On bibliographic networks

Vladimir Batagelj*, Monika Cerišek**

University of Ljubljana

Abstract

In the paper we show that the bibliographic data can be transformed into a collection of compatible networks. Using network multiplication different interesting derived networks can be obtained. In defining them an appropriate normalization should be considered. The proposed approach can be applied also to other collections of compatible networks. We also discuss the question when the multiplication of sparse networks preserves sparseness. The proposed approaches are illustrated with analyses of collection of networks on the topic ”social network” obtained from the Web of Science.

Keywords: co-authorship, collaboration, two-mode network, network multiplication, sparse network, normalization

2010 MSC: 91D30, 62H30, 68W40, 93A15

1. Introduction

A collaboration network is usually defined in the following way. The set of network’s nodes consists of authors. There exists an edge (undirected link) between authors u and v iff they produced a joint work (paper, book, report, etc.). Its weight \( w(u,v) \) is equal to the number of works to which u and v both contributed.

In this case a more basic network is a two-mode network linking the set of works with the set of authors. There is an arc (directed link) from the work
p to the author u iff u is an author of the work p. It is well known that if we represent this two-mode network with a matrix $W_A$ then we can compute the matrix of the corresponding collaboration network as $W_A^T \ast W_A$ using matrix multiplication.

The problem with matrices of large networks is that they require in their standard representation too much computer memory although most of their entries are zero. For this reason we introduce a ‘parallel’ operation of network multiplication that deals only with nonzero elements.

For a given set of works, besides the two-mode network $W_A$ on works $\times$ authors, we can construct other two-mode networks such as $W_K$ on works $\times$ keywords, $W_C$ on works $\times$ classifications, $W_J$ on works $\times$ journals, etc. Since these networks have the same first set – the set of works, we can obtain from them using multiplication different derived networks. For example $W_A^T \ast W_K = A_K$ gives us the two-mode network $A_K$ on authors $\times$ keywords with the weight of the arc $(u, k)$ counting in how many works the author $u$ used the keyword $k$. Additional derived networks can be produced considering also the one-mode citation network $C_i$ between works.

In the paper we first show that we can transform any data table into a collection of corresponding two-mode networks. Afterwards we introduce the network multiplication and discuss the question when it preserves the sparsity of networks. Since the networks from the collection are compatible – they share a common set – we can obtain, using multiplication, different derived networks. The main part of the paper deals with the problem of ‘normalization’ of the weights in the derived networks which is illustrated with the case of collaboration networks. The described approach can be used also for other derived networks. In the last part of the paper some other derived networks for the case of bibliographic networks are presented.

The introduced concepts are illustrated on the network data set SN5 obtained in 2008 from the Web of Science for a query "social network*" and expanded with existing descriptions of the most frequent references and the bibliographies of around 100 social networkers. Using the program WoS2Pajek (Batagelj, 2007) the corresponding collection of network data was produced: the networks works $\times$ authors, works $\times$ keywords, ..., citation network; partition of works by publication year, and the DC partition distinguishing between works with complete description and the cited only works. The sizes of the sets are as follows: works $|W| = 193376$, works with complete description $|C| = 7950$, authors $|A| = 75930$, journals $|J| = 14651$, keywords $|K| = 29267$. The data set was used for the Viszards session at the
Analyses were made in a program Pajek (Pajek wiki (2012)), a tool for analysis and visualization of large networks.

2. Two-mode networks and network multiplication

2.1. Two-mode networks from data tables

A data table $\mathcal{T}$ is a set of records $\mathcal{T} = \{ T_k : k \in K \}$, where $K$ is the set of keys. A record has the form $T_k = (k, q_1(k), q_2(k), \ldots, q_r(k))$ where $q_i(k)$ is the value of the property (attribute) $q_i$ for the key $k$.

Suppose that the property $q$ has the range $2^Q$. For example, for Wasserman and Faust (1994):

| work     | authors                        | year |
|----------|--------------------------------|------|
| SNA      | S. Wasserman, K. Faust         | 1994 |
| GenCores | V. Batagelj, M. Zaveršnik      | 2011 |
| Islands  | M. Zaveršnik, V. Batagelj      | 2004 |
| ESNA2    | W. de Nooy, A. Mrvar, V. Batagelj | 2012 |
| IFCS09   | N. Kejžar, S. Korenjak, V. Batagelj | 2010 |

Here work is a key, and authors and year are properties.

If $Q$ is finite we can assign to the property $q$ a two-mode network $K \times q = (K, Q, \mathcal{A}, w)$ where $(k, v) \in \mathcal{A}$ iff $v \in q(k)$, and $w(k, v) = 1$. Note that the set $Q$ can always be transformed into a finite set by partitioning it and recoding the values.

Single-valued properties can be represented more compactly by a partition.

For data from the Web of Science (Knowledge) we can obtain the corresponding networks using the program WoS2Pajek (Batagelj, 2007). Similar
programs exist also for other bibliographic data sources/formats: BiBT\TeX, DBPL, IMDB, Zentralblatt Math, and others.

