Some Analyses of Erdős Collaboration Graph

Vladimir Batagelj* and Andrej Mrvar
University of Ljubljana

Abstract

Patrick Ion (Mathematical Reviews) and Jerry Grossman (Oakland University) maintain a collection of data on Paul Erdős, his co-authors and their co-authors. These data can be represented by a graph, also called the Erdős collaboration graph.

In the paper some techniques for analysis of large networks (different approaches to identify 'interesting' individuals and groups, analysis of internal structure of the main core using pre-specified blockmodeling and hierarchical clustering) and visualizations of their parts, are presented on the case of Erdős collaboration graph, using program **`Pajek`**.

Keywords: large networks, cores, visualization, clustering, blockmodeling

Math. Subj. Class. (1991): 92H30, 93A15, 68R10, 05C90, 92G30, 62H

*The corresponding author:** Vladimir Batagelj, University of Ljubljana, FMF, Department of Mathematics, Jadranska 19, 1000 Ljubljana, Slovenia. e-mail: `vladimir.batagelj@uni-lj.si`

1 Introduction

The current level of development of computer technology allows us to deal with large (having thousands to several hundreds of thousands of lines – arcs and/or edges) networks already on PCs. The basic problem is that such networks can't be grasped in a single view – we have to either produce a global view/characteristics omitting the details, or make a detailed inspection of some selected part of the network of moderate size (some tens of vertices), or something in between. The Erdős collaboration graph is an example of a large network on which we can present some techniques that can be used for analysis of large networks. The obtained results are of their own interest for graph theory community.

Paul Erdős was one of the most prolific mathematicians in the history, with more than 1500 papers to his name. He was born March 26, 1913 in Budapest, Hungary and died September 20, 1996 in Warsaw, Poland. Paul Erdős won many prizes including Cole Prize of the American Mathematical Society in 1951 and the Wolf Prize in 1983. He is also known as a promoter of collaboration and as a mathematician with the largest number of different co-authors. This was a motivation for the introduction of the Erdős number.

2 Erdős collaboration graph

The *Erdős number* $\mathbf{n}_E$ of an author is defined as follows: Paul Erdős himself has $\mathbf{n}_E = 0$; people who have written a joint paper with Paul Erdős have $\mathbf{n}_E = 1$; and their co-authors, with Erdős number not yet defined, have $\mathbf{n}_E = 2$; etc.

Often on the home pages of people interested in or related to combinatorics we find the statement:

My Erdős number is ...

Patrick Ion (Mathematical Reviews) and Jerry Grossman (Oakland University) collected the related data (Grossman and Ion, 1995, Grossman, 1996) and made them available at the URL:

`http://www.oakland.edu/~grossman/erdoshp.html`

These data can be represented as a graph called the *Erdős collaboration graph* – $\mathcal{E} = (V, E)$. The set of its vertices V consists of known authors with $\mathbf{n}_E \leq 2$, and its edges connect two authors, if they wrote a joint paper, and at least one of them has $\mathbf{n}_E \in \{0, 1\}$ — the data about collaboration among authors with $\mathbf{n}_E = 2$ are not (yet?) available.

The data are updated annually. Table 1 shows the 'growth' of the Erdős collaboration graph.

By removing Paul Erdős himself and connections to him from the graph $\mathcal{E}$ we get the *truncated Erdős collaboration graph* $\mathcal{E}'$. The last, 1999 edition of this graph contains 6100 vertices and 9939 edges.

Table 1: The growth of Erdős collaboration graph

	$\mathbf{n}_E = 0, 1$		$\mathbf{n}_E = 2$		Total	
Year	$\lvert V_{01}\rvert$	$\lvert E_{01}\rvert$	$\lvert V_2\rvert$	$\lvert E_2\rvert$	$\lvert V\rvert$	$\lvert E\rvert$
1997	473	1786	5016	7658	5489	9444
1998	486	1866	5337	8124	5823	9990
1999	493	1909	5608	8522	6101	10431

The names of authors with $\mathbf{n}_E = 1$ are written in capitals and the names of authors with $\mathbf{n}_E = 2$ are only capitalized.

We used program **`Pajek`** to make some analyses and get layouts of selected parts of Erdős collaboration graph. **`Pajek`** is a program, for Windows, for analysis and visualization of large networks (Batagelj and Mrvar, 1998). It is freely available for noncommercial use at:

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

3 Some basic analyses

In Table 2 some basic statistics about the number of co-authors – vertex degrees in Erdős collaboration graph $\mathcal{E}$ are presented.

