Weighted two-mode projections

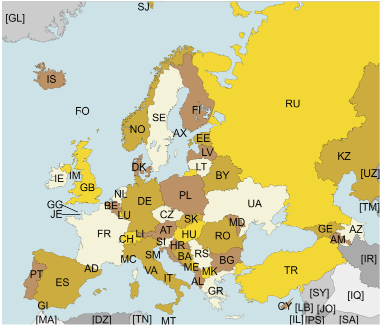
V. Batagelj

On weighted two-mode network projections

Vladimir Batagelj
UP FAMNIT Koper

1338. sredin seminar
Ljubljana, August 9, 2023

Outline



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

Current version of slides (August 9, 2023 at 17 : 30): PDF

European science

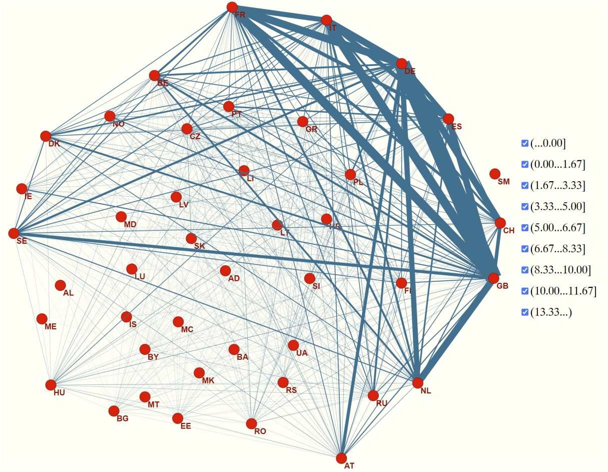
Motivation

In the paper [4] we studied the collaboration (co-authorship) between scientists from different post-Soviet countries. We decided to repeat the study on the European countries. It turned out that there are different ways how we can define a network describing the co-authorship collaboration between countries. Some options are discussed in this paper.

From the Centre for Science & Technology Studies (CWTS), Leiden University, NL we obtained data for the years 2000-2022.

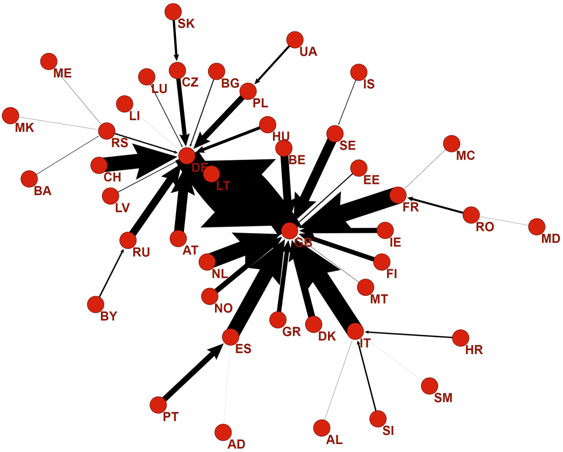
Some exploratory analyses

Europe 2019 by levels



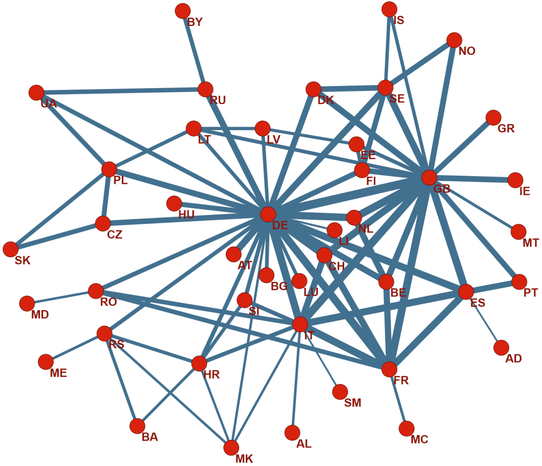
Europe by levels

Europe 2019 closest neighbor skeleton



Europe closest neighbor

Europe 2019 PathFinder skeleton



Europe PathFinder skeleton

Two-mode networks

In a *two-mode* (affiliation or bipartite) network $\mathcal{N} = ((U, V), L, w)$ the set of nodes is split into two disjoint sets (*modes*) U and V. Each link $e \in L$ has one end-node $u \in U$ and the other end-node $v \in V$, $\text{ext}(e) = (u, v)$. The function $w : L \to \mathbb{R}$ assigns to each link its weight.