2.2. Multiplication of networks

The product of two compatible networks is essentially the network corresponding to the product of matrices corresponding to the given networks; or in more formal words:

To a simple (no parallel arcs) two-mode network \( N = (I, J, A, w) \); where \( I \) and \( J \) are sets of nodes, \( A \) is a set of arcs linking \( I \) and \( J \), and \( w : A \rightarrow \mathbb{R} \) is a weight; we can assign a network matrix \( W = [w_{i,j}] \) with elements: 

\[
w_{i,j} = w(i, j) \quad \text{for} \quad (i, j) \in A \quad \text{and} \quad w_{i,j} = 0 \quad \text{otherwise}.
\]

Given a pair of compatible two-mode networks \( N_A = (I, K, A_A, w_A) \) and \( N_B = (K, J, A_B, w_B) \) with corresponding matrices \( A_{I \times K} \) and \( B_{K \times J} \) we call a product of networks \( N_A \) and \( N_B \) a network \( N_C = (I, J, A_C, w_C) \), where \( A_C = \{(i, j) : i \in I, j \in J, c_{i,j} \neq 0\} \) and \( w_C(i, j) = c_{i,j} \) for \( (i, j) \in A_C \). The product matrix \( C = [c_{i,j}]_{I \times J} = A * B \) is defined in the standard way

\[
c_{i,j} = \sum_{k \in K} a_{i,k} \cdot b_{k,j}
\]

In some applications we have to consider the product on other semirings than the standard \((\mathbb{R}, +, \cdot, 0, 1)\) (Batagelj, 1994).

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

Note that in the expression (1) to the value \( c_{i,j} \) contribute only the terms \( a_{i,k} \cdot b_{k,j} \) in which both factors \( a_{i,k} \) and \( b_{k,j} \) are nonzero. For \( N_A(i) \cup N_B^{-}(j) \neq \emptyset \) we have

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

where \( N_A(i) \) are the successors of node \( i \) in network \( N_A \) and \( N_B^{-}(j) \) are the predecessors of node \( j \) in network \( N_B \).

Therefore, if all weights in networks \( N_A \) and \( N_B \) are equal to 1 then the product \( a_{i,k} \cdot b_{k,j} \in \{0, 1\} \) and the value of \( c_{i,j} \) counts the number of ways we can go from \( i \in I \) to \( j \in J \) passing through \( K \).

The standard matrix multiplication has the complexity \( O(|I| \cdot |K| \cdot |J|) \) – it is too slow to be used for large networks. Most of large networks are sparse – their matrices contain much more zero elements than nonzero elements. For sparse large networks we can multiply much faster considering only nonzero elements.
for $k$ in $\mathcal{K}$ do
    for $(i, j)$ in $\mathcal{N}_A^- (k) \times \mathcal{N}_B (k)$ 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.

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
$K_{\mathcal{N}_A^- (k), \mathcal{N}_B (k)}$ (or, in the case $\mathcal{I} = \mathcal{J}$, a complete subgraph $K_{\mathcal{N}_A (k)}$). If both
degrees $\deg_A (k) = |\mathcal{N}_A^- (k)|$ and $\deg_B (k) = |\mathcal{N}_B (k)|$ are large then already
the computation of this complete subgraph has a quadratic (time and space)
complexity – the result ‘explodes’.

It is easy to see that if at least one of the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$
has small maximal degree on $\mathcal{K}$ then also the resulting product network $\mathcal{N}_C$
is sparse.

We shall prove a stronger result that if for the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$
there are in $\mathcal{K}$ only some vertices 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.
Let
\[ d_{\text{min}}(k) = \min(\deg_A(k), \deg_B(k)) \quad \text{and} \quad d_{\text{max}}(k) = \max(\deg_A(k), \deg_B(k)). \]
Then
\[ \deg_A(k) \cdot \deg_B(k) = d_{\text{min}}(k) \cdot d_{\text{max}}(k) \]
Define also \( \Delta_{\text{min}} = \max_{k \in K} d_{\text{min}}(k) \) and
\[ K(d) = \{ k \in K : d_{\text{max}}(k) \geq d \} \]
Let us denote \( d^* = \arg\min_{d}(|K(d)| \leq d) \) and \( K^* = K(d^*) \). Then \( |K^*| \leq d^* \) and the number of nonzero elements in the product
\[ C \leq \sum_{k \in K} \deg_A(k) \cdot \deg_B(k) = \sum_{k \in K} d_{\text{min}}(k) \cdot d_{\text{max}}(k) \]
\[ = \sum_{k \in K^*} d_{\text{min}}(k) \cdot d_{\text{max}}(k) + \sum_{k \in K \setminus K^*} d_{\text{min}}(k) \cdot d_{\text{max}}(k) \]
\[ \leq \Delta_{\text{min}} \cdot \sum_{k \in K^*} d_{\text{max}}(k) + d^* \cdot \sum_{k \in K \setminus K^*} d_{\text{min}}(k) \]
\[ \leq d^* \cdot (\Delta_{\text{min}} \cdot \max(|I|, |J|) + \min(|A_A|, |A_B|)) \]
Therefore:
If for the sparse networks \( N_A \) and \( N_B \) the quantities \( \Delta_{\text{min}} \) and \( d^* \) are small then also the resulting product network \( N_C \) is sparse.

That is equivalent to the claimed result.

3. Collaboration

3.1. Co-authorship networks
Let \( WA \) be the works \( \times \) authors two-mode co-authorship network; \( wa_{pi} \in \{0, 1\} \) is describing the authorship of author \( i \) of work \( p \). Then for each work \( p \in W \):
\[ \sum_{i \in A} wa_{pi} = \text{outdeg}(p) \]
The outdeg\( (p) \) is equal to the number of authors of work \( p \).
Let $N$ be its normalized version with $n_{pi}$ describing the share of contribution of author $i$ to work $p$ such that for each work $p \in W$:

$$\sum_{i \in A} n_{pi} \in \{0, 1\}$$

The sum has value 0 for works without authors.

