Network weight compatibility normalizations 2

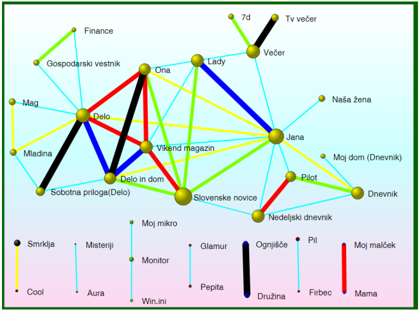
Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, and NRU HSE Moscow

1311. and 1312. Sredin seminar

on Zoom, January 19 and 26, 2022

Outline



Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

Current version of slides (January 26, 2022 at 05 : 24): slides PDF

https://github.com/bavla/NormNet/blob/main/docs/

Coauthorship between post-Soviet countries

Nataliya Matveeva collected from WoS a data set on coauthorships between researchers from post-Soviet countries for the years 1993, 1998, 2003, 2008, 2013, and 2018. Here is the matrix for the year 2018:

	Aze	Arm	Bel	Est	Geo	Kaz	Kyr	Lat	Lit	Mol	Rus	Tad	Tur	Ukr	Uzb
Azerbaijan	255	148	123	4	121	13	3	3	2	7	216	1	0	47	3
Armenia	148	245	256	157	274	12	11	146	146	5	375	1	0	199	26
Belarus	123	256	269	167	269	19	1	158	191	10	642	2	0	211	25
Estonia	4	157	167	587	163	16	13	255	238	6	290	0	0	188	23
Georgia	121	274	269	163	114	12	7	150	151	10	302	2	0	175	29
Kazakhstan	13	12	19	16	12	357	30	5	10	3	277	5	1	41	15
Kyrgyzstan	3	11	1	13	7	30	17	3	6	2	40	3	1	13	6
Latvia	3	146	158	255	150	5	3	242	342	4	302	0	0	289	23
Lithuania	2	146	191	238	151	10	6	342	1143	5	329	0	0	328	23
Moldova	7	5	10	6	10	3	2	4	5	37	44	2	0	11	3
Russia	216	375	642	290	302	277	40	302	329	44	23928	21	0	867	65
Tadjikistan	1	1	2	0	2	5	3	0	0	2	21	13	0	2	4
Turkmenistan	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1
Ukraina	47	199	211	188	175	41	13	289	328	11	867	2	0	1649	32
Uzbekistan	3	26	25	23	29	15	6	23	23	3	65	4	1	32	72

The matrix contains also the data about coauthorships inside each country. For example, in 2018 for Russia there were 23928 internal coauthorships and only 3770 coauthorships with other countries; compare this with Belorus, inter = 269, other = 2074. This makes Russia a big outlier.

Coauthorship between post-Soviet countries

Because we are interested in strict coauthorship we decided to remove the diagonal values / loops. We denote the corresponding matrix $\mathbf{C}$. Nonexisting links are represented with the value 0. The matrix $\mathbf{C}$ is symmetric.

The weights in inter-country coauthorship networks depend on the size of the countries and other factors. They span a wide range of values. To compare the countries we have to apply some normalization.

"History"

- Pat – citations between journals; bistochastic normalization; Preddvor 199? [15]
- Kropivnik: Slovenian political parties 1994 PDF
- Slovenian journals 2000
- Reuters Terror News 2001 PDF
- bibliographic networks 2013 arXiv
- post-Soviet 2021

Slovenian political parties 1994 / reordered

		1	3	6	8	9	2	4	5	7	10
SKD	1	0	114	94	176	117	-215	-89	-77	-170	-210
SDSS	3	114	0	138	177	180	-217	-203	-80	-109	-174
ZS	6	94	138	0	140	116	-150	-142	-188	-97	-106
SLS	8	176	177	140	0	235	-253	-241	-120	-184	-132
SPS-SNS	9	117	180	116	235	0	-230	-254	-160	-191	-164
ZLSD	2	-215	-217	-150	-253	-230	0	134	77	57	49
LDS	4	-89	-203	-142	-241	-254	134	0	157	173	23
ZSESS	5	-77	-80	-188	-120	-160	77	157	0	170	-9
DS	7	-170	-109	-97	-184	-191	57	173	170	0	-6
SNS	10	-210	-174	-106	-132	-164	49	23	-9	-6	0

