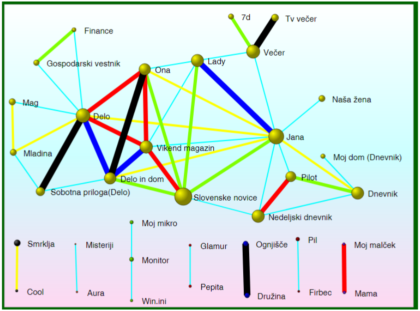
Network weight compatibility normalizations

Vladimir Batagelj
IMFM Ljubljana, IAM UP Koper, and NRU HSE Moscow

1301. Sredin seminar
on Zoom, April 14, 2021

Outline



Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (April 14, 2021 at 17 : 38): slides PDF
https://github.com/bavla/NormNet/blob/main/docs/

Slovenian magazines and journals 1999-2000

Over 100.000 people have been asked which magazines and journals they read (survey conducted in 1999 and 2000, source CATI Center Ljubljana). They listed 124 different magazines and journals. The collected data can be represented as a 2-mode network:

	Delo	Dnevnik	Sl.novice	...
Reader1	X		X	...
Reader2		X		...
Reader3	X			...
....	.....	....	....	...

From the 2-mode network Readers $\times$ Journals we generated an ordinary network on Journals

- an undirected edge with weight counting the number of readers reading both endnode journals;
- a loop on selected journal counts the number of all readers of this journal.

Slovenian magazines and journals 1999-2000

The obtained network matrix:

	Delo	Dnevnik	Sl.novice	...
Delo	20714	3219	4214	...
Dnevnik		15992	3642	...
Sl. novice			31997	...
....	......	.....	.....	...

The second ordinary network on Readers would be huge (more than 100.000 nodes) containing large cliques (readers of a particular journal).

Slovenske novice (31997), Jana (24674), Delo (20714), Večer (18933), Vikend magazin (17122), Nedeljski dnevnik (16596), Lady (16225), Dnevnik (15992), Delo in dom (14606), Ona (14066), . . . Moj pes (53), Grosupeljski odmevi (48), Svet knjige (42), Krog-MK (41), Ljubljana (7).

Reuters Terror news

The Reuters terror news network was produced by Steve Corman and Kevin Dooley at Arizona State University. It is based on all stories released during 66 consecutive days by the news agency Reuters concerning the September 11 attack on the U.S., beginning at 9:00 AM EST 9/11/01. The nodes of a network are words (terms); there is an edge between two words iff they appear in the same text unit (sentence). The weight of an edge is its frequency. The network has n = 13332 nodes (different words in the news) and m = 243447 edges, 50859 with value larger than 1. There are no loops in the network.

The network DaysAll.net contains the main connected component of the network obtained by transforming the Reuters terror news network into a combined network for all 66 days (union of all time points). It has 13308 vertices.

united_states (15000), attack (10348), taliban (6266), people (5286), afghanistan (5176), bin_laden (4885), new_york (4832), pres_bush (4506), washington (4047), official (3902), . . . radioactivity (1), negotiate (1)

2-mode networks

The construction of "Journals" and "Terror news" networks is a special case of the following scheme of analysis of 2-mode networks by transforming them into ordinary (1-mode) networks that are analyzed further using standard network analysis methods.

A *2-mode network* is a structure $\mathcal{N} = (U, V, A, w)$, where U and V are disjoint sets of *nodes*, A is the set of *arcs* (directed links) with the initial node in the set U and the terminal node in the set V, and $w : A \to \mathbb{R}$ is a *weight*. If no weight is defined we can assume a constant weight $w(u, v) = 1$ for all arcs $(u, v) \in A$. The set A can be viewed also as a relation $A \subseteq U \times V$.

For the "Journals" network $U =$ Readers, $V =$ Journals,
$A(u, v) =$ reader u is reading journal v;
and for the "Terror news" network $U =$ News units, $V =$ Words,
$A(u, v) =$ word v appears in news unit u.
Both networks have a constant weight $w = 1$.