The contributions $n_{pi}$ can be determined by some rules or, assuming that each author contributed equally to the work, it can be computed from $WA$ as

$$n_{pi} = \frac{wa_{pi}}{\max(1, \text{outdeg}(p))}.$$ 

A similar normalization of collaboration links, but with $\text{outdeg}(p) - 1$ instead of $\text{outdeg}(p)$, was proposed already by Newman (2001). He is interpreting the weight as a proportion of time spent for the collaboration with each co-author.

Row-normalization $n(N)$ is a network obtained from $N$ in which the weight of each arc $a$ is divided by the sum of weights of all arcs having the same initial node as the arc $a$. For binary network $A$ on $I \times J$

$$n(A) = \text{diag} \left( \frac{1}{\max(1, \text{outdeg}(i))} \right)_{i \in I} \ast A$$

Therefore we can obtain the normalized co-authorship network as

$$N = n(WA)$$

In some sense reverse transformation is the binarization $b(N)$ of the $N$: it is the original network in which all weights are set to 1. It holds

$$WA = b(N)$$

and if $N$ was obtained from $WA$ also $WA = b(n(WA))$.

Another useful transformation is the transposition. Transposition $N^T$ or $t(N)$ is a network obtained from $N$ in which to all arcs their direction is reversed. For bibliographic networks we introduce the abbreviations $AW = WA^T$, $KW = WK^T$, etc.
3.2. The first collaboration network

A standard way to obtain the collaboration network \( \mathbf{Co} \) from the co-authorship network using network multiplication is

\[
\mathbf{Co} = \mathbf{A} \mathbf{W} \ast \mathbf{W} \mathbf{A}
\]

From

\[
co_{ij} = \sum_{p \in W} wa_{pi} wa_{pj} = \sum_{p \in N(i) \cap N(j)} 1
\]

we see that \( co_{ij} \) is equal to the number of works that authors \( i \) and \( j \) wrote together.

The weights in the first collaboration network are symmetric

\[
co_{ij} = \sum_{p \in W} wa_{pi} wa_{pj} = \sum_{p \in W} wa_{pj} wa_{pi} = co_{ji}
\]

One can search for authors with most collaborators. Such authors in the set \( \text{SN5} \) are listed in Table 1. On the other hand Table 2 shows the distribution of output degree of authors in the set \( \text{SN5} \). Output degree of each author is equal to the number of works he/she co-authored.

Table 1: List of the authors with the largest number of different collaborators in \( \text{SN5} \)

| i  | author          | collaborators |
|----|-----------------|---------------|
| 1  | Snijders,T      | 77            |
| 2  | Krackhardt,D    | 71            |
| 3  | Wasserman,S     | 65            |
| 4  | Ferligoj,A      | 63            |
| 5  | Berkman,L       | 63            |
| 6  | van Duijn,M     | 63            |
| 7  | Donovan,D       | 62            |
| 8  | Friedman,S      | 60            |
| 9  | Latkin,C        | 59            |
| 10 | Faust,K         | 59            |

The obvious question is: who are the most collaborative authors? The standard answer is provided by \( k \)-cores, Batagelj and Zaveršnik (2011).

A subset \( \mathcal{U} \subseteq \mathcal{V} \) of nodes determines a \( k \)-core \( \mathcal{C} = (\mathcal{U}, \mathcal{L}\mathcal{U}) \) in the network \( \mathcal{N} = (\mathcal{V}, \mathcal{L}) \) iff for each node \( u \in \mathcal{U} \) it holds \( \deg_{\mathcal{C}}(u) \geq k \) and the set \( \mathcal{U} \) is the maximal such set. The subset of links \( \mathcal{L}\mathcal{U} \) consists of links from \( \mathcal{L} \) that have both end-nodes in \( \mathcal{U} \).
Table 2: Outdegree distribution in $\text{WA(SN5)}$

| outdeg | frequency | outdeg | frequency | paper                          |
|--------|-----------|--------|-----------|-------------------------------|
| 1      | 2637      | 12     | 8         |                               |
| 2      | 2143      | 13     | 4         |                               |
| 3      | 1333      | 14     | 3         |                               |
| 4      | 713       | 15     | 2         |                               |
| 5      | 396       | 21     | 1         | Pierce et al. (2007)          |
| 6      | 206       | 22     | 1         | Allen et al. (1998)           |
| 7      | 114       | 23     | 1         | Kelly et al. (1997)           |
| 8      | 65        | 26     | 1         | Semple et al. (1993)          |
| 9      | 43        | 41     | 1         | Magliano et al. (2006)        |
| 10     | 24        | 42     | 1         | Doll et al. (1992)            |
| 11     | 10        | 48     | 1         | Snijders et al. (2007)        |

In a collaboration network a $k$-core is the largest subnetwork with the property that each its author wrote a joint work with at least $k$ other authors from the core.

In Figure 2 the cores of orders 20-47 are presented. From this figure we can see a serious drawback of directly applying cores for analysis of collaboration networks. A work with $k$ authors contributes a complete subgraph on $k$ vertices to a collaboration network. For the bibliographies with works with large number of authors the cores procedure identifies as the highest level cores the complete subgraphs corresponding to these works, and not the groups of really the most collaborative authors, as one would expect.