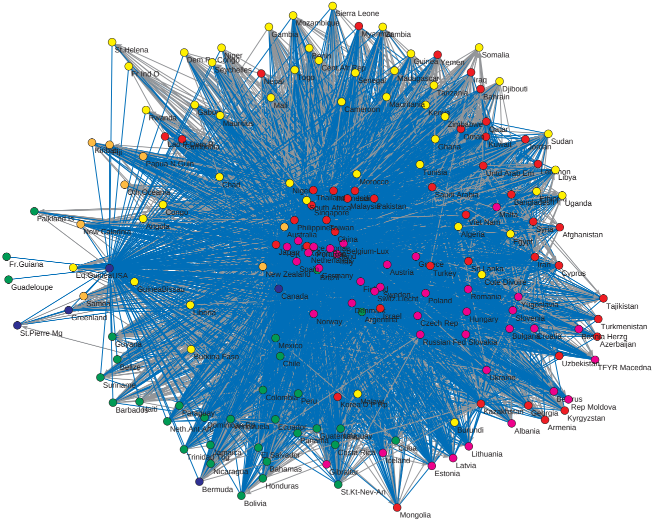
S. Kropivnik, A. Mrvar: An Analysis of the Slovene Parliamentary Parties Network. in Developments in data analysis, MZ 12, FDV, Ljubljana, 1996, p. 209-216.

Weighted network

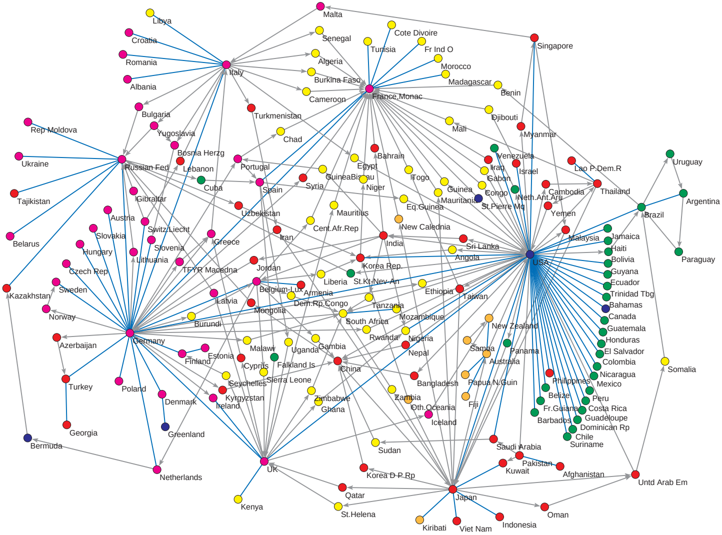
A network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$, $w : \mathcal{L} \to \mathbb{R}$.

Using link cuts, islands, cores, skeletons (spanning tree, Pathfinder, k-neighbors), community detection, hubs and authorities, etc. we can identify the most active subnetworks. We can also apply clustering and blockmodeling methods. [7]

World Trade 1999 spring embedder



World Trade 1999 Pathfinder skeleton



Normalizations

Often some nodes/links prevail. How to make weights comparable? [5, 7, p. 94] We assume $w : \mathcal{L} \to \mathbb{R}_0^+$

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

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

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

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

. . . 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

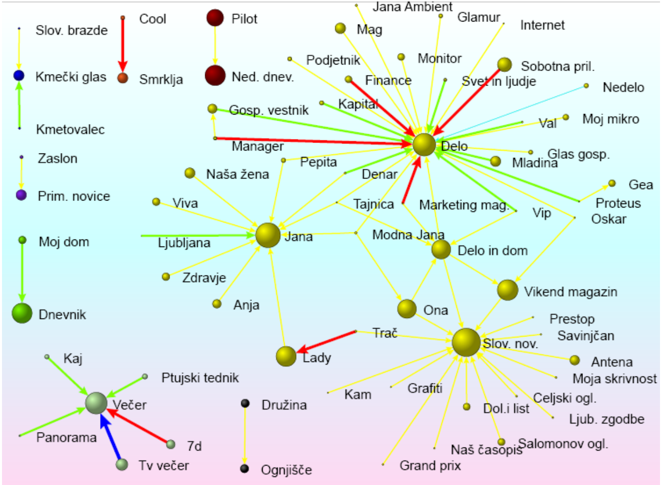
$$w_{vv} = \frac{1}{2}(\sum_u w_{vu} + \sum_u w_{uv})$$

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.

