On Fractional Approach to Analysis of Linked Networks

Vladimir Batagelj[1,2,3]
[1]Institute of Mathematics, Physics and Mechanics,
Jadranska 19, 1000 Ljubljana, Slovenia
[2]University of Primorska, Andrej Marušič Institute, 6000 Koper, Slovenia
[3] National Research University Higher School of Economics,
Myasnitskaya, 20, 101000 Moscow, Russia.
vladimir.batagelj@fmf.uni-lj.si

February 11, 2019

Abstract

In this paper, we present the outer product decomposition of a product of compatible linked networks. It provides a foundation for the fractional approach in network analysis. We discuss the standard and Newman's normalization of networks. We propose some alternatives for fractional bibliographic coupling measures.

Keywords: social network analysis, linked networks, bibliographic networks, network multiplication, fractional approach, Newman's normalization, bibliographic coupling.

1 Introduction

The fractional approach was proposed by Lindsey (1980). For example in the analysis of coauthorship the contributions of all coauthors to a work has to add to 1. Usually the contribution is then estimated as 1 divided by the number of coauthors. An alternative rule, Newman's normalization, was given in Newman (2001) and Newman (2004) which excludes the selfcollaboration. Recently several papers (Batagelj and Cerinšek, 2013; Cerinšek and Batagelj, 2015; Perianes-Rodriguez et al., 2016; Prathap and Mukherjee, 2016; Leydesdorff and Park, 2017; Gauffriau, 2017) reconsidered the background of the fractional approach. The details are presented and discussed in Subsection 6.2. In this paper we propose a theoretical framework based on the outer product decomposition to get the insight into the structure of bibliographic networks obtained with network multiplication.

2 Linked networks

Linked or multi-modal networks are collections of networks over at least two sets of nodes (modes) and consist of some one-mode networks and some two-mode networks linking different modes. For example: modes are Persons and Organizations. Two one-mode networks describe collaboration among Persons and among Organizations. The linking two-mode network describes membership of Persons to different Organizations.

Linked networks are the basis of the MetaMatrix approach developed by Krackhardt and Carley (Krackhardt and Carley, 1998; Carley, 2003). For an example see the Table 3 in Diesner and Carley (2004, p. 89).

Another example of linked networks are bibliographic networks. From special bibliographies (BibTEX) and bibliographic services (Web of Science, Scopus, SICRIS, CiteSeer, Zentralblatt MATH, Google Scholar, DBLP Bibliography, US patent office, IMDb, and others) we can construct some two-mode networks on selected topics: authorship on works $\times$ authors (**WA**), keywordship on works $\times$ keywords (**WK**), journalship on works $\times$ journals/publishers (**WJ**), and from some data also the classification network on works $\times$ classification (**WC**) and the one-mode citation network on works $\times$ works (**Ci**); where works include papers, reports, books, patents, movies, etc. Besides this we get also the partition of works by the publication year, and the vector of number of pages (WoS, 2018; Batagelj, 2007).

An important tool in analysis of linked networks is the use of derived networks obtained by network multiplication.

3 Network multiplication

Given a pair of *compatible* two-mode networks $\mathcal{N}_A = (\mathcal{I}, \mathcal{K}, \mathcal{A}_A, w_A)$ and $\mathcal{N}_B = (\mathcal{K}, \mathcal{J}, \mathcal{A}_B, w_B)$ with corresponding matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$ we call a *product of networks* $\mathcal{N}_A$ and $\mathcal{N}_B$ a network $\mathcal{N}_C = (\mathcal{I}, \mathcal{J}, \mathcal{A}_C, w_C)$, where $\mathcal{A}_C = \{(i,j) : i \in \mathcal{I}, j \in \mathcal{J}, c_{i,j} \neq 0\}$ and $w_C(i,j) = c_{i,j}$ for $(i,j) \in \mathcal{A}_C$. The product matrix $\mathbf{C} = [c_{i,j}]_{\mathcal{I}\times\mathcal{J}} = \mathbf{A} \cdot \mathbf{B}$ is defined in the standard way

$$c_{i,j} = \sum_{k \in \mathcal{K}} a_{i,k} \cdot b_{k,j}$$

In the case when $\mathcal{I} = \mathcal{K} = \mathcal{J}$ we are dealing with ordinary one-mode networks (with square matrices).

In the following we will often identify networks by their matrices.