For the SN5 bibliography the components of the cores of orders 20-47 are induced by the papers Snijders et al. (2007); Doll et al. (1992); Magliano et al. (2006); Semple et al. (1993); Kelly et al. (1997); Allen et al. (1998); Pierce et al. (2007) that correspond to the works with the largest number of authors (21-48), see Table 2 and Appendix A. In the picture only the names of authors that are the end-nodes of links with weight larger than 1 are displayed.

An approach to deal with this problem would be to remove all links with weight 1 (or up to some other small threshold) and apply cores on the so reduced network.

A better solution is to identify the works with (too) many authors – very high outdegree in the network $\text{WA}$ – and, for this analysis, remove them from the network $\text{WA}$. We can review the removed works separately.

Yet another approach is to apply on the collaboration network $\text{Co}$ the $p_S$-cores (Batagelj and Zaveršnik, 2011) – a generalization of the ordinary
cores in which the degree $\text{deg}_C(u)$ is replaced by the sum of weights of links from $u$ to other nodes in $U$

$$p_S(u, U) = \sum_{v \in U} w(u, v)$$

A subset $U \subseteq V$ of nodes determines a $p_S$-core at level $t$ $C = (U, L|U)$ in the network $N = (V, L)$ iff for each node $u \in U$ it holds $p_S(u, U) \geq t$ and the set $U$ is the maximal such set.

In Figure 3 the $p_S$-core at level 20 is presented. Each author belonging to it has at least 20 collaborations with other authors inside the core.

Again in the network $SN5$ the cliques corresponding to papers with the largest number of authors appear in the $p_S$-core. Besides them we get also some strongly collaborating groups such as: \{ S. Borgatti, M. Everett \}, \{ H. Bernard, P. Killworth, C. McCarty, E. Johnsen, G. Shelley \}, \{ R. Rotenberg, S. Muth, J. Potterat, D. Woodhouse \}, \{ L. Magliano, M. Maj, C. Malangon, A. Fiorillo \}, and others.
Figure 3: $p_{S}$-core at level 20 of $\textbf{Co}(\text{SN5})$
To neutralize the overrating of the contribution of works with many authors we can try with alternative definitions of collaboration networks using the normalized co-authorship network. The structure (graph) of the collaboration network remains the same, but the weights change.

3.3. The second collaboration network

is defined as

\[ C_n = AW \ast N \]

The value of the weight \( cn_{ij} \)

\[ cn_{ij} = \sum_{p \in W} w_{api} n_{pj} = \sum_{p \in N(i) \cap N(j)} n_{pj} \]

is equal to the contribution of author \( j \) to works, that he/she wrote together with the author \( i \).

In general the entries of matrix \( C_n \) need not to be symmetric (\( cn_{ij} = cn_{ji} \)). In the case when \( n_{pi} = \frac{w_{api}}{\text{outdeg}_{WA}(p)} \) they are

\[ cn_{ij} = \sum_{p \in N(i) \cap N(j)} n_{pj} = \sum_{p \in N(i) \cap N(j)} \frac{1}{\text{outdeg}_{WA}(p)} = \sum_{p \in N(j) \cap N(i)} n_{pi} = cn_{ji} \]

The total contribution of terms \( w_{api} n_{pj} \) for a work \( p \) from the definition of \( cn_{ij} \)

\[ \sum_{j \in A} \sum_{i \in A} w_{api} n_{pj} = \sum_{i \in N(p)} \sum_{j \in A} n_{pj} = \sum_{i \in N(p)} 1 = \text{outdeg}_{WA}(p) \]

is equal to the number of authors of the work \( p \).

Similarly, for an author \( i \) the total contribution of entries \( cn_{ij} \)

\[ \sum_{j \in A} cn_{ij} = \sum_{j \in A} \sum_{p \in W} w_{api} n_{pj} = \sum_{p \in W} w_{api} \sum_{j \in A} n_{pj} = \sum_{p \in W} w_{api} = \text{indeg}_{WA}(i) \]

is equal to the number of works that the author \( i \) co-authored; and the (diagonal) entry

\[ cn_{ii} = \sum_{p \in N(i)} n_{pi} \]
Table 3: List of the "best" authors in SN5