Table 2: Basic statistics on degrees in Erdős collaboration graph

	$\mathbf{n}_E = 1$	$\mathbf{n}_E \in \{0, 1\}$	$\mathbf{n}_E = 2$
minimum	1	1	1
median	18	18	1
average degree	24.08	25.03	1.52
maximum	277	492	18
maximizer	Harary	Erdős	Lesniak

There exist 17 connected components in $\mathcal{E}'$. One of them is very large (it contains 6045 authors), others are smaller (they contain 12 authors at most, see Figure 1).

The diameter of the large component in graph $\mathcal{E}'$ is 12 with 3 diametric pairs of vertices

(Jakob, Matthias : An, Chung Ming)
(Jakob, Matthias : Corsten, L. C. A.)
(Jakob, Matthias : Stux, Ivan E.)

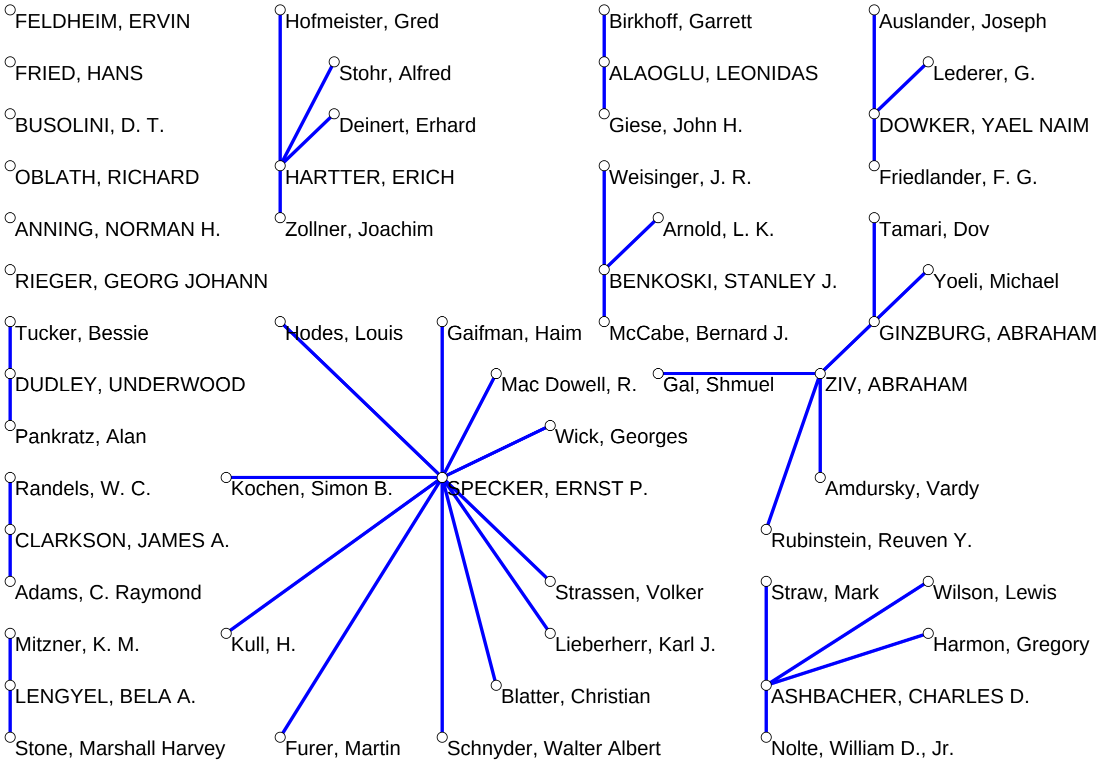

Figure 1: Small components.