In the paper Batagelj and Cerinšek (2013) it is shown that $c_{i,j}$ is equal to the value of all two step paths from $i \in \mathcal{I}$ to $j \in \mathcal{J}$ passing through $\mathcal{K}$. In a special case, if all weights in networks $\mathcal{N}_A$ and $\mathcal{N}_B$ are equal to 1 the value of $c_{i,j}$ counts the number of ways we can go from $i \in \mathcal{I}$ to $j \in \mathcal{J}$ passing through $\mathcal{K}$: $c_{i,j} = |N_A(i) \cap N_B^-(j)|$; where $N_A(i)$ is the set of nodes in $\mathcal{K}$ linked by arcs from node i in the network $\mathcal{N}_A$, and $N_B^-(j)$ is the set of nodes in $\mathcal{K}$ linked by arcs to node j in the network $\mathcal{N}_B$.

The standard matrix multiplication has the complexity $O(|\mathcal{I}| \cdot |\mathcal{K}| \cdot |\mathcal{J}|)$ – it is too slow to be used for large networks. For sparse large networks we can multiply much faster considering only nonzero elements.

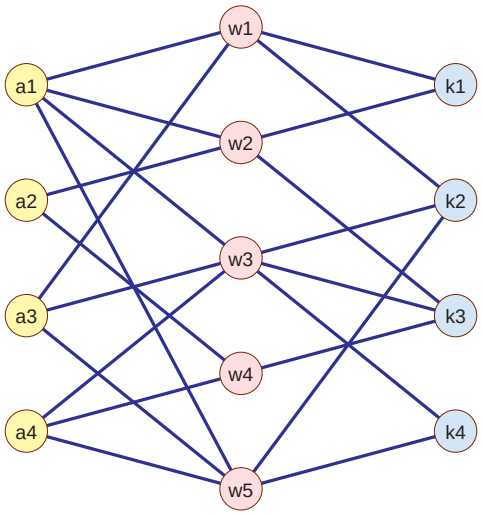

Figure 1: $\mathbf{WA}^T \cdot \mathbf{WK}$

for k **in** $\mathcal{K}$ **do**
  for (i,j) **in** $N_A^-(k) \times N_B(k)$ **do**
    if $\exists c_{i,j}$ **then** $c_{i,j} := c_{i,j} + a_{i,k} \cdot b_{k,j}$
    else new $c_{i,j} := a_{i,k} \cdot b_{k,j}$

In general the multiplication of large sparse networks is a 'dangerous' operation since the result can 'explode' – it is not sparse. If for the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$ there are in $\mathcal{K}$ only few nodes with large degree and no one among them with large degree in both networks then also the resulting product network $\mathcal{N}_C$ is sparse.

From the network multiplication algorithm we see that each intermediate node $k \in \mathcal{K}$ adds to a product network a complete two-mode subgraph $K_{N_A^-(k),N_B(k)}$ (or, in the case $\mathbf{B} = \mathbf{A}^T$, where $\mathbf{A}^T$ is the transposition of $\mathbf{A}$, a complete subgraph $K_{N(k)}$). If both degrees $\deg_A(k) = |N_A^-(k)|$ and $\deg_B(k) = |N_B(k)|$ are large then already the computation of this complete subgraph has a quadratic (time and space) complexity – the result 'explodes'. For details see the paper Batagelj and Cerinšek (2013).

4 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

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

then we can express the previous observation about the structure of product network as the *outer product decomposition*

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

For binary (weights) networks we have $\mathbf{H}_k = K_{N_A^-(k),N_B(k)}$.

Example A: As an example let us take the binary network matrices **WA** and **WK**:

$$\mathbf{WA} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \end{array}, \quad \mathbf{WK} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

and compute the product $\mathbf{H} = \mathbf{WA}^T \cdot \mathbf{WK}$. We get a network matrix $\mathbf{H}$ which can be decomposed as

$$\begin{array}{c} \mathbf{H} \\ \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 2 & 3 & 2 & 2 \\ 1 & 0 & 2 & 0 \\ 1 & 3 & 1 & 2 \\ 0 & 2 & 2 & 2 \end{bmatrix} \end{array} = \begin{array}{c} \mathbf{H}_1 \\ \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} + \begin{array}{c} \mathbf{H}_2 \\ \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} +$$

$$\begin{array}{c} \mathbf{H}_3 \\ \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array} + \begin{array}{c} \mathbf{H}_4 \\ \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array} + \begin{array}{c} \mathbf{H}_5 \\ \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

5 Derived networks

We can use the multiplication to obtain new networks from existing *compatible* two-mode networks. For example, from basic bibliographic networks **WA** and **WK** we get

$$\mathbf{AK} = \mathbf{WA}^T \cdot \mathbf{WK}$$

a network relating authors to keywords used in their works, and

$$\mathbf{Ca} = \mathbf{WA}^T \cdot \mathbf{Ci} \cdot \mathbf{WA}$$

is a network of citations between authors.

