Network Analysis of Reuters News about the Terrorist Attack on September 11, 2001

Vladimir Batagelj, Andrej Mrvar
University of Ljubljana

Methodology and Statistics 2002

16 - 18 September 2002, Ljubljana, Slovenia

Outline

- Reuters terror news network
- 2-mode networks to 1-mode networks
- Valued cores – main themes
- Normalized valued networks – secondary themes

Reuters terror news network

Centering Resonance Analysis (CRA) is a new text analysis technique developed by Steve Corman and Kevin Dooley at Arizona State University. For demonstration of CRA they produced and analyzed several networks. Among them also the ***Reuters terror news network*** that 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.

This network was selected by *Viszards* (network visualization group) for the presentation at the visualization session on the Sunbelt XXII International Sunbelt Social Network Conference, New Orleans, USA, 13-17. February 2002.

We first combined all 66 CRA networks into a single Pajek's temporal network **`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.

Temporal networks in Pajek

```
*vertices  13332
 ...
 6842 "label"        [3,31,45,56,57,61]
 6843 "labor"        [2,6,20,25,31,43,53]
 6844 "laboratory"   [2,4,5,15,18,24,28,30,31,37,39-41,45-49,51-53,55,56,60]
 6845 "labour"       [21,22]
 6846 "lace"         [17,59]
 ...
10574 "science"      [2,4,5,14-16,37,40-43,47,53,60]
10575 "scientific"   [15,49,53,60]
10576 "scientist"    [2,6,11,18,21,30,37,41,47,48,55-57,60,63]
10577 "scoff"        [13]
10578 "scope"        [4-7,10,15,17,28,53]
10579 "scorch"       [59]
10580 "score"        [1-3,6,8,9,13,15,26,32,33,35,50,54]
10581 "scoreboard"   [29,50]
 ...
*edges
 ...
 8545   9227   4 [60]
 9227  10935   2 [60]
 1076  11885   2 [60]
 6844  10575   3 [60]
  417   6844   1 [60]
 9288  11741   2 [60]
 ...
```

2-mode networks

A kind of *network based data mining* ...

A *2-mode network* is a structure (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 $w : A \rightarrow \mathbb{R}$ is a weight. If no weight is defined we can assume a constant weight $w(u, v) = 1$ for all arcs $(u, v) \in A$. A 2-mode network can be formally represented by rectangular matrix $\mathbf{A} = [a_{uv}]_{U \times V}$.

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

An approach to analyze a 2-mode network is to transform it into an ordinary (1-mode) network $(U, \mathbf{A}\mathbf{A}^T, w_1)$ or/and $(V, \mathbf{A}^T\mathbf{A}, w_2)$, which can be analyzed separately using standard techniques.

From 2-mode networks over large sets U and/or V we can get smaller ordinary networks by first partitioning U and/or V and shrinking the clusters.

2-mode networks ...

There are also different ways to define matrix multiplication – we can select different semirings. Instead of usual operations $(+, \cdot)$ we can also use $(\vee, \wedge)$ for $w = 1$, (max, min), (min, $+$) ... In this way we get different ordinary 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 network (words, co-apperance, frequency).

Valued cores

Used to identify dense parts of the network.

Let (V, E, w) be a network, V is the set of vertices, E is the set of edges, and $w : E \rightarrow \mathbb{R}$ the weight of edges. 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), and $N(v)$ is the neighbourhood (set of neighbors) of v.

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 $O(m \log n)$ algorithm to determine t-cores which is also implemented in **`Pajek`**.

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 *normalization* of the weights. Because of the huge differences in frequencies of different words it is very hard to compare values on edges according to the raw data. First we have to normalize the network. There exist several ways how to do this. Some of them are presented in the following table.

Normalizations

$$\text{Geo}_{uv} = \frac{w_{uv}}{\sqrt{w_{uu} w_{vv}}} \qquad \text{GeoDeg}_{uv} = \frac{w_{uv}}{\sqrt{\deg_u \deg_v}}$$

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

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

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

Normalization ...

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)

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

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.

Normalization ...

If we cut a normalized network at selected level t

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

we get a subnetwork $(V(E'), E', w')$ that contains a set of components – *islands*, determining different themes. Their number and sizes depend on t. Usually there are many small components. To obtain interesting themes we extract only components of size at least k. The values of thresholds t and k are determined by inspecting the distribution of normalized weights and the distribution of component sizes,

The normalization approach 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.

We demonstrate the use of GeoDeg, Maxdir and MinDir normalizations.

Addresses

http://locks.asu.edu/terror/

http://vlado.fmf.uni-lj.si/pub/networks/doc/Terror/core500.htm

http://vlado.fmf.uni-lj.si/pub/networks/doc/Terror/deg50/deg50001.htm

http://vlado.fmf.uni-lj.si/pub/networks/doc/Terror/GeoDeg.htm

http://vlado.fmf.uni-lj.si/pub/networks/doc/Terror/MaxDir.htm

http://vlado.fmf.uni-lj.si/pub/networks/doc/Terror/MinDir.htm

http://vlado.fmf.uni-lj.si/pub/networks/doc/Seminar/Krebs.pdf

http://vlado.fmf.uni-lj.si/pub/networks/pajek/

vladimir.batagelj@uni-lj.si