MinDir of Slovenian journals 2000



Over 100000 people were asked in the years 1999 and 2000 about the journals they read. They mentioned 124 different journals. (source Cati)

Nature of weights

1. (directly) measured: road traffic, baboons, ...
2. derived – computed from existing data
   1. projections of two-mode networks
      1. binary
      2. binary with NA
      3. nonnegative reals
      4. general
   2. network weight indexes: SPC weights, preferential attachment $w(u:v) = \deg(u) \cdot \deg(v)$, link betweeness, short cycles counts; (dis)similarity between end-nodes
   3. ... football
3. signed networks

many zeros, threshold, only for links of a given network

Binary projections

2-mode network $\mathcal{N} = ((\mathcal{U}, \mathcal{V}), \mathcal{L})$ represented with a *matrix* $\mathbf{UV} = [UV[u,v]]$. $UV[u,v] = 1$ if $(u,v) \in \mathcal{L}$, otherwise $UV[u,v] = 0$. *Neighbors* $N_U(v) = \{u \in U : (u,v) \in \mathcal{L}\}$.

A weighted network on the set of nodes $\mathcal{V}$ is determined by a *projection*

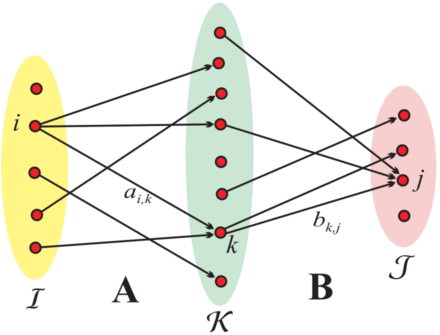
$$\mathbf{Co}(\mathbf{UV}) = \mathbf{UV}^T \cdot \mathbf{UV}$$

$$Co[v,z] = \sum_{u \in U} UV[u,v] \cdot UV[u,z] = |N_U(v) \cap N_U(z)|$$

$$Co[v,v] = |N_U(v)| = \deg(v)$$

A weighted network we intend to analyze was often essentially obtained this way.

Multiplication of networks



$$c_{i,j} = \sum_{k \in N_A(i) \cap N_B^-(j)} a_{i,k} \cdot b_{k,j}$$

If all weights in networks $\mathcal{N}_A$ and $\mathcal{N}_B$ are equal to 1 the value of $c_{i,j}$ of $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$ counts the number of ways we can go from $i \in \mathcal{I}$ to $j \in \mathcal{J}$ passing through $\mathcal{K}$, $c_{i,j} = |N_A(i) \cap N_B^-(j)|$.

Hamming distance

Co is a similarity measure. For clustering we need a dissimilarity. A related dissimilarity between sets is the *Hamming distance*

$$d_H(A,B) = |A \oplus B| = |A \cup B \setminus A \cap B| = |A| + |B| - 2|A \cap B|$$

d_H is a distance

1. $d_H(A,B) = 0 \Longrightarrow A = B$
2. $d_H(A,B) = d_H(B,A)$
3. $d_H(A,B) + d_H(B,C) \geq d_H(A,C)$

or in our case

$$D_H(v,z) = d_H(N_U(v), N_U(z)) = |N_U(v) \oplus N_U(z)|$$

$$= |Co[v,v]| + |Co[z,z]| - 2|Co[v,z]|$$

Note: $D_H(v,z) = 0$ implies $N_U(v) = N_U(z)$, but not $v = z$; v and z are structurally equivalent in the 2-mode network.

Binary similarity measures

The use of Jaccard similarity and some other similarities between binary vectors for analysis of two-mode networks was already proposed in SNA literature [11, p.420-424]. In principle, we could consider any similarity measure between binary vectors [2, 12].

For all these similarities the corresponding matrices can be computed from the events co-affiliation or co-appearances matrix (ordinary column projection) **Co**. The similarities between vectors v and z are expressed in terms of the quantities a, b, c, and d

$$
\begin{array}{cc|c|c|c}
 & & \multicolumn{2}{c}{z} & \\
 & & 1 & 0 & \\
\hline
\multirow{2}{*}{v} & 1 & a & b & a+b \\
\cline{2-5}
 & 0 & c & d & c+d \\
\hline
 & & a+c & b+d & |\mathcal{U}|
\end{array}
$$

The quantity a counts the number of cases (indices) for which both vectors v and z have value 1, etc.