Networks obtained from existing networks using some operations are called *derived* networks. They are very important in analysis of collections of *linked* networks.

What is the meaning of the product network? In general we could consider weights, addition and multiplication over a selected semiring (Cerinšek and Batagelj, 2017). In this paper we will limit our attention to the traditional addition and multiplication of real numbers.

The weight $\mathbf{AK}[a, k]$ is equal to the number of times the author a used the keyword k in his/her works.

The weight $\mathbf{Ca}[a, b]$ counts the number of times a work authored by the author a is citing a work authored by the author b; or shorter, how many times the author a cited the author b.

Using network multiplication we can also transform a given two-mode network, for example **WA**, into corresponding ordinary one-mode networks (*projections*)

$$\mathbf{WW} = \mathbf{WA} \cdot \mathbf{WA}^T \qquad \text{and} \qquad \mathbf{AA} = \mathbf{WA}^T \cdot \mathbf{WA}$$

The obtained projections can be analyzed using standard network analysis methods. This is a traditional recipe how to analyze two-mode networks. Often the weights are not considered in the analysis; and when they are considered we have to be very careful about their meaning.

The weight $\mathbf{WW}[p, q]$ is equal to the number of common authors of works p and q.

The weight $\mathbf{AA}[a, b]$ is equal to the number of works that author a and b coauthored. In a special case when $a = b$ it is equal to the number of works that the author a wrote. The network $\mathbf{AA}$ is describing the *coauthorship* (collaboration) between authors and is also denoted as $\mathbf{Co}$ – the "first" coauthorship network.

In the paper Batagelj and Cerinšek (2013) it was shown that there can be problems with the network $\mathbf{Co}$ when we try to use it for identifying the most collaborative authors. By the outer product decomposition the coauthorship network $\mathbf{Co}$ is composed of complete subgraphs on the set of work's coauthors. Works with many authors produce large complete subgraphs, thus bluring the collaboration structure, and are over-represented by its total weight. To see this, 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

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

In general each term $\mathbf{H}_w$ in the outer product decomposition of the product $\mathbf{C}$ has different total weight $T(\mathbf{H}_w) = \sum_{a,k} (\mathbf{H}_w)_{ak}$ leading to over-representation of works with large values. In the case of coautorship network $\mathbf{Co}$ we have $S(\mathbf{WA}[w, .]) = \text{outdeg}_{\mathbf{WA}}(w)$ and therefore $T(\mathbf{H}_w) = \text{outdeg}_{\mathbf{WA}}(w)^2$. To resolve the problem we apply the fractional approach.

6 Fractional approach

To make the contributions of all works 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 $T(\mathbf{H}'_w) = 1$ for all works w.

In the case of two-mode networks $\mathbf{WA}$ and $\mathbf{WK}$ we denote

$$S_w^{\mathbf{WA}} = \begin{cases} \sum_a \mathbf{WA}[w, a] & \text{outdeg}_{\mathbf{WA}}(w) > 0 \\ 1 & \text{outdeg}_{\mathbf{WA}}(w) = 0 \end{cases}$$

(and similarly $S_w^{\mathbf{WK}}$) and define the *normalized* matrices

$$\mathbf{WAn} = \text{diag}(\frac{1}{S_w^{\mathbf{WA}}}) \cdot \mathbf{WA}, \quad \mathbf{WKn} = \text{diag}(\frac{1}{S_w^{\mathbf{WK}}}) \cdot \mathbf{WK}$$

In real life networks $\mathbf{WA}$ (or $\mathbf{WK}$) it can happen that some work has no author. In such a case $S_w^{\mathbf{WA}} = \sum_a \mathbf{WA}[w, a] = 0$ which makes problems in the definition of the normalized network $\mathbf{WAn}$. We can bypass the problem by setting $S_w^{\mathbf{WA}} = 1$, as we did in the above definition.