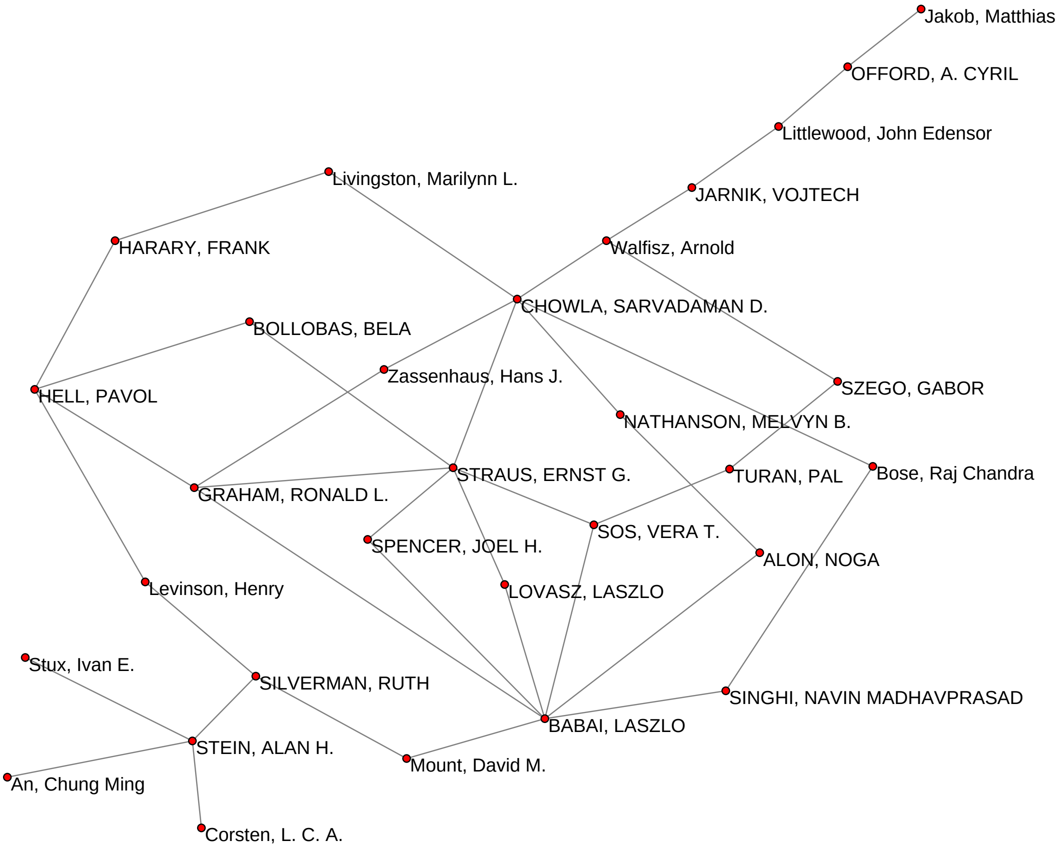

Figure 2: Diametric Geodesics.

Table 3: Top ten authors according to number of co-authors

Author	No. of co-authors	core
ERDOS, PAUL	492	10
HARARY, FRANK	277	10
ALON, NOGA	168	10
SHELAH, SAHARON	146	7
GRAHAM, RONALD L.	126	10
COLBOURN, CHARLES J.	121	7
KLEITMAN, DANIEL J.	117	10
ODLYZKO, ANDREW M.	113	8
TUZA, ZSOLT	107	10
HOFFMAN, ALAN J.	93	7

The corresponding geodesic subgraph is presented in Figure 2.

The top ten authors according to the number of co-authors are presented in Table 3. Frank Harary and Noga Alon, the two authors with the highest degree in $\mathcal{E}'$, did not (?) write an article together. But there exist 15 authors with whom both of them are co-authors. The common co-authors are:

ERDOS, PAUL	TROTTER, WILLIAM T., JR.
BOLLOBAS, BELA	TUZA, ZSOLT
DUKE, RICHARD A.	Akiyama, Jin
FAUDREE, RALPH J.	Brualdi, Richard A.
GRAHAM, RONALD L.	Dewdney, Alexander Keewatin
NESETRIL, JAROSLAV	Fellows, Michael R.
RODL, VOJTECH	Karp, Richard M.
THOMASSEN, CARSTEN	Welsh, Dominic J. A.

The distributions of distances of other authors from Harary and Alon are given in Table 4. We see that the Alon's co-authors are more collaborative.