Projections

A 2-mode network can be formally represented by a rectangular matrix $\mathbf{UV} = [UV[u,v]]_{U \times V}$

$$UV[u,v] = \begin{cases} w(u,v) & (u,v) \in A \\ 0 & \text{otherwise} \end{cases}$$

It can be transformed to the corresponding one-mode networks by:
column projection: $\mathbf{UU} = \mathrm{col}(\mathbf{UV}) = \mathbf{UV} \cdot \mathbf{UV}^T$
row projection: $\mathbf{VV} = \mathrm{row}(\mathbf{UV}) = \mathbf{UV}^T \cdot \mathbf{UV}$

$$\mathrm{row}(\mathbf{UV}) = \mathbf{UV}^T \cdot \mathbf{UV} = \mathbf{UV}^T \cdot (\mathbf{UV}^T)^T = \mathrm{col}(\mathbf{UV}^\mathbf{T})$$

For a binary network $\mathbf{UV}$, both projection networks are weighted networks with weights $UU[u,t] = |UV(u) \cap UV(t)|$ and $VV[v,z] = |VU(v) \cap VU(z)|$, where $UV(u) = \{v \in V : UV[u,v] = 1\}$ is the set of out-neighbors and $VU(v) = UV^T(v)$.

Projections

Compatibility normalizations

V. Batagelj

$$
\begin{aligned}
UU[u,t] &= \sum_{v \in V} UV[u,v] \cdot UV^T[v,t] = \sum_{v \in V} UV[u,v] \cdot UV[t,v] \\
&= \sum_{v \in V} UV[t,v] \cdot UV^T[v,u] = UU[t,u]
\end{aligned}
$$

$$UU[u,t] = \sum_{v \in UV(u) \cap UV(t)} 1 = |UV(u) \cap UV(t)|$$

$$UU[u,u] = |UV(u)|$$

Derived 1-mode networks

The projections determine the corresponding ordinary (1-mode) networks $\mathrm{col}(\mathcal{N}) = (U, E_c, w_c)$ or/and $\mathrm{row}(\mathcal{N}) = (V, E_r, w_r)$.

The set of edges E_c and the weight w_c are determined by the matrix $\mathbf{UU} = \mathrm{col}(\mathbf{UV})$: there is an *edge* $(u : t) \in E_c$, $(u : t) = (t : u)$ iff $UV(u) \cap UV(t) \neq \emptyset$. Its weight is $w_c(u : t) = UU[u, t]$.

The set of edges E_r and the weight w_r are determined similarly by the matrix $\mathbf{VV} = \mathrm{row}(\mathbf{UV})$.

Outer-product decomposition

For vectors $x = [x_1, x_2, \ldots, x_n]$ and $y = [y_1, y_2, \ldots, y_m]$ their *outer product* $x \circ y$ is defined as a matrix

$$x \circ y = [x_i \cdot y_j]_{n \times m}$$

Let $S_x = \sum_i x_i$ and $S_y = \sum_j y_j$ then the *contribution* of the outer product $x \circ y$ is equal

$$T(x \circ y) = \sum_{i,j} (x \circ y)_{ij} = \sum_i \sum_j x_i \cdot y_j = \sum_i x_i \cdot \sum_j y_j = S_x \cdot S_y$$

We can express the product of two compatible matrices **A** and **B** as the *outer product decomposition*

$$\mathbf{C} = \mathbf{A} \cdot \mathbf{B} = \sum_k \mathbf{H}_k \quad \text{where} \quad \mathbf{H}_k = \mathbf{A}[\cdot, k] \circ \mathbf{B}[k, \cdot]$$

$$T(\mathbf{C}) = T(\sum_k \mathbf{H}_k) = \sum_k T(\mathbf{H}_k)$$

Outer-product decomposition

In the case of **UU** we get for a binary network

$$\mathbf{H}_v = \mathbf{UV}[\cdot, v] \circ \mathbf{UV}^T[v, \cdot] = \mathbf{UV}[\cdot, v] \circ \mathbf{UV}[\cdot, v] = \mathbf{K}_{VU(v)}$$

$$S(\mathbf{UV}[., v]) = \sum_{u \in U} UV[u, v] = \text{indeg}_{UV}(v)$$

$$T(\mathbf{H}_v) = \text{indeg}_{UV}(v)^2$$

In words

- a derived network is a sum of complete subgraphs;
- the contribution to the total of a node $v \in V$ from the other set is $\text{indeg}_{UV}(v)^2$.

This means that the nodes with large indegree in V are over-represented in the projection.