| i | author       | cn_{11} | total | $K_i$ |
|---|--------------|---------|-------|-------|
| 1 | Burt, R      | 43.83   | 53    | 0.173 |
| 2 | Newman, M    | 36.77   | 60    | 0.387 |
| 3 | Doreian, P   | 34.44   | 47    | 0.267 |
| 4 | Bonacich, P  | 30.17   | 41    | 0.264 |
| 5 | Marsden, P   | 29.42   | 37    | 0.205 |
| 6 | Wellman, B   | 26.87   | 41    | 0.345 |
| 7 | Leydesdorf, L| 24.37   | 35    | 0.304 |
| 8 | White, H     | 23.50   | 33    | 0.288 |
| 9 | Friedkin, N  | 20.00   | 23    | 0.130 |
|10 | Borgatti, S  | 19.20   | 41    | 0.532 |
|11 | Everett, M   | 16.92   | 31    | 0.454 |
|12 | Litwin, H    | 16.00   | 21    | 0.238 |
|13 | Freeman, L   | 15.53   | 20    | 0.223 |
|14 | Barabasi, A  | 14.99   | 35    | 0.572 |
|15 | Snijders, T  | 14.99   | 30    | 0.500 |
|16 | Valente, T   | 14.80   | 34    | 0.565 |
|17 | Breiger, R   | 14.44   | 20    | 0.278 |
|18 | Skvoretz, J  | 14.43   | 27    | 0.466 |
|19 | Krackhardt, D| 13.65   | 25    | 0.454 |
|20 | Carley, K    | 12.93   | 28    | 0.538 |
|21 | Pattison, P  | 12.10   | 27    | 0.552 |
|22 | Wasserman, S | 11.72   | 26    | 0.549 |
|23 | Berkman, L   | 11.21   | 30    | 0.626 |
|24 | Moody, J     | 10.83   | 15    | 0.278 |
|25 | Scott, J     | 10.47   | 15    | 0.302 |
|26 | Latkin, C    | 10.14   | 37    | 0.726 |
|27 | Morris, M    | 9.98    | 20    | 0.501 |
|28 | Rothenberg, R| 9.82    | 28    | 0.649 |
|29 | Kadushin, C  | 9.75    | 11    | 0.114 |
|30 | Faust, K     | 9.72    | 18    | 0.460 |
|31 | Batagelj, V  | 9.69    | 20    | 0.516 |
|32 | Mizruchi, M  | 9.67    | 15    | 0.356 |
|33 | [Anon]       | 9.00    | 9     | 0.000 |
|34 | Johnson, J   | 8.89    | 21    | 0.577 |
|35 | Fararo, T    | 8.83    | 16    | 0.448 |
|36 | Lazega, E    | 8.50    | 12    | 0.292 |
|37 | Knopke, D    | 8.33    | 11    | 0.242 |
|38 | Ferligoj, A  | 8.19    | 19    | 0.569 |
|39 | Brewer, D    | 8.03    | 11    | 0.270 |
|40 | Klovdahl, A  | 7.96    | 17    | 0.532 |
|41 | Hammer, M    | 7.92    | 10    | 0.208 |
|42 | White, D     | 7.83    | 15    | 0.478 |
|43 | Holme, P     | 7.42    | 14    | 0.470 |
|44 | Boyd, J      | 7.37    | 13    | 0.433 |
|45 | Kilduff, M   | 7.25    | 16    | 0.547 |
|46 | Small, H     | 7.00    | 7     | 0.000 |
|47 | Iacobucci, D | 7.00    | 12    | 0.417 |
|48 | Pappi, F     | 6.83    | 10    | 0.317 |
|49 | Chen, C      | 6.78    | 12    | 0.435 |
|50 | Seidman, S   | 6.75    | 9     | 0.250 |
is equal to the total contribution of author $i$ to his/her works.

Therefore we can base on the entries of matrix $C_n$ the *self-sufficiency* index

$$S_i = \frac{cn_{ii}}{\text{indeg}_{WA}(i)}$$

as the proportion of author’s contribution to his/her works and the total number of works he/she co-authored.

The *collaborativeness* index is complementary to it

$$K_i = 1 - S_i$$

All $cn_{ij}$ values

$$\sum_{i \in A} \sum_{j \in A} cn_{ij} = \sum_{i \in A} \text{indeg}_{WA}(i) = m_{WA}$$

sum up to the number of all links in the network $WA$.

In Table 3 the 50 authors with the largest self-contribution $cn_{ii}$ to the topic of 'social network analysis' are presented together with the total number of works on the topic that an author co-authored, and his/her collaborativeness index.

### 3.4. The third collaboration network

is defined as

$$C_t = N^T \ast N$$

The weight $ct_{ij}$ is equal to the total contribution of collaboration of authors $i$ and $j$ to works.

The total contribution of a complete subgraph corresponding to the authors of a work is 1:

$$\sum_{i \in A} \sum_{j \in A} n_{ip}^T n_{pj} = \sum_{i \in A} n_{pi} \sum_{j \in A} n_{pj} = \sum_{i \in A} n_{pi} \cdot 1 = 1$$

The weights $ct_{ij}$ are symmetric

$$ct_{ij} = \sum_{p \in W} n_{ip}^T n_{pj} = \sum_{p \in W} n_{jp}^T n_{pi} = ct_{ji}$$

and the sum

$$\sum_{j \in A} ct_{ij} = \sum_{j \in A} \sum_{p \in W} n_{pi} n_{pj} = \sum_{p \in W} n_{pi} \sum_{j \in A} n_{pj} = \sum_{p \in W} n_{pi}$$
is equal to the total contribution of author $i$ to works from $W$.

The sum of all weights $c_{ij}$

$$\sum_{i \in A} \sum_{j \in A} c_{ij} = \sum_{i \in A} \sum_{p \in W} n_{pi} = \sum_{p \in W} \sum_{i \in A} n_{pi} = \sum_{p \in W} 1 = |W|$$

is equal to the number of all works.

We can also introduce the *author's contribution to the field* as

$$ac_i = \frac{|A|}{|W|} \sum_{p \in W} n_{pi}$$

with the property

$$\sum_{i \in A} ac_i = |A|$$

Therefore the average $ac$ is 1.

Note also that

$$b(\text{Co}) = b(\text{Cn}) = b(\text{Ct})$$

In Figure 4 the $p_s$-core of order 0.75 in the third collaboration network $\text{Ct(SN5)}$ is presented. In this core the large cliques disappear. The largest core’s component consists of the main stream social networks researchers with the most intensive collaboration pairs: Borgatti and Everett, Killworth and Bernard, Bonacich and Bienstock, Ferligoj and Batagelj, Pattison and Robins, etc. The second largest component consists of physicists with more intensive collaboration pairs: Newman and Park, Barabasi and Albert, and Masuda and Konno. In the smaller components we find additional three pairs: Lienhard and Holland (with Lienhard represented with two nodes), Metzke and Steinhaus, and Chou and Chi.