4 Cores

Let $G = (V, E)$ be a graph. The notion of core was introduced by Seidman (1983). A maximal subgraph $H_k = (W, E|W)$ induced by the set $W \subseteq V$ is a k-*core*, or *core of order* k, iff $\forall v \in W : \deg_H(v) \geq k$, see Figure 3. The core of maximum order is also called the *main* core. The cores have two important properties:

- The cores are nested: $i < j \implies H_j \subseteq H_i$

Table 4: Distributions of distances from Harary and Alon

Distance	from Harary	from Alon
0	1	1
1	276	167
2	938	1124
3	2757	2764
4	1514	1416
5	538	473
6	12	99
7	9	1
Sum	6045	6045
∞	55	55
Average	3.193	3.199

- There exists an efficient algorithm of order $O(|E|)$ for determining the cores (Batagelj, Mrvar and Zaveršnik, 1999).

We denote the *neighborhood* of vertex $v \in V$ by $N(v)$:

$$N(v) = \{u \in V : (v : u) \in E\} \tag{1}$$

and the *rooted neighborhood* of vertex $v \in V$ by $N^+(v)$:

$$N^+(v) = N(v) \cup \{v\} \tag{2}$$

In $\mathcal{E}$ the main core is of order 10, and of order 9 in $\mathcal{E}'$. In Table 5 the distribution in $\mathcal{E}$ of number of authors in k-cores (second column), and the distributions of number of co-authors in cores for selected members of the main core, are given. The authors belonging to the main core and some of their characteristics are presented in Table 6, where $\overline{\text{deg}}$ is the average degree of all co-authors.

$$\overline{\text{deg}}(v) = \begin{cases} 0 & N(v) = \emptyset \\ \dfrac{1}{|N(v)|} \displaystyle\sum_{u \in N(v)} \text{deg}(u) & \text{otherwise} \end{cases} \tag{3}$$

and $\overline{\text{core}}$ is the average core number of all co-authors.

$$\overline{\text{core}}(v) = \begin{cases} 0 & N(v) = \emptyset \\ \dfrac{1}{|N(v)|} \displaystyle\sum_{u \in N(v)} \text{core}(u) & \text{otherwise} \end{cases} \tag{4}$$

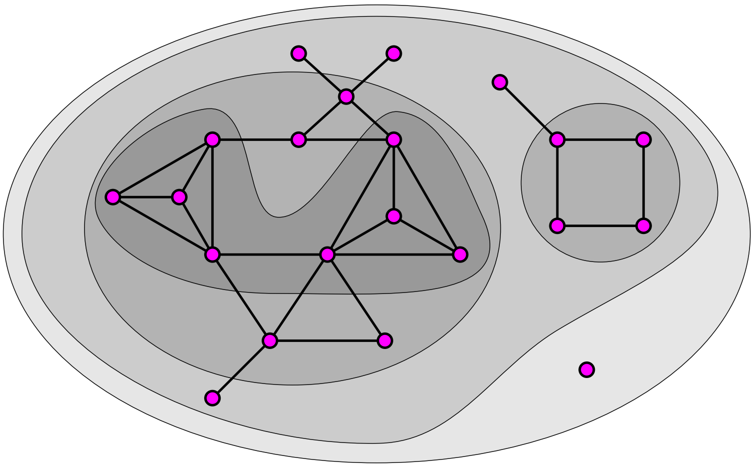

Figure 3: 0, 1, 2 and 3 core.

Table 5: Distribution of number of co-authors in cores

core	No. of authors	ERDŐS	HARARY	ALON	Lesniak	RODL	SIMON.	LEHEL
10	37	35	14	18	10	19	12	16
9	14	12	4	2	3	7	2	1
8	27	19	4	11	1	7	3	0
7	73	45	16	15	2	13	3	1
6	96	50	17	8	1	4	2	2
5	178	94	14	8	0	4	2	5
4	231	89	11	19	1	4	2	1
3	410	71	27	22	0	8	0	3
2	853	62	38	23	0	12	2	2
1	4182	15	132	42	0	10	1	1
Sum	6101	492	277	168	18	88	29	32
Average	1.76	4.88	2.98	4.22	8.83	5.89	7.52	7.22

Table 6: Authors and number of their co-authors in the main core, total number of co-authors, average core and average degree of all their co-authors, and their collaborativeness

