Network weight compatibility normalizations

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, and NRU HSE Moscow

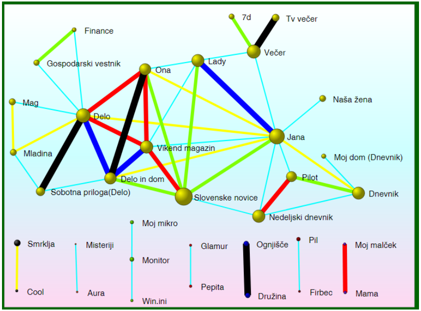
1301. Sredin seminar

on Zoom, April 14, 2021

Outline

1 2-model networks

2 References



Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (April 11, 2021 at 21 : 58): slides PDF
https://github.com/bavla/TQ/tree/master/docs

Analysis of weighted networks

properties / weights
higher value - more important
cuts
islands
clustering (corrected dissimilarities)

A similar approach can be used for a selected node property $p \in P$ in a network $N = (V, L, P, W)$ asumming that p is reflecting the importance of nodes with respect to our question – node-cuts, node-islands.

Slovenian magazines and journals 1999-2000

distribution of values (diagonals)

Reuters Terror news

diagonals

2-mode networks

The construction of news networks is a special case of the following scheme, a kind of *network based data mining*, 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 *vertices*, A is the set of *arcs* (directed links) with the initial vertex in the set U and the terminal vertex 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 $A \subseteq U \times V$.

A 2-mode network can be formally represented by rectangular matrix $\mathbf{A} = [a_{uv}]_{U \times V}$.

$$a_{uv} = \begin{cases} w_{uv} & (u, v) \in A \\ 0 & \text{otherwise} \end{cases}$$

Derived 1-mode networks

We denote by $N(u) = \{v \in V : (u,v) \in A\}$ the *set of neighbors* of vertex $u \in U$; similary, $N(v) = \{u \in U : (u,v) \in A\}$ is the set of neighbors of vertex $v \in V$.

An approach to analyze a 2-mode network is to transform it into an ordinary (1-mode) network $\mathcal{N}_1 = (U, E_1, w_1)$ or/and $\mathcal{N}_2 = (V, E_2, w_2)$, where E_1 and w_1 are determined by the matrix $\mathbf{A}^{(1)} = \mathbf{A}\mathbf{A}^T$, $a^{(1)}_{uv} = \sum_{z\in V} a_{uz} \cdot a^T_{zv}$. Evidently $a^{(1)}_{uv} = a^{(1)}_{vu}$. There is an *edge* $(u : v) \in E_1$, $(u : v) = (v : u)$, in $\mathcal{N}_1$ iff $N(u) \cap N(v) \neq \emptyset$. Its weight is $w_1(u : v) = a^{(1)}_{uv}$.

The network $\mathcal{N}_2$ is determined in a similar way by the matrix $\mathbf{A}^{(2)} = \mathbf{A}^T\mathbf{A}$.

The networks $\mathcal{N}_1$ and/or $\mathcal{N}_2$ can be analyzed separately using standard techniques.

News networks

Also in our case there is a 2-mode network in the background. The construction of CRA networks can be viewed as a transformation of a 2-mode network (units of text, words, contains) into the corresponding 1-mode *news network* (words, co-apperance, frequency).

Edge-cuts

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

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

is a subnetwork $\mathcal{N}(t) = (V(E'), E', w)$, $V(E')$ is the set of all endpoints of the edges from E'. 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 edge-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.

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

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

and $E(V')$ is the set of edges from E that have both endpoints in V'.

Analysis of the combined network

We first combined all 66 CRA networks into a single Pajek's temporal network stored on the file `Days.net`. It has $n = 13332$ vertices (different words in the news) and $m = 243447$ edges, 50859 with value larger than 1. There are no loops in the network.

It consists of 22 components – one large of size 13308, 3 of size 2, and 18 isolated vertices. We continue the analysis on the large component, saved as `DaysAll.net`, and the corresponding temporal subnetwork, saved as `DaysCom.net`.

To analyze the combined network we shall use the valued cores. They identify the dense parts of a network.

Valued cores

Let $\mathcal{N} = (V, E, w)$ be a network. For $v \in V$ and $C \subseteq V$ we define the vertex value p

$$p(v; C) = \sum_{u \in N(v) \cap C} w(v, u)$$

where $w(v, u)$ is the frequency of edge (v, u).
The *t-core* of the network is the maximum subset C such that for all $v \in C$ it holds $p(v; C) \geq t$.
In the *Terror news* network the weight w is the frequency of co-appearance of given two words (endpoints of the edge). In this case a t-core is the maximum subnetwork in which each its vertex co-appeared in the text at least t times with other vertices from the subnetwork.
There exists a very efficient algorithm to determine t-cores which is also implemented in Pajek. Using it we produce the valued cores partition with step 25.