Association coefficients

Association Coefficients

measure		definition	range	class
Russel and Rao (1940)	s_1	$\frac{a}{m}$	$[1,0]$	
Kendall, Sokal-Michener (1958)	s_2	$\frac{a+d}{m}$	$[1,0]$	S
Rogers and Tanimoto (1960)	s_3	$\frac{a+d}{m+b+c}$	$[1,0]$	S
Hamann (1961)	s_4	$\frac{a+d-b-c}{m}$	$[1,-1]$	S
Sokal & Sneath (1963), un_3^{-1}, S	s_5	$\frac{b+c}{a+d}$	$[0,\infty]$	S
Jaccard (1900)	s_6	$\frac{a}{a+b+c}$	$[1,0]$	T
Kulczynski (1927), T^{-1}	s_7	$\frac{a}{b+c}$	$[\infty,0]$	T
Dice (1945), Czekanowski (1913)	s_8	$\frac{a}{a+\frac{1}{2}(b+c)}$	$[1,0]$	T
Sokal and Sneath	s_9	$\frac{a}{a+2(b+c)}$	$[1,0]$	T
Kulczynski	s_{10}	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	$[1,0]$	
Sokal & Sneath (1963), un_4	s_{11}	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$	$[1,0]$	
Q_0	s_{12}	$\frac{bc}{ad}$	$[0,\infty]$	Q
Yule (1912), ω	s_{13}	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$[1,-1]$	Q
Yule (1927), Q	s_{14}	$\frac{ad-bc}{ad+bc}$	$[1,-1]$	Q
– bc –	s_{15}	$\frac{4bc}{m^2}$	$[0,1]$	
Driver & Kroeber (1932), Ochiai (1957)	s_{16}	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$[1,0]$	
Sokal & Sneath (1963), un_5	s_{17}	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1,0]$	
Pearson, ϕ	s_{18}	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1,-1]$	
Baroni-Urbani, Buser (1976), S^{**}	s_{19}	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$[1,0]$	
Braun-Blanquet (1932)	s_{20}	$\frac{a}{\max(a+b,a+c)}$	$[1,0]$	
Simpson (1943)	s_{21}	$\frac{a}{\min(a+b,a+c)}$	$[1,0]$	
Michael (1920)	s_{22}	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$[1,-1]$	

Association coefficients and co-appearance

For example, the *Jaccard similarity*

$$J = \frac{|N_U(v) \cap N_U(z)|}{|N_U(v) \cup N_U(z)|} = \frac{a}{a+b+c}.$$

The following equalities hold

$$\begin{aligned} a &= |N_U(v) \cap N_U(z)| = Co[v,z] \\ a+b &= |N_U(v)| = \deg(v) = Co[v,v] \\ a+c &= |N_U(z)| = \deg(z) = Co[z,z] \\ a+b+c+d &= |U|. \end{aligned}$$

From them we get

$$\begin{aligned} b &= |N_U(v) \setminus N_U(z)| = Co[v,v] - Co[v,z] \\ c &= |N_U(z) \setminus N_U(v)| = Co[z,z] - Co[v,z] \\ d &= |U| + Co[v,z] - Co[v,v] - Co[z,z] \\ a+b+c &= |N_U(v) \cup N_U(z)| = Co[v,v] + Co[z,z] - Co[v,z] \end{aligned}$$

Jaccard

For Jaccard similarity we get [10]

$$Jcol[v,z] = \frac{Co[v,z]}{Co[v,v] + Co[z,z] - Co[v,z]}$$

In this sense, the similarity measures (matrices) can be seen as a kind of compatibility normalization of the weights obtained with the standard projection [5].

The corresponding Jaccard network $\mathcal{J} = (V, L_J, Jcol)$ is undirected with loops removed.

The Jaccard dissimilarity

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \oplus B|}{|A \cup B|}$$

is also a distance. Note: $b + c = |N_U(v) \oplus N_U(z)|$.

Co-appearances and similarities

- **networks obtained by projection have a special structure – they are sums of complete subgraphs** [8].
- in creating the co-appearances network **Co** it is important to include the loops.
- the (a,b,c,d) (dis)similarities can be computed from the co-appearances network **Co**.
- projections can be used to convert survey data (two-mode network) into a weighted network.