Author	Co-authors	All co-authors	$\overline{\text{core}}$	$\overline{\text{deg}}$	coll
ERDOS PAUL	35	* 492	4.88	24.08	2.05
FAUDREE, RALPH J.	19	51	6.33	38.63	1.58
GRAHAM, RONALD L.	19	126	4.94	26.64	2.03
RODL, VOJTECH	19	88	5.89	32.53	1.70
ALON, NOGA	18	* 168	4.22	17.63	* 2.37
GYARFAS, ANDRAS	18	41	6.90	40.85	1.45
JACOBSON, MICHAEL S.	17	55	5.87	30.60	1.70
TUZA, ZSOLT	17	107	4.21	22.73	* 2.38
CHUNG, FAN RONG K.	16	81	5.12	30.79	1.95
GOULD, RONALD J.	16	37	6.73	35.00	1.49
FUREDI, ZOLTAN	16	65	5.95	36.45	1.68
LEHEL, JENO	16	32	* 7.22	42.69	1.39
SCHELP, RICHARD H.	16	42	6.21	33.12	1.61
SPENCER, JOEL H.	16	67	6.10	38.99	1.64
BURR, STEFAN ANDRUS	15	29	7.14	* 58.58	1.40
SZEMEREDI, ENDRE	15	68	5.87	30.13	1.70
HARARY, FRANK	14	* 277	2.98	9.62	* 3.36
WEST, DOUGLAS B.	14	81	4.14	20.10	* 2.42
CHARTRAND, GARY	13	81	5.27	23.33	1.90
LOVASZ, LASZLO	13	91	4.74	23.30	2.11
NESETRIL, JAROSLAV	13	65	5.31	33.12	1.88
PACH, JANOS	13	78	4.88	25.33	2.05
BABAI, LASZLO	12	75	4.99	27.32	2.01
FRANKL, PETER	12	64	5.41	33.67	1.85
SIMONOVITS, MIKLOS	12	29	* 7.52	* 50.76	1.33
TROTTER, WILLIAM T., JR.	12	54	5.48	39.15	1.82
OELLERMANN, ORTRUD R.	11	40	6.63	38.45	1.51
SOS, VERA T.	11	37	6.84	47.00	1.46
BOLLOBAS, BELA	10	78	4.92	30.10	2.03
CHEN, GUANTAO	10	29	5.76	33.83	1.74
GODDARD, WAYNE D.	10	48	6.10	29.56	1.64
HAJNAL, ANDRAS	10	50	5.74	36.66	1.74
KLEITMAN, DANIEL J.	10	117	4.03	18.11	* 2.48
KUBICKA, EWA MARIE	10	19	7.11	* 64.95	1.41
KUBICKI, GRZEGORZ	10	25	7.08	* 50.68	1.41
Lesniak, Linda M.	10	18	* 8.83	34.89	1.13
ROUSSEAU, CECIL CLYDE	10	31	5.10	34.68	1.96

Table 7: Top ten most collaborative authors

	author	coll
1	HARARY, FRANK	3.358
2	SHELAH, SAHARON	3.106
3	HOFFMAN, ALAN J.	3.056
4	SALAT, TIBOR	2.674
5	CHUI, CHARLES KAM-TAI	2.538
6	JANSON, SVANTE	2.537
7	FISHBURN, PETER C.	2.521
8	FRAENKEL, AVIEZRI S.	2.514
9	VAN LINT, JACOBUS HENDRICUS	2.503
10	KRANTZ, STEVEN GEORGE	2.493

The authors with the highest values of $\overline{\text{deg}}$ and $\overline{\text{core}}$ are indicated by a star.

We have to be very careful in interpretation of $\overline{\text{deg}}$ and $\overline{\text{core}}$. Their high values imply that a 'central' author is mainly collaborating with other 'central' authors. Therefore we propose as a measure of *collaborativeness* the quantity

$$\text{coll}(v) = \frac{\text{core}(v)}{\overline{\text{core}}(v)} \tag{5}$$

that measures the openness of author v towards 'peripheral' authors. If $\text{core}(v) = 0$, also $\text{coll}(v) = 0$.

The most collaborative authors in the main core are Frank Harary, Daniel Kleitman, Douglas West, Zsolt Tuza and Noga Alon. But, it turns out that among the top ten most collaborative authors Frank Harary is the only one from the main core (see Table 7).

Note that this is valid only relatively to the graph $\mathcal{E}$ since for authors with $\mathbf{n}_E = 2$ their core numbers are underestimated, because of incomplete data about their collaboration.

5 Lords

We call 'lords' vertices that have 'strong influence' to their neighborhoods. At the beginning we assign to each vertex its degree as its initial power. The final distribution of power is the result of 'transferring' the power from weaker to stronger vertices.

