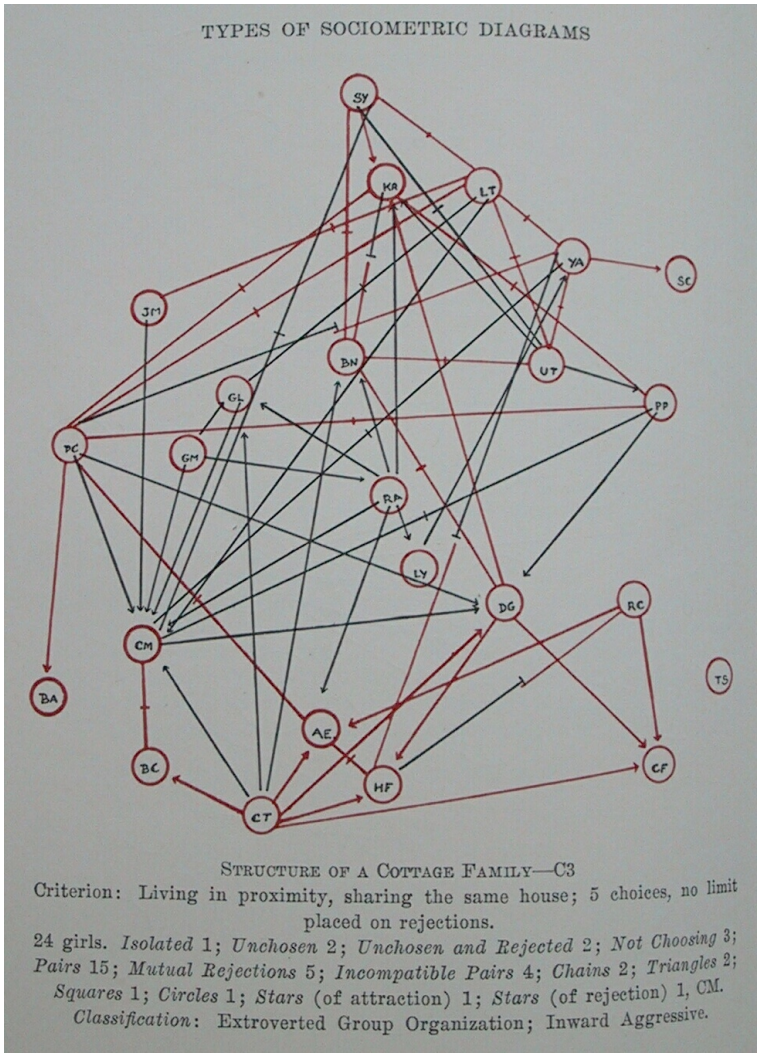


Course on Social Network Analysis

Weights

Vladimir Batagelj

University of Ljubljana

Slovenia

Padova, April 10-11, 2003

Outline

Line-Cuts

If we *line-cut* a network $\mathbf{N} = (V, L, w)$, $w : L \rightarrow \mathbb{R}$ at selected level t

$$L' = \{e \in L : w(e) \geq t\}$$

we get a subnetwork $\mathbf{N}(t) = (V(L'), L', w)$, $V(L')$ is the set of all end-vertices of the lines from L'.

We then look at the components of $\mathbf{N}(t)$. Their number and sizes depend on t. Usually there are many small components. Often we consider only components of size at least k.

The values of thresholds t and k are determined by inspecting the distribution of weights and the distribution of component sizes.

Vertex-Cuts

In some networks we can have also a function $p : V \rightarrow \mathbb{R}$ that describes some property of vertices. Its values can be obtained by measuring, or they are computed (for example, centrality indices).

The *vertex-cut* of a network $\mathbf{N} = (V, L, p)$ at selected level t is a network $\mathbf{N}(t) = (V', L(V'), p)$, determined by the set

$$V' = \{v \in V : p(v) \geq t\}$$

and $L(V')$ is the set of lines from L that have both end-vertices in V'.

Vertex weights

The most important distinction between vertex weights is based on the decision whether the graph is considered directed or undirected. This gives us two main types of weights:

- undirected case: *measures of centrality*;
- directed case: *measures of importance*; with two subgroups: *measures of influence*, based on outgoing lines; and *measures of support*, based on incoming lines.

For undirected graphs all three types of measures coincide.

Another important division of vertex weights is:

- *local* measures, which consider only immediate neighbors of a given vertex;
- *global* measures, which consider all vertices connected by paths with a given vertex.