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2. Miss Laura Mandeville	×	×	×		×	×	×	×						
3. Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4. Miss Brenda Rogers	×		×	×	×	×	×	×						
5. Miss Charlotte McDowd			×	×	×		×							
6. Miss Frances Anderson			×		×	×		×						
7. Miss Eleanor Nye					×	×	×	×						
8. Miss Pearl Oglethorpe						×		×	×					
9. Miss Ruth DeSand					×		×	×	×					
10. Miss Verne Sanderson							×	×	×			×		
11. Miss Myra Liddell								×	×	×		×		
12. Miss Katherine Rogers								×	×	×		×	×	×
13. Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14. Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15. Mrs. Helen Lloyd							×	×		×	×	×		
16. Mrs. Dorothy Murchison								×	×					
17. Mrs. Olivia Carleton									×		×			
18. Mrs. Flora Price									×		×			

Davis: Southern women, 1941

Two-mode network matrix and some notions

In general, the weight can be measured on different measurement scales (counts, ratio, interval, ordinal, nominal, binary,TQ, etc.).

We extend the semiring $(\mathbb{R} \cup \{\square\}, +, \cdot)$ with a new zero $\square$ with rules $\square + a = a$ and $\square \cdot a = a \cdot \square = \square$.

The network *matrix* $\mathbf{M} = [m[u,v]]$ of a two-mode network $\mathcal{N}$ is defined as

$$m[u,v] = \begin{cases} \sum_{e \in L : \mathrm{ext}(e)=(u,v)} w(u,v) & \exists e \in L : \mathrm{ext}(e) = (u,v) \\ \square & \text{otherwise} \end{cases}$$

Multiplication of networks

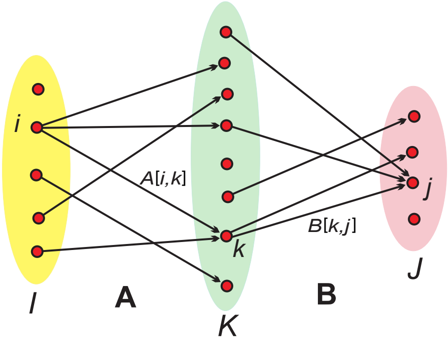
The *product* $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$ of two compatible matrices $\mathbf{A}_{I \times K}$ and $\mathbf{B}_{K \times J}$ is defined in the standard way

$$C[i,j] = \sum_{k \in K} A[i,k] \cdot B[k,j]$$

(it can be extended to semirings !!!)

The product of two compatible networks $\mathcal{N}_A = ((I,K), L_A, a)$ and $\mathcal{N}_B = ((K,J), L_B, b)$ is the network $\mathcal{N}_C = ((I,J), L_C, c)$ where $L_C = \{(i,j) : c[i,j] \neq \square\}$ and the weight c is determined by the matrix $\mathbf{C}$, $c(i,j) = C[i,j]$.

Multiplication of networks



In binary networks $\mathcal{N}_A$ and $\mathcal{N}_B$, the value of $C[i,j]$ of $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$ counts the number of ways we can go from the node $i \in I$ to the node $j \in J$ passing through K, $C[i,j] = |N_A(i) \cap N_B(j)|$.

$$C[i,j] = \sum_{k \in N_A(i) \cap N_B(j)} A[i,k] \cdot B[k,j]$$

Projections of binary two-mode networks

Most of the bibliometric networks are obtained by a projection of a non-weighted network represented by a binary matrix. For example from the authorship network WA describing the authorship relation of the set of works (papers, books, reports, etc.) W by the authors from the sets A. It is represented by a matrix $\mathbf{WA} = [wa[w, a]]$ where $wa[w, a] = 1$ iff a is an author of the work w and 0 otherwise. We get the co-authorship (counting) network Co_A determined by the projection

$$\mathbf{Co_A} = \mathbf{WA}^T \cdot \mathbf{WA}$$

. . . Projections of binary two-mode networks

As we know [2]