To determine this distribution we order vertices in the increasing order according to their degrees and in this order we deal the power of the current vertex to its stronger neighbors proportionally. The result of applying this procedure on $\mathcal{E}'$ is given in Table 8.

In an alternative version of this procedure each vertex obtains equal initial power 1, and the vertex is considered stronger if it has greater power or, in the case of equal powers,

Table 8: Lords

	Author	Power
1	HARARY, FRANK	7043.46
2	ALON, NOGA	6773.39
3	RUBEL, LEE A.	1643.83
4	COLBOURN, CHARLES J.	1151.59
5	SHELAH, SAHARON	753.12
6	KAC, MARK	252.39
7	CHUI, CHARLES KAM-TAI	235.28
8	MAULDIN, R. DANIEL	214.22
9	JOO, ISTVAN	198.73
10	BRENNER, JOEL LEE	169.29

if it has greater degree. Again, we obtain for 'lords' a permutation of the same authors. In $\mathcal{E}$ the only lord is Paul Erdős.

6 Blockmodeling

To uncover the internal structure of the main core of $\mathcal{E}'$ we applied *pre-specified blockmodeling* (Batagelj, 1997, Batagelj, Ferligoj and Doreian, 1998) to it. As a block model we selected:

- cliques as diagonal blocks,
- complete or empty out-diagonal blocks.

We obtained a partition into 9 clusters – the smallest contains only 1 author, the largest 8 authors. All clusters are cliques.

$\mathbf{C}_1 = \{$ ALON, CHUNG, FRANKL, FUREDI, GRAHAM, KLEITMAN $\}$
$\mathbf{C}_2 = \{$ BABAI, SIMONOVITS, SOS, SZEMEREDI $\}$
$\mathbf{C}_3 = \{$ BOLLOBAS, SPENCER, WEST $\}$
$\mathbf{C}_4 = \{$ BURR, HARARY, NESETRIL, RODL $\}$
$\mathbf{C}_5 = \{$ CHARTRAND, GODDARD, KUBICKA, KUBICKI, OELLERMANN $\}$
$\mathbf{C}_6 = \{$ CHEN, FAUDREE, GOULD, GYARFAS, JACOBSON, LEHEL, SCHELP, Lesniak $\}$
$\mathbf{C}_7 = \{$ HAJNAL, LOVASZ, PACH $\}$
$\mathbf{C}_8 = \{$ ROUSSEAU, TUZA $\}$
$\mathbf{C}_9 = \{$ TROTTER $\}$