Outer-product decomposition

For example: $U =$ Works, $V =$ Authors, $A(u, v) = v$ is co-author of work u.
The works with many co-authors contribute large complete subgraphs to the “collaboration” matrix **VV**.

$VV[v, z] =$ number of works that authors v and z co-authored.

This measure is often used in evaluations of research productivity.

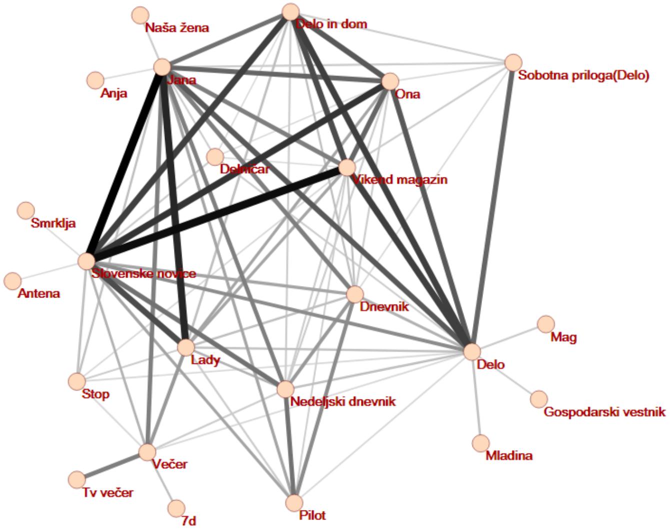
Analysis of weighted networks

Assume that the weight w in the network $\mathcal{N} = (V, L, w)$ is "compatible" with our research question:

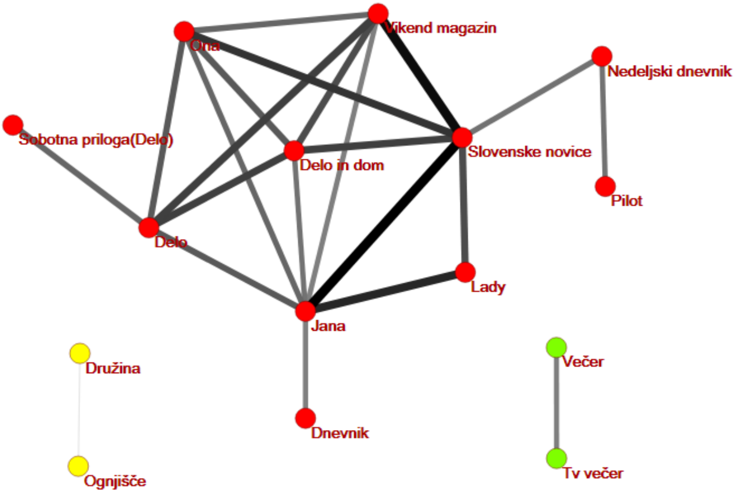
larger is the value of the weight $w(e) \Rightarrow$
more important is the link $e \in L$ with respect to our question

To identify important nodes or subnetworks we ca use: cuts, islands, (valued) cores, clustering (corrected dissimilarities), etc.

Edge-cut of Journals at level 1500



Simple islands in Journals



Pathfinder skeleton of Journals

Normalized weighted networks

Till now, we were interested in identifying and displaying the most important parts of a weighted network.

An alternative approach is based on *compatibility normalization* of the weights. Because of the huge differences in frequencies of different words it is not possible to compare values on edges according to the raw data. First we have to normalize the network to make the weights comparable. There exist several ways how to do this. Some of them are presented in the following table.

The normalization approach was developed for quick inspection of (1-mode) networks derived from 2-mode networks.

Weight normalizations

$$\mathrm{Geo}_{uv} = \frac{w_{uv}}{\sqrt{w_{uu}w_{vv}}} \qquad \mathrm{GeoDeg}_{uv} = \frac{w_{uv}}{\sqrt{\deg\ u \cdot \deg\ v}}$$

$$\mathrm{Input}_{uv} = \frac{w_{uv}}{w_{vv}} \qquad \mathrm{Output}_{uv} = \frac{w_{uv}}{w_{uu}}$$

$$\mathrm{Min}_{uv} = \frac{w_{uv}}{\min(w_{uu}, w_{vv})} \qquad \mathrm{Max}_{uv} = \frac{w_{uv}}{\max(w_{uu}, w_{vv})}$$

$$\mathrm{MinDir}_{uv} = \begin{cases} \frac{w_{uv}}{w_{uu}} & w_{uu} \leq w_{vv} \\ 0 & \text{otherwise} \end{cases} \qquad \mathrm{MaxDir}_{uv} = \begin{cases} \frac{w_{uv}}{w_{vv}} & w_{uu} \leq w_{vv} \\ 0 & \text{otherwise} \end{cases}$$