Notes on the missing diagonal values

Setting a missing diagonal value $Co[v,v]$ for node v to the corresponding out-diagonal row sum $R'(v)$ is not an exact solution. We know that $Co[v,v] = \deg_U(v)$.

$$R'(v) + Co[v,v] = \sum_{z \in V} Co[v,z] = \sum_{z \in V} \sum_{u \in U} UV[u,v] \cdot UV[u,z] =$$

$$= \sum_{u \in N_U(v)} \sum_{z \in V} UV[u,z] = \sum_{u \in N_U(v)} \deg_V(u)$$

If $\deg_V(u) \geq 2$ for all $u \in N_U(v)$ then

$$R'(v) = \deg_U(v) + \sum_{u \in N_U(v)} (\deg_V(u) - 2) \geq \deg_U(v)$$

. . . Notes on the missing diagonal values

Let $\mathbf{Co}'$ denote the corrected matrix and $\mathbf{Co}$ the correct one; $a = Co[v,z]$ and $a' = Co'[v,z]$. Then $a' = a + k$ can be considered as a (1,1)-counter for vectors $v' = (v,x)$ and $z' = (z,x)$ where x is a vector with k values 1. The corresponding (dis)similarity measure is still a measure of the same type. For example, $J' \in [0,1]$.

What is the relation between J and J' ?

$$J = \frac{a}{a+b+c} \quad ? \quad \frac{a+k}{a+k+b+c} = J'$$

$$a \cdot (a+k+b+c) \quad ? \quad (a+b+c)\cdot(a+k)$$

$$0 \quad ? \quad (b+c)\cdot k$$

Therefore $J \leq J'$.

Fractional approach

To make in a projection the contributions of all nodes from the set $\mathcal{U}$ equal we apply the *fractional approach* by normalizing the weights [8].

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

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

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

Normalized projection $\quad \mathbf{Cn} = n(\mathbf{UV})^T \cdot n(\mathbf{UV})$

Newman's normalization

In bibliometric applications on network **WA**, the derived network **Co** describes the co-authorship (collaboration) among authors. The network **Cn** 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.

Stochastic or Markov normalization

Let $R(v) = \sum_z w[v, z]$ denote the row sum for a node v; and $Q(z) = \sum_v w[v, z]$ the column sum for a node z. For $R(v) > 0$, we define a *stochastic normalization* with

$$\mathbf{S}[v, z] = \frac{w[v, z]}{R(v)}$$

In real life networks it can happen that for some node v we have $R(v) = 0$. In such a case all entries in the row corresponding to v are 0, $w[v, z] = 0$. Then also $\mathbf{S}[v, z] = 0$.

For $R(v) > 0$ we have $\sum_z \mathbf{S}[v, z] = 1$ – the matrix entries can be interpreted as probabilities that a node v coappears with a node z. Usually the matrix $\mathbf{S}$ is not symmetric.

Balassa normalization

Let $T = \sum_{v,z} w[v,z]$ the total sum of weights in the network. If the network is undirected/symmetric then $R(v) = Q(v)$.

Then $R(v)/T$ is the proportion/probability of activity of the node v, and the expected activity $\hat{w}[v,z]$ from v to z is equal to

$$\hat{w}[v,z] = \frac{R(v)}{T} \cdot Q(z)$$

The measured weight $w[v,z]$ may deviate for a factor $a[v,z]$ from the expected value $w[v,z] = a[v,z] \cdot \hat{w}[v,z]$ or

$$a[v,z] = \frac{w[v,z] \cdot T}{R(v) \cdot Q(z)}$$

If $a[v,z] > 1$ the measured weight is larger than expected, ...
The deviation measure a is called the *Balassa index* or the "revealed comparative advantage" [1].

. . . Balassa normalization

The codomain of Balassa index is not 'symmetric' – the larger values have range $(1, \infty)$, the smaller values have range $(1, 0)$.

To 'symmetrize' it we apply, as suggested by [22], a logarithmic function to it. For easier interpretation, we selected base 2 logarithms

$$b[v, z] = \log_2 a[v, z], \quad \text{for} \quad a[v, z] > 0$$

In our analysis of post-Soviet countries we used the index b.

Salton or cosine similarity

In the case of general weights in the two-mode network **UV** the Salton index is often used in construction of the network on the set of nodes $\mathcal{V}$

$$s[v,z] = \frac{\sum_u UV[u,v]\cdot UV[u,z]}{\sqrt{\sum_u UV[u,v]^2 \cdot \sum_u UV[u,z]^2}}$$

In general $s \in [-1,1]$. If $UV \geq 0$ then $s \in [0,1]$.

Using s we can define the angular distance θ wp

$$\theta[v,z] = \frac{\arccos(s[v,z])}{\pi}$$

[17]

We could use any (dis)similarity measure between appropriately normalized columns of **UV** [14, 20].