- For $a \neq b$, $co_A[a, b]$ = number of works co-authored by authors a and b.
- $co_A[a, a]$ = number of works from W written by the author a.
- The works with a large number of coauthors are "overrepresented" in the network Co_A – for example, the co-authorship of authors of a paper with 2 authors counts the same as the co-authorship between any pair of authors of the paper with 1000 co-authors; a paper with 1000 co-authors adds 1000000 links to projection network; while a single author paper only a loop. For this reason, the number $co_A[a, b]$ is not the best measure for measuring the collaboration intensity.

Weighted projections

The case of collaboration between countries is slightly different because the two-mode network WC is weighted. Actually, we could get it as $\mathbf{WC} = \mathbf{WA} \cdot \mathbf{AC}$ where AC is the author-to-country affiliation network. This view opens a possibility to deal with authors affiliated to different countries provided that $\sum_c ac[a, c] = 1$. If the affiliations are changing through time the temporal quantities can be used [3].

To obtain a collaboration network between a set of countries C based on a set of works W, we start with a two-mode network WC described by a matrix $\mathbf{WC} = [wc[w, c]]$ where

$$wc[w, c] = \text{number of authors of the work } w \text{ from the country } c$$

Authors counting collaboration

In the network WC we can consider all authors of selected works W by adding to the set of countries C also the "country" Others. Instead of countries, other partitions of the set of authors can be used, for example, institutions.

We will use $T(N) = \sum_{e \in L} w(e)$ to denote the total sum of weights of all links of the network $N = (V, L, w)$.

The *authors counting collaboration* network Co_C described by the matrix $\mathbf{Co_C}$ is obtained by projection

$$\mathbf{Co_C} = \mathbf{WC}^T \cdot \mathrm{bin}(\mathbf{WC})$$

where $\mathrm{bin}(\mathbf{WC}) = [\widehat{wc}[w,c]]$, and $\widehat{wc}[w,c] = 1$ iff $wc[w,c] \neq 0$ and 0 otherwise.

... Authors counting collaboration

What are the meaning of the entries $co_C[a, b]$ and their properties?

a For $a \neq b$, $co_C[a, b] = \sum_w wc[w, a] \cdot \widehat{wc}[w, b]$ – number of appearances of **authors** from the country a in works co-authored also by some author from the country b. We will denote this number $\text{wdeg}_{WC}(a/b)$.

b $co_C[a, a] = \sum_w wc[w, a] \cdot \widehat{wc}[w, a] = \text{wdeg}_{WC}(a)$ – number of appearances of authors from the country a in works from W; a column sum for country a in the matrix **WC**.

c From a simple example

$$\mathbf{WC} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{array} \begin{array}{c} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{bmatrix} 0 & 2 & 1 \\ 2 & 1 & 0 \\ 1 & 3 & 1 \\ 3 & 0 & 2 \\ 2 & 3 & 1 \\ 1 & 0 & 3 \end{bmatrix} \end{array} \qquad \mathbf{Co_C} = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \end{array} \begin{array}{c} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{bmatrix} 9 & 5 & 7 \\ 7 & 9 & 8 \\ 7 & 3 & 8 \end{bmatrix} \end{array}$$

we see that the matrix $\mathbf{Co_C}$ is in general not symmetric – there can exist pairs a, b such that $co_C[a, b] \neq co_C[b, a]$.

. . . Authors counting collaboration

d Consider a row sum $R(a)$ for the country a in the matrix $\mathbf{Co_C}$. We get

$$R(a) = \sum_b co_C[a, b] = \sum_w wc[w, a] \cdot \sum_b \widehat{wc}[w, b] = \sum_w wc[w, a] \cdot \mathrm{deg}_{WC}(w)$$

Since in the network WC only works W with co-authors from at least 2 countries are considered, we have $\mathrm{deg}_{WC}(w) \geq 2$ and we can continue $R(a) \geq 2 \sum_w wc[w, a] = 2\ \mathrm{wdeg}_{WC}(a)$. Now, combined with **b**, we finally get

$$\sum_{b:b\neq a} co_C[a, b] \geq \mathrm{wdeg}_{WC}(a) = co_C[a, a]$$

The sum of the out-diagonal entries in the a row of the matrix $\mathbf{Co_C}$ is larger or equal to its diagonal entry.