Degree

Simple local weights are degrees: $\deg, \mathrm{indeg}, \mathrm{outdeg}$. To ensure comparability of measures for graphs on different vertex sets we normalize them by dividing them with the maximal possible value d_{max}. This gives us weights

$$\mathrm{c}(v) = \frac{\deg(v)}{d_{max}}, \qquad \mathrm{c}_{in}(v) = \frac{\mathrm{indeg}(v)}{d_{max}}, \qquad \mathrm{c}_{out}(v) = \frac{\mathrm{outdeg}(v)}{d_{max}}$$

It holds $0 \leq \mathrm{c}(v), \mathrm{c}_{in}(v), \mathrm{c}_{out}(v) \leq 1$.

Closeness

Let G be strongly connected. Sabidussi (1966) introduced the following measure of *closeness*

$$\mathrm{cl}(v) = \frac{n-1}{\sum_{u \in V} d(v,u)}$$

where $d(v,u)$ is the length of geodesics from v to u. It holds $0 \leq \mathrm{cl}(v) \leq 1$.

Betweenness

Freeman (1977) defined *betweenness* of a vertex v by

$$\mathrm{bw}(v) = \frac{1}{(n-1)(n-2)} \sum_{\substack{u,z \in V : n(u,z) \neq 0 \\ u \neq z, v \neq u, v \neq z}} \frac{n(u,z;v)}{n(u,z)}$$

where $n(u,z)$ is the number of geodesics from u to z and $n(u,z;v)$ is the number of geodesics from u to z passing through v. It holds $0 \leq \mathrm{bw}(v) \leq 1$.

Flow measures

Degree, closeness and betweenness measures are based on the *economy assumption* – communications use the geodesics. Bonacich (1971) and Stephenson and Zelen (1989) proposed two indices which consider all possible communication paths.

Hubs and authorities

In directed networks we can usually identify two types of important vertices: hubs and authorities.

A vertex is a *good authority*, if it is pointed to by many good hubs, and it is a *good hub*, if it points to many good authorities.

Let $x(v)$ be an authority weight of vertex v and $y(v)$ its hub weight. Then

$$x(v) = \sum_{(u,v)\in L} w(u,v)y(u) \qquad \text{and} \qquad y(v) = \sum_{(v,u)\in L} w(v,u)x(u)$$

The hubs and authorities are a refinement of input and output degrees: in the case of input degree we only count incoming lines while for authorities it is important also from whom the lines are coming (important or less important vertices); the same holds for hubs and output degree.

Network centralization measures

Network centralization measures measure the extent to which the network supports/is dominated by a single node. They are usually constructed combining the corresponding vertex values. We have to consider two problems:

- is a vertex measure defined also for nonconnected networks; isolated vertices?
- is a measure comparable over different networks?

There are two general approaches.

Extremal approach

Let $p(v)$ be a node measure. We introduce the quantities

$$\begin{aligned} p^\star &= \max_{v \in V} p(v) \\ D &= \sum_{v \in V} (p^\star - p(v)) \\ D^\star &= \max\{D(G) : G \text{ is a graph on } V\} \end{aligned}$$

Then we can define

$$p_{max} = \frac{D}{D^\star}$$

It can be shown that among connected graphs for all these measures (c, cl, bw) the (directed) star (with a loop in the root) is a maximal graph ($p_{max} = 1$) and the complete graph is a minimal graph ($p_{max} = 0$).

Variational approach

The other approach is based on variance. First we compute the average vertex centrality

$$\overline{p} = \frac{1}{n} \sum_{v \in V} p(v)$$

and then define

$$p_{var} = \frac{1}{n} \sum_{v \in V} (p(v) - \overline{p})^2$$

Triangular connectivities

The notion of connectivity can be extended to connectivity by chains of triangles.