Dissimilarities on $\mathbb{R}^m$ / 1

n	measure	definition	range	note
1	Euclidean	$\sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$	$[0, \infty)$	$M(2)$
2	Sq. Euclidean	$\sum_{i=1}^{m}(x_i - y_i)^2$	$[0, \infty)$	$M(2)^2$
3	Manhattan	$\sum_{i=1}^{m}\lvert x_i - y_i\rvert$	$[0, \infty)$	$M(1)$
4	rook	$\max_{i=1}^{m}\lvert x_i - y_i\rvert$	$[0, \infty)$	$M(\infty)$
5	Minkowski	$\sqrt[p]{\sum_{i=1}^{m}(x_i - y_i)^p}$	$[0, \infty)$	$M(p)$

n	measure	definition	range	note
6	Canberra	$\sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$	$[0, \infty)$	$\mathbb{R}_0^+$
7	Heincke	$\sqrt{\sum_{i=1}^{m} (\frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert})^2}$	$[0, \infty)$	$\mathbb{R}_0^+$
8	Self-balanced	$\sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\max(x_i, y_i)}$	$[0, \infty)$	$\mathbb{R}_0^+$
9	Lance-Williams	$\frac{\sum_{i=1}^{m} \lvert x_i - y_i \rvert}{\sum_{i=1}^{m} x_i + y_i}$	$[0, \infty)$	$\mathbb{R}_0^+$
10	Correlation c.	$\frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$	$[1, -1]$	

General projections

The *binarization* $b(\mathbf{UV})$ of the network $\mathbf{UV}$ is obtained by setting weights of all its links to 1.

$$b(UV)[u,v] = 1 \Leftrightarrow (u,v) \in \mathcal{L}$$

We define, generalizing [8], the *left* and the *right contribution* $l(\mathbf{UV}) = \mathbf{UV}^T \cdot b(\mathbf{UV})$ and $r(\mathbf{UV}) = b(\mathbf{UV})^T \cdot \mathbf{UV}$

It holds

$$r(UV)[z,v] = \sum_{u\in\mathcal{U}} b(UV)^T[z,u] \cdot UV[u,v] =$$

$$\sum_{u\in\mathcal{U}} b(UV)[u,z] \cdot UV[u,v] = \sum_{u\in\mathcal{U}} UV^T[v,u] \cdot b(UV)[u,z] = l(UV)[v,z]$$

General projection

The joint contribution can be defined as some mean of both contributions

$$biCnX[v,z] = meanX(l(UV)[v,z], r(UV)[v,z]) =$$

$$= meanX(l(UV)[v,z], l(UV)[z,v])$$

We need to compute only the left contributions.

Instead of the network **UV** we can use its generalized normalized version

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

General projection

$biCnA[v,z] = \frac{1}{2}(I(UV)[v,z] + I(UV)[z,v])$ – arithmetic
$biCnm[v,z] = \min(I(UV)[v,z], I(UV)[z,v])$ – minimum
$biCnM[v,z] = \max(I(UV)[v,z], I(UV)[z,v])$ – maximum
$biCnG[v,z] = \sqrt{I(UV)[v,z] \cdot I(UV)[z,v]}$ – geometric, Salton
$biCnH[v,z] = 2(I(UV)[v,z]^{-1} + I(UV)[z,v]^{-1})^{-1}$ – harmonic, Dice
$biCnJ[v,z] = (I(UV)[v,z]^{-1} + I(UV)[z,v]^{-1} - 1)^{-1}$ – Jaccard

$$biCnJ[v,z] \leq biCnm[v,z] \leq biCnH[v,z] \leq biCnG[v,z] \leq$$
$$\leq biCnA[v,z] \leq biCnM[v,z]$$