From the example in **c** we see that this property does not hold for columns – see the column c_2.

… Authors and works counting collaboration

e For the diagonal values of the network Co_C it holds $co_C[c, c] = \mathrm{wdeg}_{WC}(c)$

$$co_C[c, c] = \sum_w wc[w, c] \cdot \widehat{wc}[w, c] = \sum_w wc[w, c] = \mathrm{wdeg}_{WC}(c)$$

Therefore $\sum_c co_C[c, c] = T(WC)$.

f In the case when also the matrix **WC** is binary, $\mathrm{bin}(\mathbf{WC}) = \mathbf{WC}$, we deal with the standard projection mentioned in the introduction $\mathbf{Co_C} = \mathbf{WC}^T \cdot \mathbf{WC}$. In the *works counting collaboration* network $\mathbf{Co_b} = \mathrm{bin}(\mathbf{WC})^T \cdot \mathrm{bin}(\mathbf{WC})$ its weight $co_b[a, b]$ counts **works**: $co_b[a, b] =$ number of works from W co-authored by authors from countries a and b, and $co_b[a, a] =$ number of works from W co-authored by authors from the country a. Note that the inequality from **d** still holds (and also for columns).

Normalizations

For binary networks, we define their normalized versions: *standard* $n(\mathbf{WA}) = [wan[w, a]]$

$$wan[w, a] = \frac{wa[w, a]}{\max(1, \deg_{WA}(w))}$$

and *strict* (or Newman's) $N(\mathbf{WA}) = [waN[w, a]]$

$$waN[w, a] = \frac{wa[w, a]}{\max(1, \deg_{WA}(w) - 1)}$$

Using the normalized networks we define the *standard fractional projection*

$$\mathbf{Co_n} = n(\mathbf{WA})^T \cdot n(\mathbf{WA})$$

and the *strict fractional projection*

$$\mathbf{Co_N} = D_0(n(\mathbf{WA})^T \cdot N(\mathbf{WA}))$$

where the function $D_0(\mathbf{M})$ sets the diagonal of a square matrix $\mathbf{M}$ to 0.

We know [1] that if $\deg_{WA}(w) > 0$, each work $w \in W$ contributes equally, a unit 1, to the total weight of links in $\mathbf{Co_n}$. The same holds for $\mathbf{Co_N}$ if $\deg_{WA}(w) > 1$.

Standard weighted projection

To extend the fractional projections to weighted two-mode networks we define for the *standard* case $n(\mathbf{WC}) = [wcn[w,c]]$

$$wcn[w,c] = \frac{wc[w,c]}{\max(1, \mathrm{wdeg}_{WC}(w))}$$

Again we have $T(\mathbf{Co_n}) = |W|$ for $\mathbf{Co_n} = n(\mathbf{WC})^T \cdot n(\mathbf{WC})$.

$$\mathbf{Co_b} = \begin{array}{c c} & \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_3 \end{array} & \begin{bmatrix} 5 & 3 & 4 \\ 3 & 4 & 3 \\ 4 & 3 & 5 \end{bmatrix} \end{array} \qquad \mathbf{Co_n} = \begin{array}{c c} & \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_3 \end{array} & \begin{bmatrix} 1.0180556 & 0.5088889 & 0.5230556 \\ 0.5088889 & 1.1655556 & 0.4255556 \\ 0.5230556 & 0.4255556 & 0.9013889 \end{bmatrix} \end{array}$$

Computation in R

Strict weighted projection

There is no obvious way how to define the strict normalization for weighted networks.