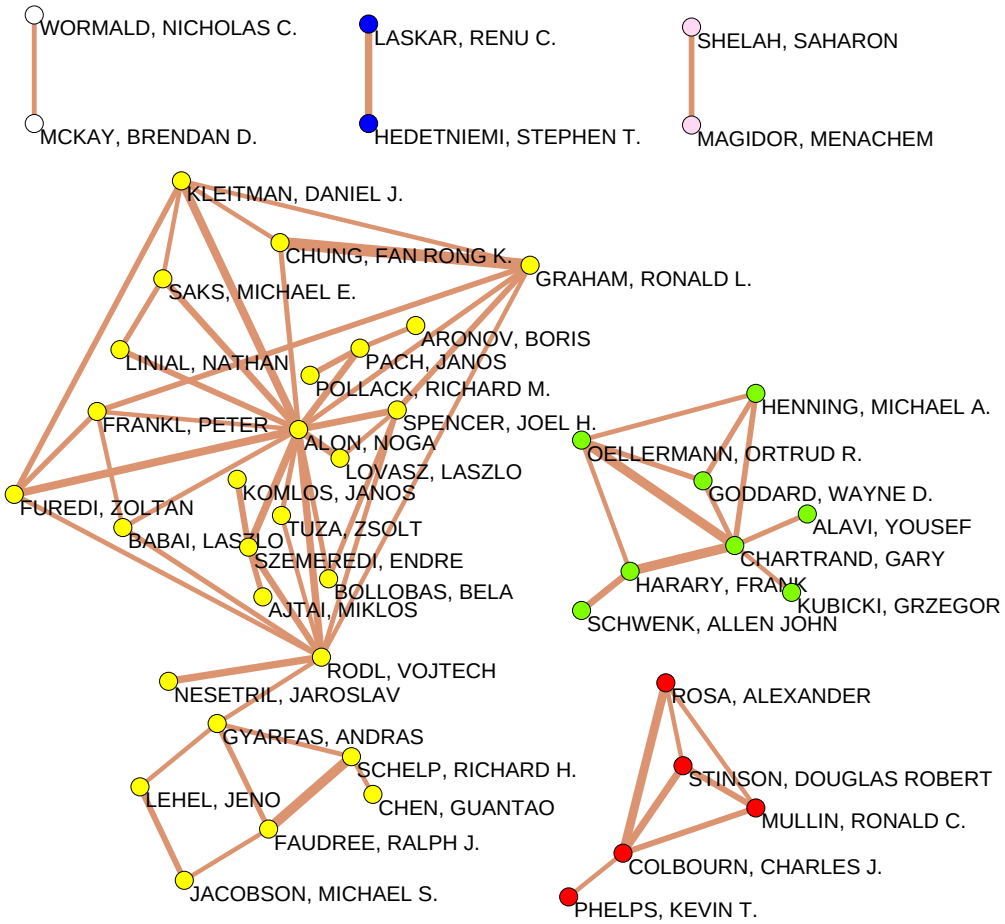
Undirected graphs

We call a *triangle* a subgraph isomorphic to K_3.

Let $\mathbf{G}$ be a simple undirected graph. A *triangular network* $\mathbf{N}_T(\mathbf{G}) = (V, E_T, w)$ determined by $\mathbf{G}$ is a subgraph $\mathbf{G}_T = (V, E_T)$ of $\mathbf{G}$ which set of edges E_T consists of all triangular edges of $E(\mathbf{G})$. For $e \in E_T$ the weight $w(e)$ equals to the number of different triangles in $\mathbf{G}$ to which e belongs.

Triangular networks can be used to efficiently identify dense clique-like parts of a graph. If an edge e belongs to a k-clique in $\mathbf{G}$ then $w(e) \geq k - 2$.

Edge-cut at level 16 of triangular network of Erdős collaboration graph



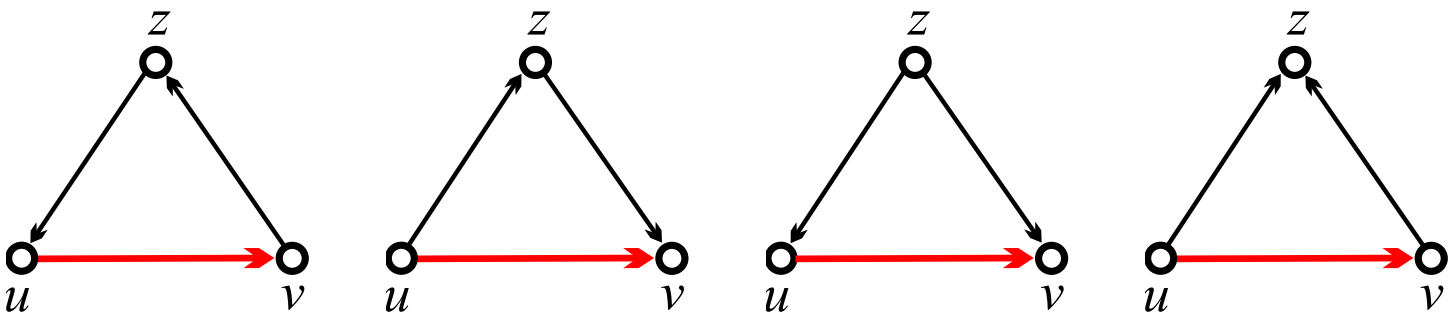
without Erdős,

$n = 6926,$

$m = 11343$

Directed graphs

If the graph $\mathbf{G}$ is mixed we replace edges with pairs of opposite arcs. In the following let $\mathbf{G} = (V, A)$ be a simple directed graph without loops. For a selected arc $(u, v) \in A$ there are four different types of directed triangles: **cyc**lic, **tra**nsitive, **in**put and **out**put.