$$biCnX[v,z] = biCnX[z,v]$$
$$I \in [0,1] \Rightarrow biCnX \in [0,1]$$

Alternative normalized projections of binary networks

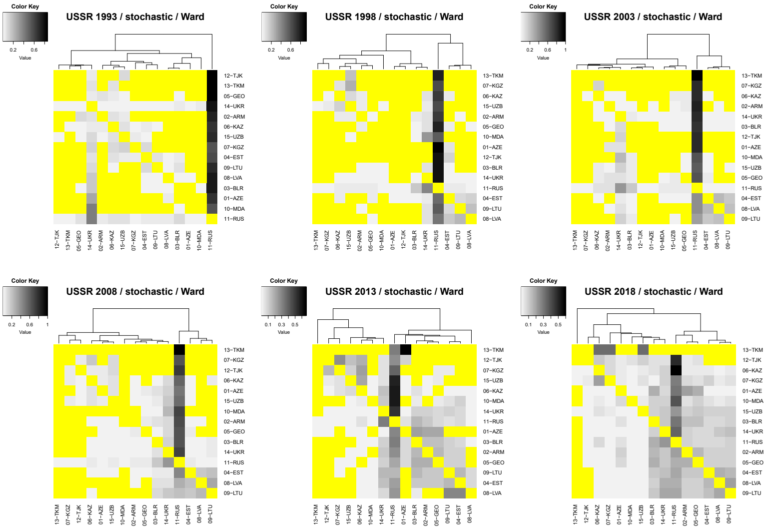
$$
\begin{aligned}
Cn_A[v,z] &= \frac{|N_U(v) \cap N_U(z)|}{2}\left(\frac{1}{|N_U(v)|} + \frac{1}{|N_U(z)|}\right) \\
Cn_m[v,z] &= \frac{|N_U(v) \cap N_U(z)|}{\max(|N_U(v)|, |N_U(z)|)} \\
Cn_M[v,z] &= \frac{|N_U(v) \cap N_U(z)|}{\min(|N_U(v)|, |N_U(z)|)} \\
Cn_G[v,z] &= \frac{|N_U(v) \cap N_U(z)|}{\sqrt{|N_U(v)| \cdot |N_U(z)|}} \\
Cn_H[v,z] &= \frac{2|N_U(v) \cap N_U(z)|}{|N_U(v)| + |N_U(z)|} \\
Cn_J[v,z] &= \frac{|N_U(v) \cap N_U(z)|}{|N_U(v) \cup N_U(z)|}
\end{aligned}
$$

Corrected Euclidean distance

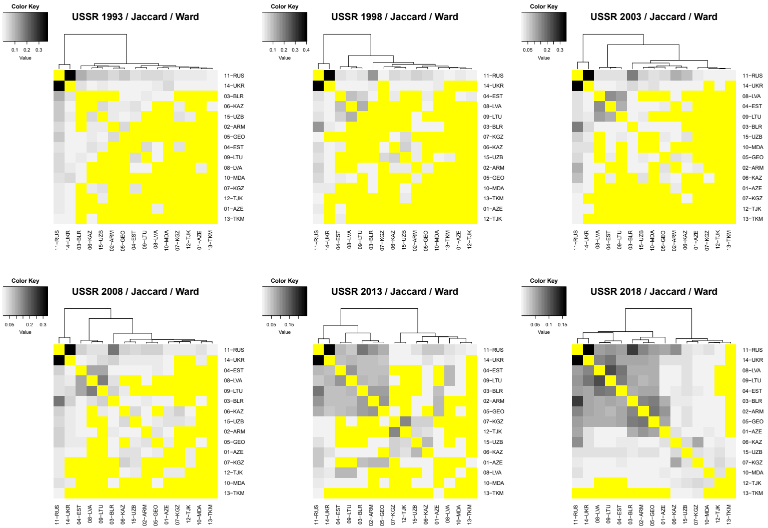
For clustering the nodes we used the normalized matrices with zero diagonals. For the index b we set the values of non-links to 0. When computing dissimilarity $\mathbf{D}[v, z]$ between nodes v and z in a network it is important to use the corrected dissimilarities that compare the entry $w_{vz} = w[v, z]$ with the entry $w_{zv} = w[z, v]$ and the entry $w[v, v]$ with the entry $w[z, z]$. We selected the *corrected Euclidean distance* [16, p. 181]

$$\mathbf{D}[v, z] = \sqrt{(w_{vz} - w_{zv})^2 + (w_{vv} - w_{zz})^2 + \sum_{y: y \neq v, y \neq z} (w_{vy} - w_{zy})^2}$$