There is another possible view on fractional projections. The definition of matrix $n(\mathbf{WA})$ can be written as $n(\mathbf{WA}) = \mathbf{d_n} \cdot \mathbf{WA}$ and similarly $N(\mathbf{WA}) = \mathbf{d_N} \cdot \mathbf{WA}$ where $\mathbf{d_n}$ is a diagonal $W \times W$ matrix with $d_n[w,w] = 1/\max(1, \deg_{WA}(w))$ and $\mathbf{d_N}$ with $d_N[w,w] = 1/\max(1, \deg_{WA}(w) - 1)$. In both cases we get $(\mathbf{d}^T = \mathbf{d})$

$$\mathbf{Co_n} = n(\mathbf{WA})^T \cdot n(\mathbf{WA}) = \mathbf{WA}^T \cdot \mathbf{d_n} \cdot \mathbf{d_n} \cdot \mathbf{WA}$$

$$\mathbf{Co_N} = n(\mathbf{WA})^T \cdot N(\mathbf{WA}) = \mathbf{WA}^T \cdot \mathbf{d_n} \cdot \mathbf{d_N} \cdot \mathbf{WA}$$

Because a product of diagonal matrices is a diagonal matrix, $\mathrm{diag}(a_w) \cdot \mathrm{diag}(b_w) = \mathrm{diag}(a_w \cdot b_w)$, both cases have a common form $\mathbf{WA}^T \cdot \mathbf{d} \cdot \mathbf{WA}$. It can be related to the weighted scalar product. Maybe this form can lead also to the extension of strict projection for weighted two-mode networks.

... Strict weighted projection

Let us look at a simple example. Assume, that a work w has authors from three countries a, b, and c. Then, since the co-authors inside the same country do not count, its contribution $T(w)$ to the total weight, see the contribution matrix

$$\mathbf{Co_C}(w) = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} \quad a \quad & \quad b \quad & \quad c \quad \end{array} \\ \left[\begin{array}{ccc} \mathbf{0} & wc[w,a]\cdot wc[w,b] & wc[w,a]\cdot wc[w,c] \\ wc[w,b]\cdot wc[w,a] & \mathbf{0} & wc[w,b]\cdot wc[w,c] \\ wc[w,c]\cdot wc[w,a] & wc[w,c]\cdot wc[w,b] & \mathbf{0} \end{array} \right] \end{array},$$

is $T(w) = \sum_{e,f \in \{a,b.c\} \wedge e \neq f} wc[w,e]\cdot wc[w,f]$. By the rule of product and the rule of sum from basic combinatorics [5], $T(W)$ is equal to twice the number of all co-authorships of authors from different countries – pairs (a,b) and (b,a) are representing co-authorship of authors a and b.

... Strict weighted projection

To make $T_N(w) = 1$ we must set the entry $d_N[w, w]$ of the diagonal matrix $\mathbf{d_N}$ for the weighted network **WC** to
$d_N[w, w] = 1/T(w) = 1/(\mathrm{wdeg}_{WC}(w)^2 - \sum_c wc[w, c]^2)$.

Note that $\sum_c wc[w, c] = \mathrm{wdeg}_{WC}(w)$ and

$$\mathrm{wdeg}_{WC}(w)^2 - \sum_c wc[w, c]^2 = \sum_{e,f:e\neq f} wc[w, e] \cdot wc[w, f]$$

The left side of this equality is computationally more convenient.

...Strict weighted projection

It is easy to see that we made a good guess – in the corresponding projection $\mathbf{Co_N} = D_0(\mathbf{WC}^T \cdot \mathbf{d_N} \cdot \mathbf{WC})$ each work contributes equally, a unit 1, to the total of link weights.

$$T_N(w) = d_N[w,w] \cdot \sum_{e,f:e \neq f} wc[w,e] \cdot wc[w,f] = 1$$

Therefore

$$T(Co_N) = \sum_w T_N(w) = |W|$$

For our example from Section 2 **c** we get

$$\mathbf{Co_N} = \begin{array}{c c} & \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_3 \end{array} & \begin{bmatrix} 0.000000 & 0.9870130 & 1.1623377 \\ 0.987013 & 0.0000000 & 0.8506494 \\ 1.162338 & 0.8506494 & 0.0000000 \end{bmatrix} \end{array}$$

with $T(Co_N) = 6$.