Then the *normalized product* matrix is

$$\mathbf{AKt} = \mathbf{WAn}^T \cdot \mathbf{WKn}$$

Denoting $\mathbf{F}_w = \frac{1}{S_w^{\mathbf{WA}} S_w^{\mathbf{WK}}} \mathbf{H}_w$ the outer product decomposition gets form

$$\mathbf{AKt} = \sum_w \mathbf{F}_w$$

Since

$$T(\mathbf{F}_w) = \begin{cases} 1 & (\text{outdeg}_{\mathbf{WA}}(w) > 0) \wedge (\text{outdeg}_{\mathbf{WK}}(w) > 0) \\ 0 & \text{otherwise} \end{cases}$$

we have further

$$\sum_{a,k} \mathbf{F}[a,k] = \sum_{a,k} \sum_w \mathbf{F}_w[a,k] = \sum_w T(\mathbf{F}_w) = |W^+|$$

where $W^+ = \{w \in W : (\text{outdeg}_{\mathbf{WA}}(w) > 0) \wedge (\text{outdeg}_{\mathbf{WK}}(w) > 0)\}$.

In the network $\mathbf{AKt}$ the contribution of each work to the bibliography is 1. These contributions are redistributed to arcs from authors to keywords.

Example B: For matrices from Example A we get the corresponding diagonal normalization matrices

$$\text{diag}(\frac{1}{S_w^{\mathbf{WA}}}) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/3 \end{bmatrix} \end{array}$$

$$\text{diag}(\frac{1}{S_w^{\mathbf{WK}}}) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \end{bmatrix} \end{array}$$

compute the normalized matrices

$$\mathbf{WAn} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \begin{bmatrix} 1/2 & 0 & 1/2 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 1/3 & 0 & 1/3 & 1/3 \end{bmatrix} \end{array}, \quad \mathbf{WKn} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{bmatrix} \end{array},$$

outer products such as

$$\mathbf{F}_1 = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{F}_5 = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 0 & 1/6 & 0 & 1/6 \\ 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 1/6 \\ 0 & 1/6 & 0 & 1/6 \end{bmatrix} \end{array}$$

and finally the product matrix

$$\mathbf{AKt} = \mathbf{WAn}^T \cdot \mathbf{WKn} = \sum_{w=1}^{5} \mathbf{F}_w = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} 0.50000 & 0.52778 & 0.36111 & 0.27778 \\ 0.25000 & 0.00000 & 0.75000 & 0.00000 \\ 0.25000 & 0.52778 & 0.11111 & 0.27778 \\ 0.00000 & 0.27778 & 0.61111 & 0.27778 \end{bmatrix} \end{array}$$

6.1 Linking through a network

Let a network $\mathbf{S}$ links works to works. The derived network $\mathbf{WA}^T \cdot \mathbf{S} \cdot \mathbf{WA}$ links authors to authors *through* $\mathbf{S}$. Again, the normalization question has to be addressed. Among different options let us consider the derived networks defined as:

$$\mathbf{C} = \mathbf{WAn}^T \cdot \mathbf{S} \cdot \mathbf{WAn}$$

It is easy to verify that:

- if $\mathbf{S}$ is symmetric, $\mathbf{S}^T = \mathbf{S}$, then also $\mathbf{C}$ is symmetric, $\mathbf{C}^T = \mathbf{C}$;

$$\mathbf{C}^T = (\mathbf{WAn}^T \cdot \mathbf{S} \cdot \mathbf{WAn})^T = \mathbf{WAn}^T \cdot \mathbf{S}^T \cdot (\mathbf{WAn}^T)^T = \mathbf{C}$$

- if $W^+ = \{w \in W : \text{outdeg}_{\mathbf{WA}}(w) > 0\} = W$, the total of weights of $\mathbf{S}$ is redistributed in $\mathbf{C}$:

$$T(\mathbf{C}) = \sum_{e \in L(\mathbf{C})} c(e) = \sum_{e \in L(\mathbf{S})} s(e) = T(\mathbf{S})$$

Since $\sum_{a \in A} wa[p,a] = \text{outdeg}_{\mathbf{WA}}(p)$ and $wan[p,a] = \begin{cases} \frac{wa[p,a]}{\text{outdeg}_{\mathbf{WA}}(p)} & \text{outdeg}_{\mathbf{WA}}(p) > 0 \\ 0 & \text{otherwise} \end{cases}$ we get

$$T(\mathbf{C}) = \sum_{e \in L(\mathbf{C})} c(e) = \sum_{a \in A} \sum_{b \in A} c[a,b] = \sum_{a \in A} \sum_{b \in A} \sum_{p \in W} \sum_{q \in W} wan[p,a] \cdot s[p,q] \cdot wan[q,b] =$$

$$= \sum_{p \in W^+} \sum_{q \in W^+} \frac{s[p,q]}{\text{outdeg}_{\mathbf{WA}}(p)\text{outdeg}_{\mathbf{WA}}(q)} \sum_{a \in A} wa[p,a] \sum_{b \in A} wa[q,b] = \sum_{p \in W^+} \sum_{q \in W^+} s[p,q]$$

and finally, if $W^+ = W$

$$\sum_{p \in W^+} \sum_{q \in W^+} s[p,q] = \sum_{e \in L(\mathbf{S})} s(e) = T(\mathbf{S})$$