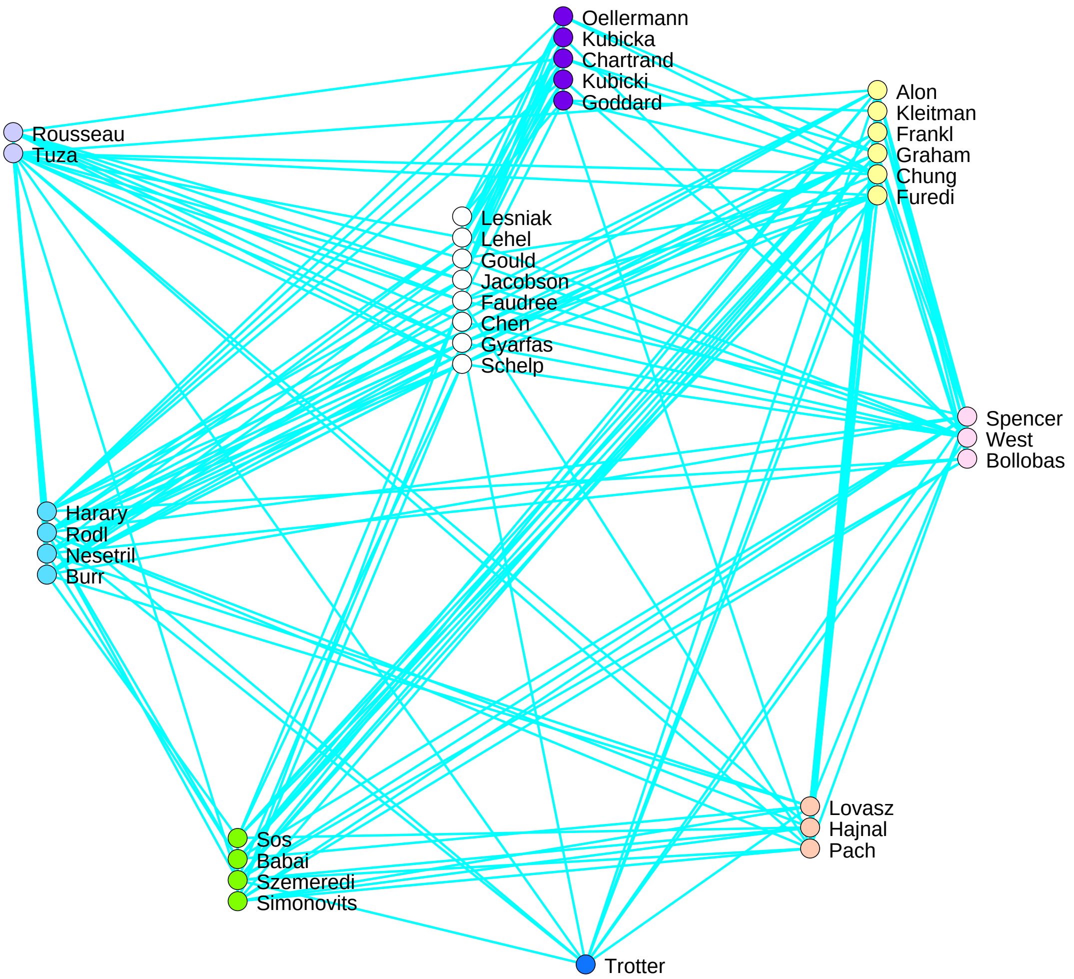

Figure 4: Cliques of the main core.