4. Derived networks

4.1. Bibliographic Coupling and Co-Citation

In WoS2Pajek the citation relation means $p \text{Ci} q \equiv p$ cites $q$. Therefore the *bibliographic coupling* network $\text{biCo}$ can be determined as (Kessler, 1963)

$$\text{biCo} = \text{Ci} \ast \text{Ci}^T$$
Figure 4: $p_S$-core of order 0.75 in the third collaboration network on $C_t(SN5)$
The corresponding weight

\[ bico_{pq} = \sum_{s \in W} c_{ips} c_{iqs} = \sum_{s \in N(p) \cap N(q)} 1 \]

is equal to the number of works cited by both works \( p \) and \( q \). It is symmetric \( bico_{pq} = bico_{qp} \).

Again we have problems with works with many citations, especially with review papers. To neutralize their impact we can introduce a normalized measure such as

\[ biCon = \frac{1}{2} (n(Ci) \ast Ci^T + Ci \ast n(Ci)^T) \]

It is easy to verify that \( bicon_{pq} \in [0,1] \) and \( bicon_{pq} = bicon_{qp} \) (symmetry). It also holds: \( bicon_{pq} = 1 \) iff the works \( p \) and \( q \) are referencing the same works.

Note that

\[ b(n(Ci) \ast Ci^T) = b(Ci \ast n(Ci)^T). \]

The \( cC_{pq} \) element of the first term represents the ‘importance’ of common \((p,q)\)-citations for the work \( p \); and the \( CC_{pq} \) element of the second term represents the ‘importance’ of common \((p,q)\)-citations for the work \( q \).

\[ bicon_{pq} = \frac{1}{2} (cC_{pq} + CC_{pq}) \]

Note that the first term in the definition of \( biCon \) is equal to the transpose of the second term

\[ (Ci \ast n(Ci)^T)^T = n(Ci) \ast Ci^T \]

and therefore \( CC_{pq} = cC_{qp} \). This can be used for more efficient computation of \( biCon \). We only need to compute the first term \( cC \). Then

\[ bicon_{pq} = \frac{1}{2} (cC_{pq} + cC_{qp}) \]

In the network \( biCon(SN5) \) the larger components with edges with \( bicon = 1 \) correspond to papers with a single reference to a book (Wasserman, S., Faust, K.: Social network analysis. Cambridge UP, 1994; Taylor, Howard F.: Balance in small groups. Van Nostrand Reinhold, 1970; Belle, D.: Childrens social networks and social supports. Wiley, 1989; Gottlieb, B. H.: Social
networks and social support. Sage, 1981; Yan, Yunxiang: The flow of gifts. Stanford UP, 1996; Zhang, L.: Strangers in the City. Stanford UP, 2001). There are also several pairs of papers with $bicon = 1$, mostly written by the same author. More interesting groups we can obtain as larger islands with values below 1.

Similarly the document co-citation network $\text{coCi}$ can be determined as

$\text{coCi} = \text{Ci}^T \ast \text{Ci}$

The corresponding weight

$$coci_{pq} = \sum_{s \in W} ci_{sp} ci_{sq} = \sum_{s \in N^{-}(p) \cap N^{-}(q)} 1$$

is equal to the number of works citing both works $p$ and $q$. $N^{-}(p)$ denotes the set of neighbors from which the node $p$ can be entered.

It holds $\text{coCi}(N) = \text{biCo}(N^T)$ and also for corresponding normalized networks $\text{coCin}(N) = \text{biCon}(N^T)$.

4.2. Other derived networks

The weight $aci_{ip}$ in the author citation network

$\text{ACi} = \text{AW} \ast \text{Ci}$

counts the number of times author $i$ cited work $p$.

The author co-citation network can be obtained as

$\text{ACo} = b(\text{ACi}) \ast b(\text{ACi})^T$

The weight $aco_{ij}$ counts the number of works cited by both authors $i$ and $j$.

The weight $ak_{ik}$ in the authors using keywords network

$$\text{AK} = \text{AW} \ast \text{WK}$$

counts the number of works in which the author $i$ used a keyword $k$. 
4.3. The cited co-authorship network

Quattrociocchi et al. (2011) proposed the cited co-authorship network:

$$AW * \text{diag}(\text{indeg}_{Ci}(p)) * WA$$

the weight of two collaborating authors is equal to the sum of numbers of citations to co-authored works where \(\text{indeg}_{Ci}(p)\) is number of citations to work \(p\).

Its normalized version would be:

$$Cc = AW * \text{diag}\left(\frac{\text{indeg}_{Ci}(p)}{\text{outdeg}_{Ci}(p)^2}\right) * WA$$

with the properties

$$\sum_{i \in A} \sum_{j \in A} w_{ai} in\text{deg}_{Ci}(p) \frac{\text{outdeg}_{Ci}(p)^2}{aw_{pj}} = \text{indeg}_{Ci}(p)$$

and

$$\sum_{i \in A} \sum_{j \in A} cc_{ij} = \sum_{p \in W} \text{indeg}_{Ci}(p) = |A_{Ci}|$$

where \(|A_{Ci}|\) is the number of arcs in the citation network \(Ci\).

4.4. Authors’ citations network

The network of citations between authors can be obtained as

$$Ca = AW * Ci * WA$$

The weight \(ca_{ij}\) counts the number of times a work co-authored by \(i\) is citing a work co-authored by \(j\), see Figure 5.