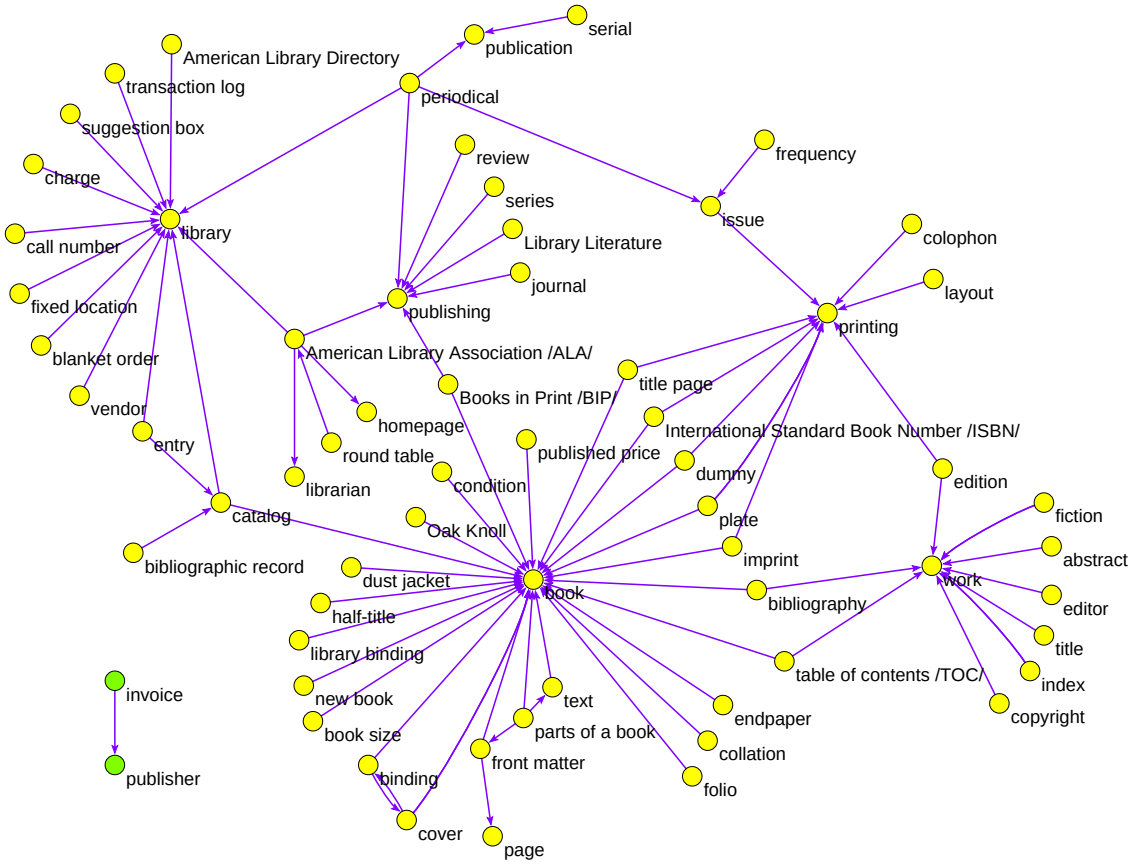


cyc tra in out

For each type we get the corresponding triangular network $\mathbf{N}_{cyc}$, $\mathbf{N}_{tra}$, $\mathbf{N}_{in}$ and $\mathbf{N}_{out}$.

These notions can be generalized to short cycle connectivity.

Line-cut at level 11 of transitive network of ODLIS dictionary graph



Citation networks

In a given set of units $\mathbf{U}$ (articles, books, works, ...) we introduce a *citing* relation $R \subseteq \mathbf{U} \times \mathbf{U}$

$$uRv \equiv v \text{ cites } u$$

which determines a *citation network* $\mathbf{N} = (\mathbf{U}, R)$.