As special cases we get for normalized author's citation networks with $W^+ = W$: for $\mathbf{S} = \mathbf{Ci}$

$$\sum_{a\in A}\sum_{b\in A} c[a,b] = \sum_{p\in W}\sum_{q\in W} ci[p,q] = |\mathbf{Ci}|$$

and for $\mathbf{S} = \mathbf{Cin}$

$$\sum_{a\in A}\sum_{b\in A} c[a,b] = \sum_{p\in W}\sum_{q\in W:\ \mathbf{outdeg_{Ci}}(q)>0} \frac{ci[p,q]}{\text{outdeg}_{\mathbf{Ci}}(p)} = \sum_{q\in W:\ \mathbf{outdeg_{Ci}}(q)>0} 1 = W^+_{\mathbf{Ci}}$$

6.2 Some notes

A. Instead of computing the normalized network $\mathbf{WAn}$ from the network $\mathbf{WA}$ we could collect the data about the real proportion $wan[w,a]$ of the contribution of each author a to a work w such that $\mathbf{WAn}$ is normalized: for every work w it holds

$$\sum_{a\in A} wan[w,a] \in \{0,1\}$$

Unfortunately in most cases such data are not available and we use the computed normalized weights as their estimates. Most of the results do not depend on the way the normalized network was obtained.

B. In general a given network matrix $\mathbf{WA}$ can be normalized in two ways: *by rows*, as used in this section, and *by columns*

$$\mathbf{WAn}' = \mathbf{WA}\cdot \text{diag}(\frac{1}{S_a^{\mathbf{WA}}}) \quad \text{where} \quad S_a^{\mathbf{WA}} = \begin{cases} \sum_w \mathbf{WA}[w,a] & \text{indeg}_{\mathbf{WA}}(a) > 0 \\ 1 & \text{indeg}_{\mathbf{WA}}(a) = 0 \end{cases}$$

In the context of bibliographic networks its meaning does not make much sense.

C. The network $\mathbf{Co}$ is symmetric: $co_{ab} = co_{ba}$. We need to compute only half of values co_{ab}, $a \le b$. The resulting network is undirected with weights co_{ab}.

D. In the paper Batagelj and Cerinšek (2013) the "second" coauthorship network $\mathbf{Cn} = \mathbf{WA}^T\cdot \mathbf{WAn}$ is considered. The weight cn_{ab} is equal to the contribution of an author a to works that (s)he wrote together with the author b. Using these weights the *selfsufficiency* of an author a is defined as:

$$S_a = \frac{cn_{aa}}{\text{indeg}_{\mathbf{WA}}(a)}$$

and *collaborativness* of an author a as its complementary measure $K_a = 1 - S_a$.

E. In the "third" coauthorship network $\mathbf{Cn} = \mathbf{WAn}^T \cdot \mathbf{WAn}$ the weight ct_{ab} is equal to the total fractional contribution of 'collaboration' of authors a and b to works. Each work w with $S_w^{\mathbf{WA}} > 0$ contributes 1 to the total of weights in $\mathbf{Cn}$. This is the network to be used in analysis of collaboration between authors (Batagelj and Cerinšek, 2013; Leydesdorff and Park, 2017; Prathap and Mukherjee, 2016). To identify the most collaborative groups we can use methods such as P_S-cores and link islands (Batagelj et al., 2014).

The product $\mathbf{Cn}$ is symmetric. Note $\mathbf{C}$ applies. We transform it to the corresponding undirected network – pairs of opposite arcs are replaced by an edge with doubled weight. In analyses we usually analyze separately the vector of weights on loops (selfcontribution) and the network $\mathbf{Cn}$ without loops.

F. An alternative normalization $\mathbf{WAn}'$ of a binary autorship matrix $\mathbf{WA}$ was proposed in Newman (2004)

$$wan'_{wa} = \frac{wa_{wa}}{\max(1, \text{outdeg}_{\mathbf{WA}}(w) - 1)}$$

in which only collaboration with coauthors is considered – no selfcollaboration. Note that using the network construction proposed on page 5 of Newman (2001) we get a network in which works with many coauthors are still over-represented. The same idea is used in the fractional counting co-authorship matrix $\mathbf{U}^*$ proposed in equation (5) in Perianes-Rodriguez et al. (2016).