… Weight normalizations

In the case of networks without loops we define the diagonal weights for undirected networks as the sum of out-diagonal elements in the row (or column)

$$w_{vv} = \sum_u w_{vu}$$

and for directed networks (for example, trade among world countries) as some mean value of the row and column sum, for example

$$w_{vv} = \frac{1}{2}(\sum_u w_{vu} + \sum_u w_{uv})$$

or

$$w_{vv} = \sqrt{\sum_u w_{vu} \cdot \sum_u w_{uv}}$$

Usually we assume that the network does not contain any isolated node.

GeoDeg normalization

Using the Geo normalization we divide elements of the matrix by geometric mean of both diagonal elements. The standard Geo normalization attains its maximal values on components consisting of a single edge, and is high for strongly 'correlated' nodes – in most of their appearences they appear together. Its application reveals very specific themes.

The display of an edge cut was obtained manually.

Nodes with large degree have little chance to appear as endnodes of edges with large normalized weight. To give a chance also to these nodes we decided to use the GeoDeg variant of Geo normalization in which the diagonal is filled with degrees of nodes.

MinDir normalization

The MinDir normalization transforms an undirected network into a directed one – an arc points from the word with a lower activity to the word with a higher activity; the weight on the arc corresponds to the proportion of activity on the arc to the total activity of the first (initial) node.

In the case of Journals this normalization measures the dependance between reading preferences. Large value (close to 1) of the MinDir weight implies that the readers of the initial journals are almost all reading also the second journal.

Also in this case the final layout was obtained manually.

Fractional approach

To make in a projection the contributions of all nodes from the second set equal we can apply the *fractional approach* by normalizing the weights: setting $x' = x/S_x$ and $y' = y/S_y$ we get $S_{x'} = S_{y'} = 1$ and therefore (if $S_v > 0$) $T(\mathbf{H}_v) = 1$ for all nodes $v \in V$.

Let $U^+ = \{u \in U : \text{outdeg}(u) > 0\}$.

The normalized 2-mode network $n(\mathbf{UV})$ has weights

$$n(UV)[u,v] = \begin{cases} \dfrac{UV[u,v]}{\text{outdeg}_{UV}(u)} & u \in U^+_{UV} \\ 0 & u \notin U^+_{UV} \end{cases}$$

$$T(n(\mathbf{UV})) = \sum_{u \in U^+} \sum_{v \in V} \frac{UV[u,v]}{\text{outdeg}_{UV}(u)} = \sum_{u \in U^+} 1 = |U^+|$$

Normalized projection

$$\mathbf{VV} = row(n(\mathbf{UV}))$$

Discuss meaningfulness – citation networks.

Newman's normalization

In bibliometric applications on network **WA** the derived network **Co** $=$ row(**WA**) describes the co-authorship (collaboration) among authors. The network **Cn** $=$ row(n(**WA**)) is the corresponding fractional version.

Mark Newman proposed an alternative normalization that considers only co-authorship between different authors – single author works and self co-authorship are excluded.

The Newman's normalized 2-mode network $n'(\mathbf{UV})$ has weights

$$n'(UV)[u,v] = \begin{cases} \dfrac{UV[u,v]}{\text{outdeg}_{UV}(u) - 1} & \text{outdeg}_{UV}(u) > 1 \\ 0 & \text{otherwise} \end{cases}$$

Newman's projection

$$\mathbf{Cn}' = (n(\mathbf{WA}))^T \cdot n'(\mathbf{WA})$$

Network: symmetrize with sum and remove loops.