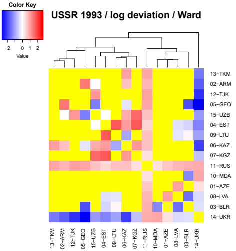
Stochastic normalization

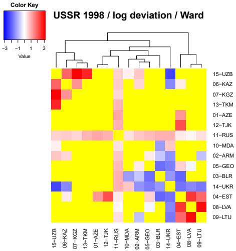


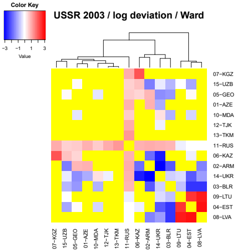
Jaccard normalization

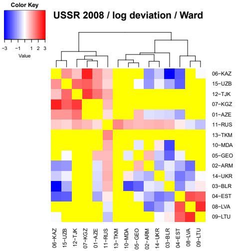


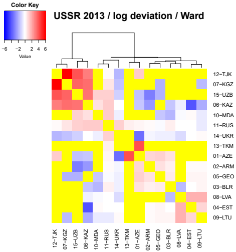
Logarithmic deviations

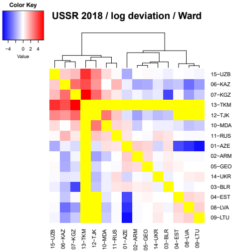




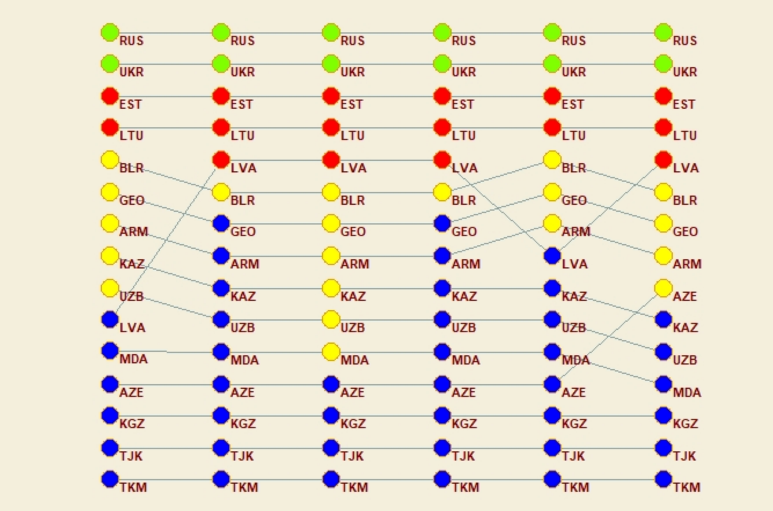








Jaccard transitions



Conclusions

- use a normalization that is expressing your research question
- generalization to weighted 2-mode networks
- projections of two-mode networks with NAs

References I

- Balassa, B. (1965). Trade Liberalisation and Revealed Comparative Advantage. The Manchester School, 33, 99-123.
- Batagelj, V., Bren, M. (1995) Comparing resemblance measures. J Classif 12 (1): 73-90.
- 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. (2003). Density-based approaches to network analysis; Analysis of Reuter's terror news network. Workshop on Link Analysis for Detecting Complex Behavior (LinkKDD2003), August 27, 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.

References II

Batagelj, V.: On fractional approach to the analysis of linked networks. Scientometrics 123 (2020), pages621–633. Springer

Batagelj, V., Maltseva, D.: Temporal bibliographic networks. Journal of Informetrics, 14 (2020) 1, 101006.

Batagelj, V. (2021). Analysis of the Southern women network using the fractional approach. Social Networks 68, 229–236.

Borgatti, S.P., Halgin, D.S. (2014) Analyzing Affiliation Networks. Chapter 28 in The SAGE Handbook of Social Network Analysis. John Scott, Peter J. Carrington (Eds.), Sage.

Cibulková, J., Šulc, Z., Řezanková, H., Sirota, S. (2020). Associations among similarity and distance measures for binary data in cluster analysis. Metodološki zvezki, Vol. 17, No. 1, 33–54. PDF

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.

Deza, M.M., Deza, E. (2013) Encyclopedia of Distances. Springer.

References III

- Doreian, P. (1985). A measure of standing of journals in stratified networks. Scientometrics 8(5): 341-363. PDF
- Doreian, P., Batagelj, V., Ferligoj, A. (2005). Generalized Blockmodeling. Cambridge University Press.
- Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. Journal of the American Society for Information Science and Technology, 59(1), 77-85.
- Pajek datasets: Journals / Slovenian magazines and journals 1999 and 2000. WWW
- Pajek datasets: days / Reuters terror news network. WWW
- SAS Institute Inc (2013). SAS/STAT® 13.1 User's Guide / The DISTANCE procedure. Cary, NC: SAS Institute Inc. PDF
- Van Cutsem, B. (ed.) (1994) Classification and Dissimilarity Analysis. Lecture Notes in Statistics 93. Springer.

References IV

Vollrath, T.L. (1991). A theoretical evaluation of alternative trade intensity measures of revealed comparative advantage. Weltwirtschaftliches Archiv 127, 265–280.

Zheng, Q., Skillicorn, D.: Social Networks with Rich Edge Semantics. CRC 2017.