To treat all works equally using the Newman's normalization the "fourth" coauthorship network was proposed in Cerinšek and Batagelj (2015). To compute it we first compute

$$\mathbf{Ct}' = \mathbf{WAn}^T \cdot \mathbf{WAn}'$$

The weight ct'_{ab} is equal to the total contribution of "strict collaboration" of authors a and b to works. The obtained product is symmetric. Again note $\mathbf{C}$ applies. We transform it to the corresponding undirected network – pairs of opposite arcs are replaced by an edge with doubled weight. The loops are removed. The contribution of each work with at least two coauthors is equal to 1. A kind of the outer product decomposition exists also for the network $\mathbf{Ct}'$ with a diagonal set to 0.

7 Bibliographic Coupling and Co-citation

glej Ludo

Bibliographic coupling occurs when two works each cite a third work in their bibliographies, see Figure 2, left. The idea was introduced by Kessler (1963) and has been used extensively since then. See figure where two citing works, p and q, are shown. Work p cites five works and q cites seven works. The key idea is that there are three works cited by both p and q. This suggests some content communality for the three works cited by both p and q. Having more works citing pairs of prior works increases the likelihood of them sharing content.

We assume that the citation relation means p $\mathbf{Ci}$ $q \equiv$ work p cites work q. Then the *bibliographic coupling* network $\mathbf{biCo}$ can be determined as

$$\mathbf{biCo} = \mathbf{Ci} * \mathbf{Ci}^T$$

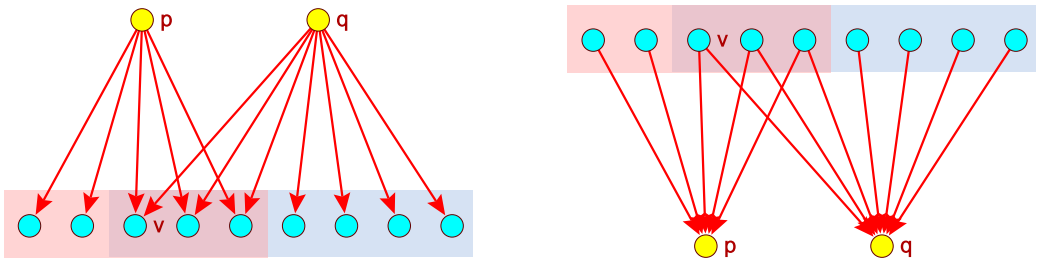

Figure 2: Bibliographic coupling (left) and Co-citation (right)

The weight $bico_{pq}$ is equal to the number of works cited by both works p and q; $bico_{pq} = |\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|$. Bibliographic coupling weights are symmetric: $bico_{pq} = bico_{qp}$:

$$\mathbf{biCo}^T = (\mathbf{Ci} \cdot \mathbf{Ci}^T)^T = \mathbf{Ci} \cdot \mathbf{Ci}^T = \mathbf{biCo}$$

Co-citation is a concept with strong parallels with bibliographic coupling (Small and Marshakova 1973), see Figure 2, right. The focus is on the extent to which works are co-cited by later works. The basic intuition is that the more earlier works are cited, the higher the likelihood that they have common content. The *co-citation* network $\mathbf{coCi}$ can be determined as

$$\mathbf{coCi} = \mathbf{Ci}^T \cdot \mathbf{Ci}.$$

The weight $coci_{pq}$ is equal to the number of works citing both works p and q. The network $\mathbf{coCi}$ is symmetric $coci_{pq} = coci_{qp}$:

$$\mathbf{coCi}^T = (\mathbf{Ci}^T \cdot \mathbf{Ci})^T = \mathbf{Ci}^T \cdot \mathbf{Ci} = \mathbf{coCi}$$

An important property of co-citation is that $\mathbf{coCi}(\mathbf{Ci}) = \mathbf{biCo}(\mathbf{Ci}^T)$:

$$\mathbf{biCo}(\mathbf{Ci}^T) = \mathbf{Ci}^T \cdot (\mathbf{Ci}^T)^T = \mathbf{Ci}^T \cdot \mathbf{Ci} = \mathbf{coCi}(\mathbf{Ci})$$

Therefore the constructions proposed for bibliographic coupling can be applied also for co-citation.