2-mode networks

For a binary network **UV** let $l(\mathbf{UV}) = n(\mathbf{UV}) \cdot \mathbf{UV}^T$ (*left fractional contribution*)

$$l(UV)[u,t] = \frac{1}{\text{outdeg}(u)} \sum_{v \in V} UV[u,v] \cdot UV[t,v] = \frac{|UV(u) \cap UV(t)|}{|UV(u)|} \leq 1$$

and $r(\mathbf{UV}) = \mathbf{UV} \cdot n(\mathbf{UV})^T$ (*right fractional contribution*)

$$\begin{aligned} r(UV)[u,t] &= \frac{1}{\text{outdeg}(t)} \sum_{v \in V} UV[u,v] \cdot UV[t,v] \\ &= \frac{|UV(t) \cap UV(u)|}{|UV(t)|} = l(UV)[t,u] \end{aligned}$$

In general $l(UV)[u,t] \neq r(UV)[u,t]$. A symmetric measure $biCnX[u,t]$ can be constructed as some $meanX$ of these two quantities

$$\begin{aligned} biCnX[u,t] &= meanX(l(UV)[u,t], r(UV)[u,t]) \\ &= meanX(l(UV)[u,t], l(UV)[t,u]) \end{aligned}$$

2-mode networks

$$biCnA[u,t] = \frac{1}{2}(I(UV)[u,t] + I(UV)[t,u])$$ – arithmetic mean

$$biCnm[u,t] = \min(I(UV)[u,t], I(UV)[t,u])$$ – minimum

$$biCnM[u,t] = \max(I(UV)[u,t], I(UV)[t,u])$$ – maximum

$$biCnG[u,t] = \sqrt{I(UV)[u,t] \cdot I(UV)[t,u]}$$ – geometric mean, Salton

$$biCnH[u,t] = 2(I(UV)[u,t]^{-1} + I(UV)[t,u]^{-1})^{-1}$$ – harmonic mean, Dice

$$biCnJ[u,t] = (I(UV)[u,t]^{-1} + I(UV)[t,u]^{-1} - 1)^{-1}$$ – Jaccard

$$biCnX[u,t] = biCnX[t,u]$$

$$biCnX[u,t] \in [0,1]$$

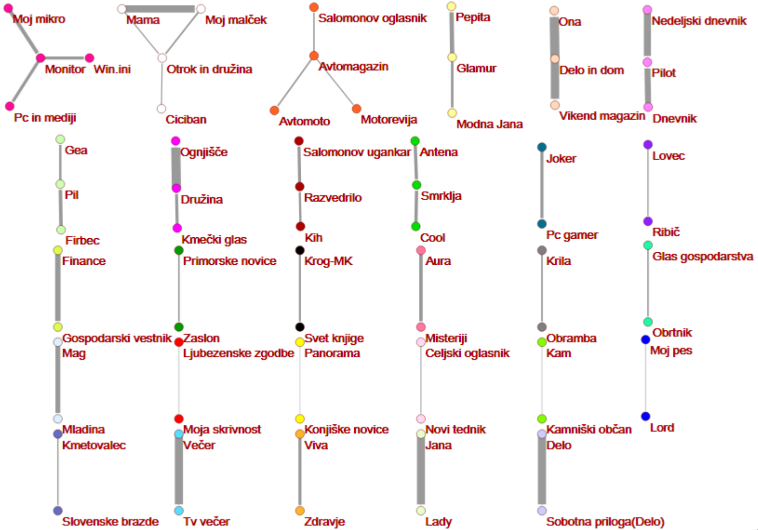
$$biCnJ[u,t] \leq biCnm[u,t] \leq biCnH[u,t] \leq biCnG[u,t] \leq biCnA[u,t] \leq biCnM[u,t]$$

