}\mathbf{A}^T = \mathbf{C}$.

# Cocitation example




$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{A^T} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{C} = \mathbf{A}\mathbf{A}^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 3 & 0 \\ 0 & 0 & 0 & 3 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$C_{ii}$ is equal to the number of edges pointing to $i$ ("the number of papers citing $i$").

# Bibliographic Coupling

The **bibliographic coupling** $B_{ij}$ of two vertices $i$ and $j$ in a directed network is the number of other vertices to which both point:

$$B_{ij} = \sum_{k=1}^{n} \mathbf{A}_{ki}\mathbf{A}_{kj} = \sum_{k=1}^{n} \mathbf{A}_{ik}^{T}\mathbf{A}_{kj}$$

We can define the **bibliographic coupling matrix** $\mathbf{B}$ to be the $n \times n$ matrix with elements $B_{ij}$ so that

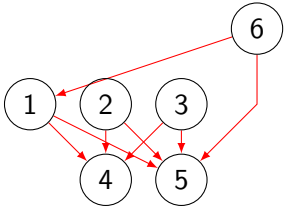
$$\mathbf{B} = \mathbf{A}^{T}\mathbf{A}$$

The diagonal elements of $\mathbf{B}$ are

$$\mathbf{B}_{ii} = \sum_{k=1}^{n} \mathbf{A}_{ki}^{2} = \sum_{k=1}^{n} \mathbf{A}_{ki}$$

Thus $\mathbf{B}_{ii}$ is equal to the number of other vertices that vertex $i$ points to ("the number of papers $i$ cites").

# Bibliographic coupling example




$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{A^T} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{B} = \mathbf{A^T A} = \begin{pmatrix} 2 & 2 & 2 & 0 & 0 & 1 \\ 2 & 2 & 2 & 0 & 0 & 1 \\ 2 & 2 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 2 \end{pmatrix}$$

$B_{ii}$ is equal to the number of other vertices that $i$ points to ("the number of papers $i$ cites").

# Co-Citation vs. Bibliographic Coupling

- Both co-citation and bibliographic coupling are affected strongly by the number of ingoing and outgoing edges that vertices have.
- For instance, two papers can only have strong co-citation if they are both well cited. Conversely, strong bibliographic coupling requires large bibliographies in papers.
- In practice, the sizes of bibliographies vary less than the number of citations papers receive and thus bibliographic coupling is a more uniform indicator of similarity between papers than citation.

# Structural Equivalence and Regular Equivalence

Both co-citation and bibliographic coupling belong to a set of measures that allow us to analyze the **similarity** between vertices. Similarity between entities can be defined in many different ways, however, here we focus on measures that utilize only the network structure.

For networks, there exist two fundamental approaches for similarity measures:

- **Structural equivalence**: two vertices share many of the same network neighbors.
- **Regular equivalence**: two vertices do not necessarily share the same set of neighbors, but they have neighbors who are themselves similar.

# Structural Equivalence: Cosine Similarity

In an undirected network the number $n_{ij}$ of common neighbors of vertices $i$ and $j$ is given by

$$n_{ij} = \sum_k A_{ik} A_{kj}$$

(This is very closely related to co-citation, which is defined for directed networks, but otherwise it is the essentially the same thing.)

- Without normalization, the number of common neighbors alone is not a good measure of similarity.
- However, simply dividing, e.g., by the total number of vertices is also not sufficient, because it penalizes low-degree vertices.

A better measure is the **cosine similarity**, which measures the angle $\phi$ between two vectors $x$ and $y$:

$$cos(\phi) = \frac{x \cdot y}{|x| \cdot |y|}$$

# Structural Equivalence: Cosine Similarity

As proposed by Salton, we can use the $i$th and $j$th rows (or columns) of the adjacency matrix as two vectors and use the cosine of the angle between them as a similarity measure:

$$\sigma_{ij} = cos(\phi) = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2} \sqrt{\sum_k A_{jk}^2}} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{k_i k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

The **cosine similarity** of $i$ and $j$ is therefore the number of their common neighbors divided by the geometric mean of their degrees.