In Figure 6 some link islands from \(Ca(SN5)\) are presented. The largest island consists of the main stream social networks researchers with some subgroups: the star around R. Burt in the top left part; the S. Borgatti and M. Everett tandem in the bottom left part; the probabilistic group in the top right part with G. Robins, P. Pattison, T. Snijders, S. Wasserman, and P. Holland as the most prominent; and others: J. Skvoretz, D. Krackhardt, P. Doreian, R. Breiger, H. White, L. Freeman, and P. Marsden.

The “scale-free” island consists mainly of physicists M. Newman, A. Barabasi, D. Watts, R. Albert, P. Holme and others. In the “medical” island
the central authors are J. Potterat, R. Rothenberg, D. Woodhouse, S. Muth, A. Klovdahl, and S. Friedman. There is also an island on "education and psychology" with T. Farmer, R. Cairns, B. Cairns, H. Xie, and P. Rodkin.

Most of the other islands are star-like, usually a professor with his Phd students.

5. Conclusions

In the paper we showed that the bibliographic data can be transformed into a collection of compatible networks. Using network multiplication different interesting derived networks can be obtained. In defining them an appropriate normalization should be considered. The proposed approach can be applied also to other collections of compatible networks (see Batagelj, 2009, pg. 8260–8262).

Note that most of the obtained derived networks are one-mode networks. To analyse them standard SNA methods can be used. For analysis of two-mode networks we can use direct methods such as (generalized) two-mode cores, two-mode hubs and authorities and 4-rings islands (Ahmed et al., 2007).

We can also transform the citation network (and other WoS networks) into temporal network using the partition of works by publication year. Using the time slices also the temporal sequences of corresponding derived networks
Figure 6: Some link islands in the network $\text{Ca}(\text{SN5})$
can be obtained.

**Pajek** allows analyses on different levels specified by a partition of the corresponding set of units and obtained using the *shrinking* of classes. For example: partition of authors by institutions, or partition of institutions by countries, partitions of authors by discipline/field/subfield, etc. Using the *extraction* of selected classes we can reduce the network to the area of our interest.

The HOW TO in **Pajek** for the described approach is available at

http://pajek.imfm.si/doku.php?id=how_to:biblio

6. Acknowledgments

The work was supported in part by the ARRS, Slovenia, grant P1-0294, as well as by grant within the EUROCORES Programme EUROGIGA (project GReGAS) of the European Science Foundation. The second author was financed in part by the European Union, European Social Fund.

Appendix A. References for analyses

Allen, J., Anton, R. F., Babor, T. F., Carbonari, J., Carroll, K. M., Connors, G. J., Cooney, N. L., Del Boca, F. K., Di-Clemente, C. C., Donovan, D., Kadden, R. M., Litt, M., Longabaugh, R., Mattson, M., Miller, W. R., Randall, C. L., Rounsaville, B. J., Rychtarik, R. G., Stout, R. L., Tonigan, J. S., Wirtz, P. W., & Zweben, A. (1998). Matching Alcoholism Treatments to Client Heterogeneity: Project Match Three-Year Drinking Outcomes. *Alcoholism-Clinical and Experimental Research, 22*(6), 1300–1311.

Doll, L.S., Petersen, L.R., White, C.R., Johnson, E.S., Ward, J.W., Williams, A., Altman, R., Becker, G., Bernarducci, J., Busch, M., Clary, N., Davis, J., Darr, F., Grindon, A., Kleinman, S., Lamberson, H., Lenes, B., Menitove, J., Molinaris, J., Ness, P., Raevsky, C., Holland, P., Shafer, A.W., Sherwood, W., Stevens, C., Vaughan, H., Arnold, E., Donovan, D.E., Kessler, D., Harper, M., Hernandez, J., Ksell, T., Melfresh, S., Myers, M.K., Montgomery, A., Nason, M., Sanchez, J., Schulze, G., Shahan, M., Stephenson, S., Theobald, J., & Wilke, D. (1992). Homosexually and non-homosexually identified men who have sex with men – a behavioral comparison. *Journal of Sex Research, 29*(1), 1–14.
Kelly, J. A., Murphy, D. A., Sikkema, K. J., McAuliffe, T. L., Roffman, R. A., Solomon, L. J., Winett, R. A., Kalichman, S. C., Heckman, T. G., Perry, M. J., Stevenson, L. Y., Hauth, A. C., Koob, J. J., Morgan, M. G., Norman, A., Lemke, A., Steiner, S., Trenary, B., Flynn, B., Ayotte, D. R., Desiderato, L. L., Lombard, D. N., & Yaffe, D. M. (1997). Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. *Lancet, 350*(9090), 1500–1505. doi: 10.1016/S0140-6736(97)07439-4

Magliano, L., Fiorillo, A., Malangone, C., De Rosa, C., Favata, G., Sasso, A., Prezioso, M., Pezzenati, L., Gentile, F., Casale, L., Bondi, E., Test, G. R., Di Lella, M., Biscusi, E., Degl’Innocent, F., Bellini, R., Di Nunzio, R., Matrella, L., Salmeri, R., Cantone, R., Gargiulo, L., Esposito, A., Delcurratolo, V., Giannini, M., Maresca, L., Cavaliere, G., Scandone, B., Folla, M., Raffaeli, M., Innocente, P., Dagianti, F., Lucania, S., Scorsino, A., Bardichia, F., Cerullo, G., Curreli, R., Miscali, S., Scordato, M., Campo, G., Maj, M.; & Gruppo Di Lavoro (2006). Family psychoeducational interventions for schizophrenia in routine settings: impact on patients’ clinical status and social functioning and on relatives’ burden and resources. *Epidemiologia e Psichiatria Sociale, 15*(3), 219–27.