What about normalizations? Searching for the most coupled works we have again problems with works with many citations, especially with review papers. To neutralize their impact we can introduce normalized measures. The fractional approach works fine for normalized co-citation

$$\mathbf{CoCit} = \mathbf{Cin}^T \cdot \mathbf{Cin}$$

where $\mathbf{Cin} = \mathbf{D} \cdot \mathbf{Ci}$ and $\mathbf{D} = \text{diag}(\frac{1}{\max(1,\text{outdeg}(p))})$. $\mathbf{D}^T = \mathbf{D}$. In the normalized network every work has value 1 and it is equally distributed to all cited works.

The fractional approach can not bi directly applied to bibliographic coupling – to get the outer product decomposition work we would need to normalize $\mathbf{Ci}$ by columns – a cited work has value 1 which is distributed equally to the citing works – the most cited works give the least. This is against our intuition. To construct a reasonable measure we can proceed as follows. Let us first look at

$$\mathbf{biC} = \mathbf{Cin} \cdot \mathbf{Ci}^T$$

we have

$$\mathbf{biC} = (\mathbf{D} \cdot \mathbf{Ci}) \cdot \mathbf{Ci}^T = \mathbf{D} \cdot \mathbf{biCo}$$

$$\mathbf{biC}^T = (\mathbf{D} \cdot \mathbf{biCo})^T = \mathbf{biCo}^T \cdot \mathbf{D}^T = \mathbf{biCo} \cdot \mathbf{D}$$

For $\mathbf{Ci}(p) \neq \emptyset$ and $\mathbf{Ci}(q) \neq \emptyset$ it holds

$$\mathbf{biC}_{pq} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(p)|} \quad \text{and} \quad \mathbf{biC}_{qp} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(q)|} = \mathbf{biC}_{pq}^T$$

and $\mathbf{biC}_{pq} \in [0, 1]$. $\mathbf{biC}_{pq}$ is the proportion of its references that the work p shares with the work q. The network $\mathbf{biC}$ is not symmetric. We have different options to construct normalized symmetric measures such as

$$\mathbf{biCoa}_{pq} = \frac{1}{2}(\mathbf{biC}_{pq} + \mathbf{biC}_{qp}) \quad \text{Average}$$

$$\mathbf{biCom}_{pq} = \min(\mathbf{biC}_{pq}, \mathbf{biC}_{qp}) \quad \text{Minimum}$$

$$\mathbf{biCoM}_{pq} = \max(\mathbf{biC}_{pq}, \mathbf{biC}_{qp}) \quad \text{Maximum}$$

or, may be more interesting

$$\mathbf{biCog}_{pq} = \sqrt{\mathbf{biC}_{pq} \cdot \mathbf{biC}_{qp}} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{\sqrt{|\mathbf{Ci}(p)| \cdot |\mathbf{Ci}(q)|}} \quad \begin{matrix}\text{Geometric mean} \\ \text{Salton cosinus}\end{matrix}$$

$$\mathbf{biCoh}_{pq} = 2 \cdot (\mathbf{biC}_{pq}^{-1} + \mathbf{biC}_{qp}^{-1})^{-1} = \frac{2|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(p)| + |\mathbf{Ci}(q)|} \quad \text{Harmonic mean}$$

$$\mathbf{biCoj}_{pq} = (\mathbf{biC}_{pq}^{-1} + \mathbf{biC}_{qp}^{-1} - 1)^{-1} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(p) \cup \mathbf{Ci}(q)|} \quad \text{Jaccard index}$$

All these measures are similarities.

It is easy to verify that $biCoX_{pq} \in [0, 1]$ and: $biCoX_{pq} = 1$ iff the works p and q are referencing the same works, $\mathbf{Ci}(p) = \mathbf{Ci}(q)$.

From $m \leq H \leq G \leq A \leq M$ and $J \leq m$, ($\frac{|P \cap Q|}{|P \cup Q|} \leq \min(\frac{|P \cap Q|}{|P|}, \frac{|P \cap Q|}{|Q|})$) we get

$$\mathbf{biCoj}_{pq} \leq \mathbf{biCom}_{pq} \leq \mathbf{biCoh}_{pq} \leq \mathbf{biCog}_{pq} \leq \mathbf{biCoa}_{pq} \leq \mathbf{biCoM}_{pq}$$

The equalities hold iff $\mathbf{Ci}(p) = \mathbf{Ci}(q)$.