- In case both vertices have degree zero, we normally set $\sigma_{ij} = 0$.
- The value of $\sigma_{ij}$ always lies in the range from 0 to 1.
- If $\sigma_{ij} = 1$, then $i$ and $j$ have exactly the same neighbors.
- If $\sigma_{ij} = 0$, then $i$ and $j$ have none of the same neighbors.

# Regular Equivalence

**Regular equivalence** is another type of similarity that (in contrast to structural equivalence) does not only observe the direct neighbors of two nodes, but also considers two nodes similar that, while they do not necessarily share neighbors, have neighbors who are themselves similar[1].

- One approach to model this kind of similarity is:

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl}$$

  (which is basically a type of Eigenvector equation: $\sigma = \alpha \mathbf{A} \sigma \mathbf{A}$)
- This basic approach has some problems, such as that it does not give a vertex high self-similarity to itself and when solved by repeated iteration, we will get a sum over even powers of the adjacency matrix.
- The issue of self-similarity can be solved by adding an extra diagonal term:

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl} + \delta_{ij}$$

[1] Source: Networks - An Introduction, M.E.J. Newman, Oxford University Press, 2010

# Regular Equivalence

Thus a better definition of regular equivalence is: vertices $i$ and $j$ are similar if $i$ has a neighbor $k$ that is itself similar to $j$:

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}$$

or

$$\sigma = \alpha \mathbf{A}\sigma + \mathbf{I}$$

If we evaluated this expression by iterating and starting from $\sigma^{(0)} = 0$ we get

$$\sigma^{(1)} = \mathbf{I}, \sigma^{(2)} = \alpha\mathbf{A} + \mathbf{I}, \sigma^{(3)} = \alpha^2\mathbf{A}^2 + \alpha\mathbf{A} + \mathbf{I}$$

In the limit of a large number of iterations this gives:

$$\sigma = \sum_{m=0}^{\infty} (\alpha\mathbf{A})^m = (\mathbf{I} - \alpha\mathbf{A})^{-1}$$

# Regular Equivalence

Thus $\sigma = (\mathbf{I} - \alpha\mathbf{A})^{-1}$ can be seen as a wighted count of all the paths between the vertices $i$ and $j$ with paths of length $r$ getting weight $\alpha^r$. As long as $\alpha < 1$, longer paths will get less weight than shorter ones.

$$\sigma = (\mathbf{I} - \alpha\mathbf{A})^{-1}$$

is also reminiscent of the Katz centrality (although Katz himself never discussed it). The Katz similarity of a vertex would then be the sum of the "Katz similarities" of that vertex to all others.

# Affiliation Networks

In an **affiliation network**, people (actors) are connected to each other through a membership relation.

- For example, two persons may be members of the same sports club, or they work in the same team, or they attend the same event.

We can represent an affiliation network as a bipartite graph, where the vertices represent actors and events, and an edge indicates that an actor participates in an event.

- Affiliation networks are also known as **two mode networks** because they consist of two different sets.

# Adjacency Matrix of an Affiliation Network

- Let $V_A$ be the set of actor vertices and let $V_E$ be the set of event vertices.
- The **actor event matrix** $\mathbf{AE}$ consists of $n_A = V_A$ rows and $n_E = V_E$ columns, where $\mathbf{AE}_{ij} = 1$ if and only if actor $i$ participates in event $j$.
- In general, the matrix $\mathbf{AE}$ is rectangular (and not square).

The number of events in which both actor $i$ and $j$ participated is given by

$$\mathbf{NE}_{ij} = \sum_{k=1}^{n_E} \mathbf{AE}_{ik} \cdot \mathbf{AE}_{jk}$$

Similarly, the number of actors participating in both event $i$ and $j$ is given by

$$\mathbf{NA}_{ij} = \sum_{k=1}^{n_A} \mathbf{AE}_{ki} \cdot \mathbf{AE}_{kj}$$