A citing relation is usually *irreflexive* and (almost) *acyclic*.

Citation weights

An approach to the analysis of citation network is to determine for each unit / arc its *importance* or *weight*. These values are used afterward to determine the essential substructures in the network. Some methods of assigning weights $w : R \rightarrow \mathbb{R}_0^+$ to arcs were proposed by Hummon and Doreian (1989):

- *node pair projection count* (NPPC) method: $w_1(u,v) = |R^{\text{inv}\star}(u)| \cdot |R^{\star}(v)|$
- *search path link count* (SPLC) method: $w_2(u,v)$ equals the number of "*all possible search paths through the network emanating from an origin node*" through the arc $(u,v) \in R$.
- *search path node pair* (SPNP) method: $w_3(u,v)$ "*accounts for all connected vertex pairs along the paths through the arc* $(u,v) \in R$".

Citation weights algorithm

To compute the SPLC and SPNP weights we introduce a related *search path count* (SPC) method for which the weights $N(u, v)$, uRv count the number of different paths from $\mathrm{Min}\, R$ to $\mathrm{Max}\, R$ through the arc (u, v).

There exists a very efficient (linear in number of arcs) algorithm to determine the citation weights.

We get the SPLC weights by applying the SPC method on the network obtained from a given standardized (added source s and sink t) network by linking the source s by an arc to each nonminimal vertex from $\mathbf{U}$; and the SPNP weights by applying the SPC method on the network obtained from the SPLC network by additionally linking by an arc each nonmaximal vertex from $\mathbf{U}$ to the sink t.

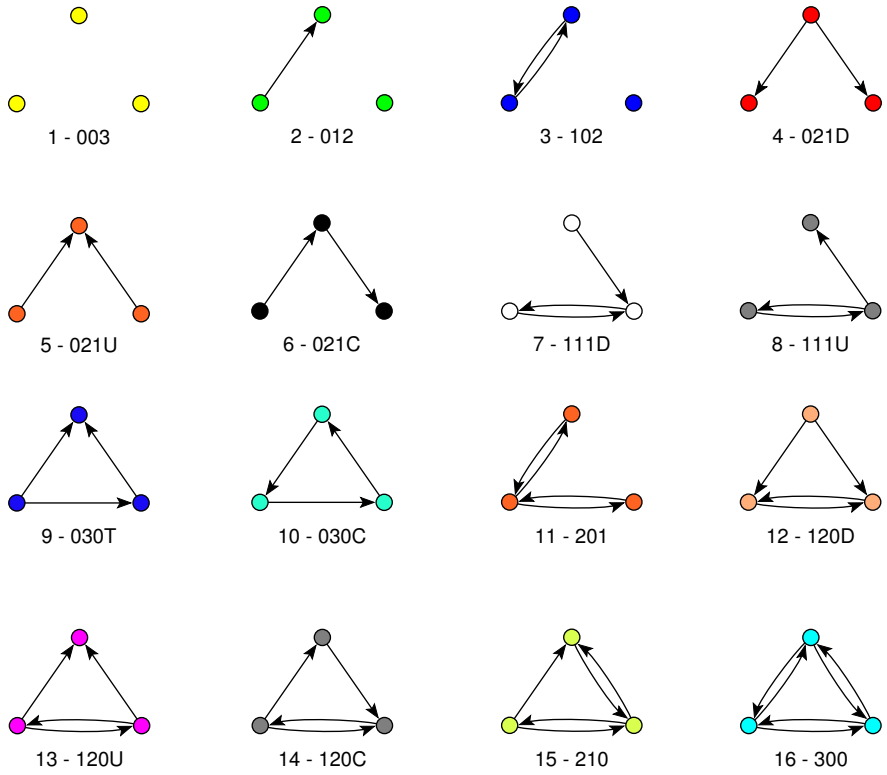
Pattern search

If a selected pattern does not occur frequently in a sparse network the standard backtracking algorithm applied for *pattern search*ing finds all appearences of the pattern very fast even in very large networks.

To speed up the search or to consider some additional properties of the pattern, a user can set some additional options:

- vertices in network should match with vertices in pattern in some nominal, ordinal or numerical property (for example, type of atom in molecula)
- values of lines must match (for example, lines representing male/female links in the case of p-graphs)
- the first vertex in the pattern can be selected only from a given subset of vertices in the network.

Applications of pattern search



Pattern searching was successfully applied to searching for patterns of atoms in molecula (carbon rings) and searching for relinking marriages in genealogies.

For counting triads a special procedure is available.