Weighted projections

C is a set of countries of our interest. If it is used, *others* $\in C$. $\mathcal{W}$ is a list of metadata about the works from the selected bibliographic data source, co-authored by authors from at least two different countries from $C \setminus \{others\}$. All four projection matrices $\mathbf{Co_b}$, $\mathbf{Co_n}$, $\mathbf{Co_C}$, and $\mathbf{Co_N}$ can be constructed in a single run through the list using the Algorithm given in the following slides.

Computing weighted projections I

function $\textsc{Projections}(\mathcal{W}, C)$

  $Cob \leftarrow Con \leftarrow CoC \leftarrow CoN \leftarrow matrix(0, nrow = |C|, ncol = |C|)$

  for $w \in \mathcal{W}$ **do**

    from the metadata of work w determine the list of pairs
    $Lw = ((a, ca))$, a – author. ca – country of a

    $Cw \leftarrow empty$

    for $(a, ca) \in Lw$ **do**

      $c \leftarrow ca$ **if** $ca \in C$ **else** $others$

      if $c \notin Cw$ **then**

        $Cw[c] \leftarrow 0$

      end if

      $Cw[c] \leftarrow Cw[c] + 1$

    end for

    write ID of w and Cw to the file `WC`

    $wdegw \leftarrow \sum_{c \in Cw} Cw[c]$

    $sqw \leftarrow \sum_{c \in Cw} Cw[c]^2$

    $dnw \leftarrow 1/wdegw^2$

    $dNw \leftarrow 1/(wdegw^2 - sqw)$

Computing weighted projections II


```
      for e ∈ Cw do
         for f ∈ Cw do
            Cob[e, f] ← Cob[e, f] + 1
            CoC[e, f] ← CoC[e, f] + Cw[e]
            Con[e, f] ← Con[e, f] + Cw[e] · Cw[f] · dnw
            if e ≠ f then
               CoN[e, f] ← CoN[e, f] + Cw[e] · Cw[f] · dNw
            end if
         end for
      end for
   end for
   return Cob, Con, CoC, CoN
end function
```

Test in R

Notes on the implementation of the algorithm

- A natural implementation of the counters Cw is as a dictionary.
- Line 13 is optional. It creates the file `WC` describing the two-mode network WC.
- Networks Co_b, Co_n, and Co_N are symmetric. They can be represented by an undirected network with the weight of an edge $(e : f)$ equal to twice the computed value, except for loops. The computation can be restricted to pairs (e, f) for which $e \leq f$.

Conclusions

In the paper, we derived the results in terms of the binary authorship network WA and the weighted network WC. The results hold in general for similar weighted two-mode networks such as

- (journals, universities, number of published articles of authors from the university u in the journal j in the selected time interval),
- (territorial units, universities, number of students from the territorial unit t studying this year at the university u),
- (web resources (movies or music tracks), types of resource, number of times the resource r of type t was downloaded in the selected time interval),
- (retail chain customers (chain card owners), (types of) products, the value of the product p bought by the customer c in the selected time interval),

etc.

References I

- Batagelj, V: On fractional approach to analysis of linked networks. Scientometrics 123 (2020) 2: 621-633
  https://doi.org/10.1007/s11192-020-03383-y

- Batagelj, V, Cerinšek, M: On bibliographic networks. Scientometrics 96 (2013) 3, 845-864.
  https://doi.org/10.1007/s11192-012-0940-1

- Batagelj, V, Maltseva, D: Temporal bibliographic networks. Journal of Informetrics, Volume 14, Issue 1, February 2020, 101006.
  https://doi.org/10.1016/j.joi.2020.101006

- Matveeva, N., Batagelj, V., Ferligoj, A.: Scientific collaboration of post-Soviet countries: the effects of different network normalizations. Scientometrics 128 (2023), 4219–4242.
  https://doi.org/10.1007/s11192-023-04752-z

- Wikipedia: Combinatorial principles.
  https://en.wikipedia.org/wiki/Combinatorial_principles