To get a dissimilarity we can use transformations $dis = 1 - sim$ or $dis = \frac{1}{sim} - 1$ or $dis = -\log sim$. For example

$$\mathbf{biCod}_{pq} = 1 - \mathbf{biCoj}_{pq} = \frac{|\mathbf{Ci}(p) \oplus \mathbf{Ci}(q)|}{|\mathbf{Ci}(p) \cup \mathbf{Ci}(q)|} \quad \text{Jaccard distance}$$

where $\oplus$ denotes the symmetric difference of sets.

Bibliographic coupling and co-citation networks are linking works to works. To get linking between authors, journals or keywords considering citation similarity we can apply the construction from Subsection 6.1 to the normalized co-citation or bibliographic coupling network.

8 Conclusions

In the paper we presented an attempt to provide a foundation of fractional approach to biblimetric networks based on the outer product decomposition of product networks. We also discussed the fractional approach to bibliographic coupling and co-citation networks. The results of application of the proposed methods to real bibliographic data will be presented in separate papers.

All described computations can be done efficiently in program Pajek (De Nooy et al., 2018) using macros such us: `norm1` – normalized 1-mode network, `norm2` – normalized 2-mode network, `norm2p` – Newman's normalization of a 2-mode network, `biCo` – bibliographic coupling network, and `biCon` – normalized bibliographic coupling network, available at GitHub (Batagelj, 2018).

Acknowledgments

The paper is based on presentations on 1274. Sredin seminar, IMFM, Ljubljana, 29. March 2017; NetGloW 2018, St Petersburg, July 4-6, 2018; and COMPSTAT 2018, Iasi, Romania, August 28-31, 2018.

This work is supported in part by the Slovenian Research Agency (research program P1-0294 and research projects J1-9187, J7-8279 and BI-US/17-18-045), COST project CRoNoS and by Russian Academic Excellence Project '5-100'.

References

Batagelj, V. (2018). Github: **biblio** – Bibliographic network analysis. https://github.com/bavla/biblio

Batagelj, V. (2007) WoS2Pajek. Networks from Web of Science. Version 1.5 (2017). http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek

Batagelj, V., Cerinšek, M. (2013). On bibliographic networks. Scientometrics 96 (3), 845-864.

Batagelj, V., Doreian P., V., Ferligoj, A., Kejžar N. (2014). Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution.

Carley K.M. (2003). Dynamic Network Analysis. in the Summary of the NRC workshop on Social Network Modeling and Analysis, Ron Breiger and Kathleen M. Carley (Eds.), National Research Council, p. 133–145.

Cerinšek, M., Batagelj, V. (2017). Semirings and Matrix Analysis of Networks. Encyclopedia of Social Network Analysis and Mining. Reda Alhajj, Jon Rokne (Eds.), Springer, New York.

Cerinšek, M., Batagelj, V. (2015). Network analysis of Zentralblatt MATH data. Scientometrics, 102 (1), 977-1001.

Clarivate Analytics (2018). https://clarivate.com/products/web-of-science/databases/

De Nooy, W., Mrvar, A., Batagelj, V. (2018). Exploratory Social Network Analysis with Pajek; Revised and Expanded Edition for Updated Software. Structural Analysis in the Social Sciences, CUP.

Diesner, J., Carley, K.M. (2004). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a Novel Method for Network Text Analysis. Chapter 4 in Causal Mapping for Research in Information Technology, V.K. Narayanan and Deborah J. Armstrong, eds. Idea Group Inc., 2005, p. 81-108.

Gauffriau, M. (2017). A categorization of arguments for counting methods for publication and citation indicators. Journal of Informetrics, 11(3), 672-684.

Kessler, M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1): 10–25.

Krackhardt, D., Carley, K.M. (1998). A PCANS Model of Structure in Organization. In Proceedings of the 1998 International Symposium on Command and Control Research and Technology Evidence Based Research: 113-119, Vienna, VA.

Leydesdorff, L., Park, H.W. (2016). Full and Fractional Counting in Bibliometric Networks. Journal of Informetrics Volume 11, Issue 1, February 2017, Pages 117–120.

Lindsey, D. (1980). Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship. Social Studies of Science, 10(2), 145–162.

Marshakova, I. (1973). System of documentation connections based on references (sci). Nauchno-Tekhnicheskaya Informatsiya Seriya, 2(6): 3–8.

Newman, M.E.J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 64(1), 016132.

Newman, M.E.J. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America 101(Suppl1), 5200–5205.

Perianes-Rodriguez, A., Waltman, L., Van Eck, N.J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics, 10(4), 1178-1195.

Prathap, G., Mukherjee, S. (2016). A conservation rule for constructing bibliometric network matrices. arXiv 1611.08592

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24(4): 265–269.