Valued cores distribution

k	num	interval	k	num	interval
0	0	0 or less	13	36	(300-325]
1	10598	(0- 25]	14	18	(325-350]
2	1081	(25- 50]	15	39	(350-375]
3	479	(50- 75]	16	27	(375-400]
4	300	(75-100]	17	4	(400-425]
5	152	(100-125]	18	14	(425-450]
6	130	(125-150]	19	9	(450-475]
7	159	(150-175]	21	4	(500-525]
8	77	(175-200]	22	2	(525-550]
9	42	(200-225]	26	4	(625-650]
10	56	(225-250]	27	4	(650-675]
11	28	(250-275]	28	2	(675-700]
12	37	(275-300]	35	6	(850-875]

Vertex-cut

On the basis of distribution of the obtained partition we decide to look at the vertex-cut at level $t = 500$. At this level there is the last gap $I_{20} = (475, 500]$ in the distribution, and also the number of words in the t-core $(= 22)$ is still small.

Afterward we draw the 500-core, representing the weights by the width of edges. An initial layout is obtained using the Kamada-Kawai procedure, and further improved manually. We export the final picture into SVG (Scalable Vector Graphics). The obtained SVG picture can be viewed in a web browser and allows the user to interactively select display of cores at different (predefined) levels.

Normalized valued networks

Till now, we were interested in identifying and displaying the most important (dense) parts of the network. But what about the other parts? One approach would be to discard the main part from the network and analyze the residuum.
In the continuation we shall present an alternative approach 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.

Weight normalizations

$$
\begin{aligned}
\mathrm{Geo}_{uv} &= \frac{w_{uv}}{\sqrt{w_{uu} w_{vv}}} & \mathrm{GeoDeg}_{uv} &= \frac{w_{uv}}{\sqrt{\deg\ u \cdot \deg\ v}} \\
\mathrm{Input}_{uv} &= \frac{w_{uv}}{w_{vv}} & \mathrm{Output}_{uv} &= \frac{w_{uv}}{w_{uu}} \\
\mathrm{Min}_{uv} &= \frac{w_{uv}}{\min(w_{uu}, w_{vv})} & \mathrm{Max}_{uv} &= \frac{w_{uv}}{\max(w_{uu}, w_{vv})} \\
\mathrm{MinDir}_{uv} &= \begin{cases} \frac{w_{uv}}{w_{uu}} & w_{uu} \leq w_{vv} \\ 0 & \text{otherwise} \end{cases} & \mathrm{MaxDir}_{uv} &= \begin{cases} \frac{w_{uv}}{w_{vv}} & w_{uu} \leq w_{vv} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

. . . 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 as some mean value of the row and column sum, for example

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

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

The normalization approach was developed for quick inspection of (1-mode) networks derived from 2-mode networks. It was the first time successfully applied in the analysis of the 2-mode network (readers, journals, is reading), $|\text{readers}| > 100000$, $|\text{journals}| = 124$ obtained from the readership survey in Slovenia, conducted in 1999 and 2000 by the CATI Center Ljubljana

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' vertices – in most of their appearences they appear together. Its application reveals very specific themes. Vertices with large degree have little chance to appear as endpoints of edges with large normalized weight. To give a chance also to these vertices we decided to use the GeoDeg variant of Geo normalization in which the diagonal is filled with degrees of vertices.

Inspecting the distribution of normalized line values we determine the threshold $t = 0.25$ and cut the network at this level. We are interested only in themes (connected components) of size at least 6. We get a network on 641 vertices with 62 components (themes). We draw them. The final layout was obtained manually.

MaxDir normalization

The MaxDir normalization transforms an undirected network into a directed one – an arc points from the word with lower frequency to the word with higher frequency; the value on the arc corresponds to the percentage of messages containing the second (terminal) word, that contain also the first (initial) one.

In our case this normalization measures the dependance between words. Large value (close to 1) of the MaxDir weight implies that the two words connected by the arc mainly co-appear.

After the MaxDir normalization the themes network is cut at level 0.1 and each its component contains at least 6 vertices. We get a network on 502 vertices with 64 components (themes). Also in this case the final layout was obtained manually.

2-mode networks

2-mode networks

2-mode networks

2-mode networks

2-mode networks

2-mode networks

2-mode networks

2-mode networks

References I

- Pajek datasets: Journals / Slovenian magazines and journals 1999 and 2000. WWW
- Batagelj, V.: 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
- Pajek datasets: days / Reuters terror news network. WWW
- Batagelj, V., Maltseva, D. (2020) Temporal bibliographic networks. Journal of Informetrics, Volume 14, Issue 1, 101006.
- Corman, S.R., Kuhn, T., McPhee, R.D., Dooley, K.J. (2002) Studying complex discursive systems: Centering resonance analysis of communication. Human Communication Research, 28(2), 157-206.