2-mode networks

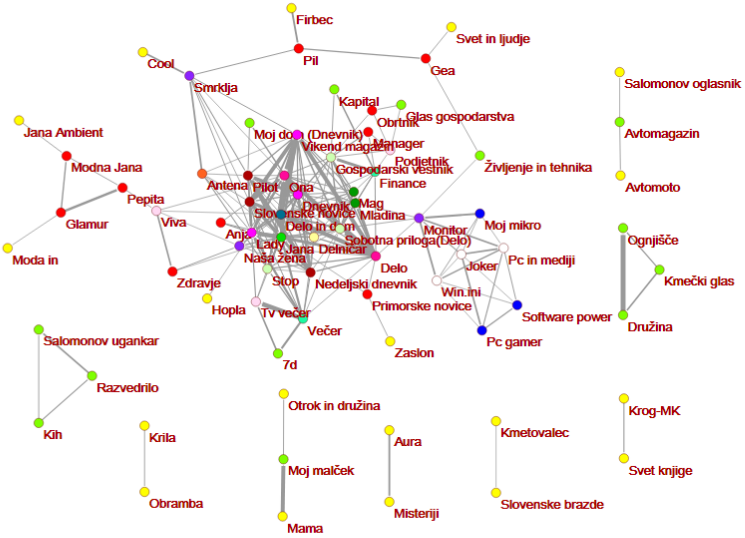
$$\begin{aligned}
biCnA[u,t] &= \frac{|UV(u)\cap UV(t)|}{2}\left(\frac{1}{|UV(u)|}+\frac{1}{|UV(t)|}\right)\\
biCnm[u,t] &= \frac{|UV(u)\cap UV(t)|}{\max(|UV(u)|,|UV(t)|)}\\
biCnM[u,t] &= \frac{|UV(u)\cap UV(t)|}{\min(|UV(u)|,|UV(t)|)}\\
biCnG[u,t] &= \frac{|UV(u)\cap UV(t)|}{\sqrt{|UV(u)|\cdot|UV(t)|}}\\
biCnH[u,t] &= \frac{2|UV(u)\cap UV(t)|}{|UV(u)|+|UV(t)|}\\
biCnJ[u,t] &= \frac{|UV(u)\cap UV(t)|}{|UV(u)\cup UV(t)|}
\end{aligned}$$

$|UV(u)\cup UV(t)| = |UV(u)| + |UV(t)| - |UV(u)\cap UV(t)|$
$|UV(u)\cap UV(t)| = UU[u,t]$ and $|UV(u)| = UU[u,u]$

Jaccard – simple islands



Jaccard – cut at level 0.04



Conclusions

- use a normalization that is expressing your research question
- generalization to weighted 2-mode networks

References I

- Pajek datasets: Journals / Slovenian magazines and journals 1999 and 2000. WWW
- Batagelj, V.: Pajek / Example – Slovenian magazines and journals. Dagstuhl seminar, 2001. WWW
- Batagelj, V.: Analiza velikih omrežij. 88. Solomonov seminar, IJS, 10. september 2002. PDF
- Batagelj, V., Mrvar, A.: Density based approaches to network analysis: Analysis of Reuters terror news network. LinkKDD 2003 PDF
- Batagelj, V, Cerinšek, M.: On bibliographic networks. Scientometrics 96 (2013), pages 845–864. Springer
- Batagelj, V., Doreian, P., Ferligoj, A., Kejžar, N.: Understanding Large Temporal Networks and Spatial Networks. Wiley 2014.
- Batagelj, V.: On fractional approach to analysis of linked networks. Scientometrics 123 (2020), pages621–633. Springer
- Pajek datasets: days / Reuters terror news network. WWW

References II

- Batagelj, V., Maltseva, D.: Temporal bibliographic networks. Journal of Informetrics, 14 (2020) 1, 101006.
- Corman, S.R., Kuhn, T., McPhee, R.D., Dooley, K.J.: Studying complex discursive systems: Centering resonance analysis of communication. Human Communication Research, 28 (2002) 2, 157–206.
- Zheng, Q., Skillicorn, D.: Social Networks with Rich Edge Semantics. CRC 2017.

Link-cuts

The *link-cut* of network $\mathcal{N} = (V, L, w)$ at selected level t

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

is a subnetwork $\mathcal{N}(t) = (V(L'), L', w)$, $V(L')$ is the set of all endnodes of the links from L'. The components of $\mathcal{N}(t)$ – *islands*, determine different themes. Their number and sizes depend on t. Usually there are many small components.
To obtain interesting themes 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.
The link-cut approach is closely related to single-linkage (minimal spanning tree) clustering method. Therefore we can expect the *chaining effect* in some results – chaining of themes with common characteristic words.

Node-cuts

In some networks we can have also a function $p : V \to \mathbb{R}$ that describes some property of vertices. Its values can be obtained by measuring, or they are computed (for example, centrality indices, clustering index, ...).
The *node-cut* of a network $\mathcal{N} = (V, L, p)$ at selected level t is a network $\mathcal{N}(t) = (V', L(V'), w)$, determined by the set

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

and $L(V')$ is the set of links from L that have both endnodes in V'.