Pierce, J. P., Natarajan, L., Caan, B. J., Parker, B. A., Greenberg, E. R., Flatt, S. W., Rock, C. L., Kealey, S., Al-Delaimy, W. K., Bardwell, W. A., Carlson, R. W., Emond, J. A., Faerber, S., Gold, E. B., Hajek, R. A., Hollenbach, K., Jones, L. A., Karanja, N., Madlensky, L., Marshall, J., Newman, V. A., Ritenbaugh, C., Thomson, C. A., Wasserman, L., & Stefanick, M. L. (2007). Influence of a Diet Very High in Vegetables, Fruit, and Fiber and Low in Fat on Prognosis Following Treatment for Breast Cancer. *JAMA, 298*(3), 289–298. doi: 10.1001/jama.298.3.289

Semple, S. J., Patterson, T. L., Temoshok, L. R., McCutchan, J. A., Straits-Troster, K. A., Chandler, J. L., Grant, I., Atkinson, J. H., Velin, R. A., Oldfield, E. C., Wallace, M. R., Malone, J., Spector, S. A., Thal, L., Heaton, R. K., Hesselink, J., Jernigan, T., Wiley, C. A., Olshen, R., Abramson, I., Butters, N., Dupont, R., Zisook, S., Jeste, D., Sieburg, H., & Weinrich, J. D. (1993). Identification of psychological stressors among HIV-positive women. *Women & Health, 20*(4), 15–36. doi: 10.1300/J013v20n04_02
Snijders, T. A. B., Robinson, T., Atkinson, A. C., Riani, M., Gormley, I. C., Murphy, T. B., Sweeting, T., Leslie, D. S., Longford, N. T., Kent, J. T., Lawrance, T., Airoldi, E. M., Besag, J., Blei, D., Fienberg, S. E., Breiger, R., Butts, C. T., Doreian, P., Batagelj, V., Ferligoj, A., Draper, D., van Duijn, M. A. J., Faust, K., Petrescu-Prahova, M., Forster, J. J., Gelman, A., Goodreau, S. M., Greenwood, P. E., Gruenberg, K., Francis, B., Hennig, C., Hoff, P. D., Hunter, D. R., Husmeier, D., Glasbey, C., Krackhardt, D., Kuha, J., Skrondal, A., Lawson, A., Liao, T. F., Mendes, B., Reinert, G., Richardson, S., Lewin, A., Titterington, D. M., Wasserman, S., Werhli, A. V., & Ghazal, P. (2007). Discussion on the paper by Handcock, Raftery and Tantrum. *Journal of the Royal Statistical Society: Series A – Statistics in Society, 170*(2), 322–354. doi: 10.1111/j.1467-985X.2007.00471.x

**References**

Ahmed, A., Batagelj, V., Fu, X., Hong, S. H., Merrick, D., & Mrvar, A. (2007). Visualisation and Analysis of the Internet Movie Database. *Proceedings of the Asia-Pacific Symposium on Visualisation (APVIS2007) 17–24*. doi: 10.1109/APVIS.2007.329304

Batagelj, V. (1994). Semirings for social networks analysis. *The journal of Mathematical Sociology, 19*(1), 53–68. doi: 10.1080/0022250X.1994.9990135

Batagelj, V. (2007). Wos2Pajek – networks from web of science [Data converter]. Retrieved from: http://pajek.imfm.si/doku.php?id=wos2pajek
Last accessed: 5. 11. 2012.

Batagelj, V. (2009). Social Network Analysis, Large-Scale. In R.A. Meyers, (Ed.), *Encyclopedia of Complexity and Systems Science* (pp. 8245–8265) New York, NY: Springer.

Batagelj, V. (2012). Pajek’s wiki. Retrieved from: http://pajek.imfm.si
Last accessed: 5. 11. 2012.

Batagelj, V., & Zaveršnik, M. (2011). Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification, 5*(2), 129–145. doi: 10.1007/s11634-010-0079-y
De Nooy, W., Mrvar, A., & Batagelj, V. (2012). *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences); Revised and Expanded Second Edition*. Cambridge; New York: Cambridge University Press.

Kejžar, N., Korenjak Černe, S., & Batagelj, V. (2010). Network Analysis of Works on Clustering and Classification from Web of Science. Classification as a Tool for Research. In H. Locarek-Junge, & C. Weihs (Eds.), *Proceedings of IFCS 2009. Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 525–536). Berlin, Germany: Springer. doi: 10.1007/978-3-642-10745-0_57

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25. doi: 10.1002/asi.5090140103

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*(1), 016132-1–016132-7. doi: 10.1103/PhysRevE.64.016132

Quattrociocchi, W., & Amblard, W. (2011). Selection in scientific networks. *Social Network Analysis and Mining, 2*(3), 229–237. doi: 10.1007/s13278-011-0043-7

Rosengren, K. E. (1968). *Sociological Aspects of the Literary System*. Stockholm, Sweden: Natur och Kultur.

Small, H. G. (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. doi: 10.1002/asi.4630240406

Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge, UK: Cambridge University Press.

White, H. D., & Griffith, B. C. (1981). Author co-citation: a literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163–171. doi: 10.1002/asi.4630320302

Zaveršnik, M., & Batagelj, V. (2004). *Islands*. In XXIV. International Sunbelt Social Network Conference, Portorož, Slovenia.