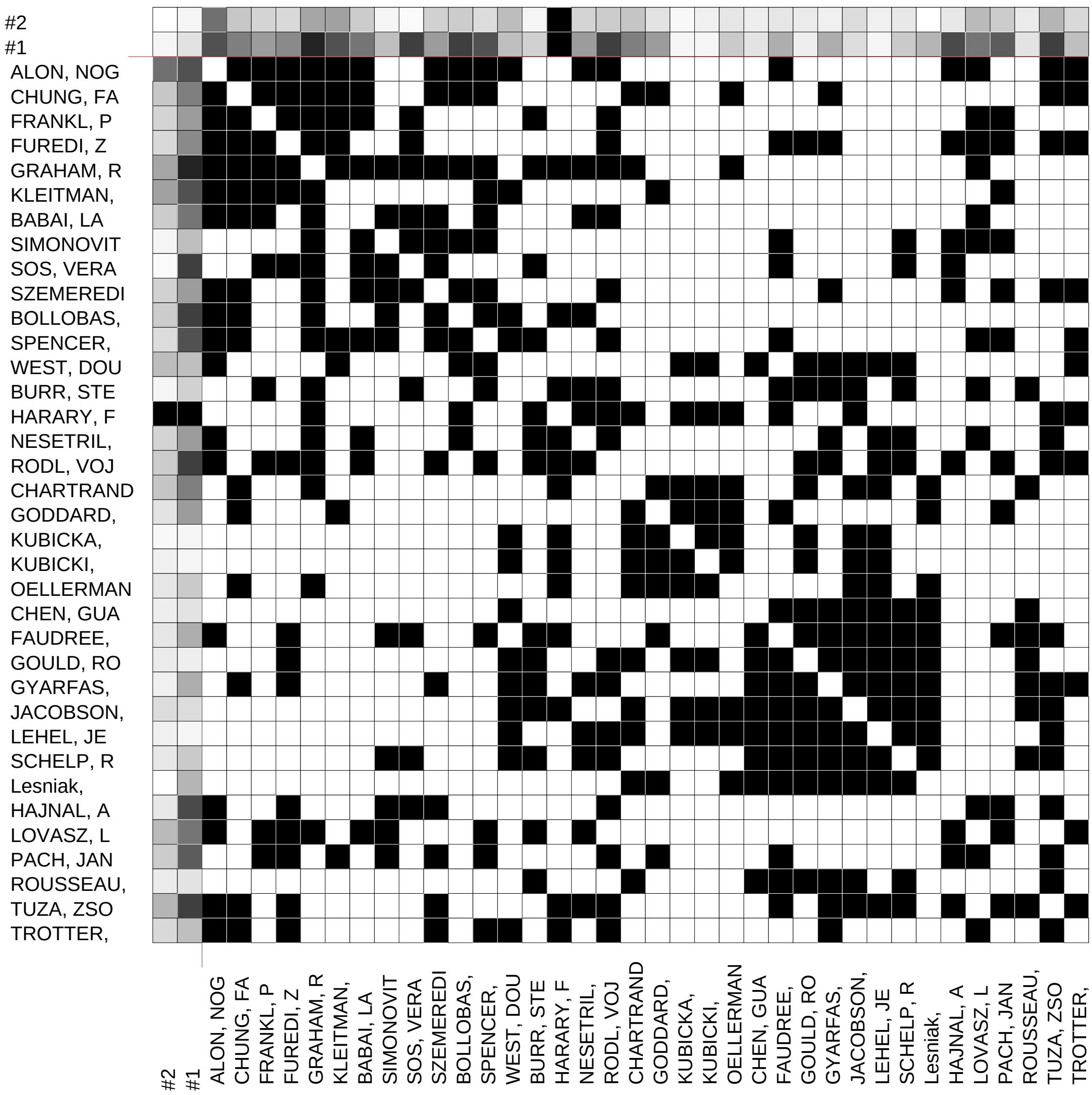

Figure 5: The main core matrix with context.

Figures 4 and 5 represent the blockmodeling results. In Figure 4 a view of 3D layout of the main core subgraph is given. A kinemage file (for **Mage** viewer) with the same layout is available at:

`http://vlado.fmf.uni-lj.si/pub/networks/doc/erdos/`

In Figure 5 an alternative visualization of this result, based on adjacency matrix (with context) reordered according the obtained clustering, is displayed. The first and the second row / column show the intensity (inside-cluster degree normalized with maximal inside-cluster degree) of collaboration of authors of the main core with the remaining authors with $\mathbf{n}_E = 2$ and $\mathbf{n}_E = 1$.

Coloring $\mathbf{C}_6$ with 8 colors, it is easy to extend this coloring to the main core and further to the remaining part of $\mathcal{E}'$. Therefore $\chi(\mathcal{E}') = 8$, and, since $\mathbf{n}_E$(Lesniak)$= 2$, $\chi(\mathcal{E}) \in \{8, 9\}$. Till now we didn't succeed to find an 8-coloring.

On cores we can build an efficient procedure for coloring large graphs that combines an exact procedure used on the main core, if it is small enough, and sequential coloring to extend the obtained core coloring to the remaining graph. The coloring order of vertices is determined by ordering them in decreasing order according to pairs $(\text{core}(v), \text{deg}(v))$.

7 Clustering

Another approach to analyze the graph is to introduce a dissimilarity d into the set V, or its subset, and apply some of multivariate techniques to it. Examples of such dissimilarities are ($\oplus$ denotes the symmetric difference and $D = \max\{|N(u)| + |N(v)| : u \neq v; u, v \in V\}$):

$$d_1(u,v) = \frac{|N(u) \oplus N(v)|}{D} \tag{6}$$

$$d_2(u,v) = \frac{|N(u) \oplus N(v)|}{|N(u) \cup N(v)|} \tag{7}$$

$$d_3(u,v) = \frac{|N(u) \oplus N(v)|}{|N(u)| + |N(v)|} \tag{8}$$

$$d_4(u,v) = \frac{\max(|N(u) \setminus N(v)|, |N(v) \setminus N(u)|)}{\max(|N(u)|, |N(v)|)} \tag{9}$$

These dissimilaries are in fact (semi)distances known in data analysis (some after transformation $d = 1 - s$ from similarity s into dissimilarity d) as dissimilarities of: d_1 – Hamming, Kendall, Sokal-Michner; d_2 – Jaccard; d_3 – Dice, Czekanowski; d_4 – Braun-Blanquet (Batagelj and Bren, 1995).

In the case $N(u) = N(v) = \emptyset$ we set for all four dissimilarities $d(u, v) = 1$. We obtain a parallel set of dissimilarities d_1^+, d_2^+, d_3^+ and d_4^+ by replacing in the above definitions neigborhoods N with rooted neighborhoods N^+.

Groups/clusters of similar units can be obtained by methods of cluster analysis (Gordon, 1981). We determined d_2^+ on $\{8, 9, 10\}$-core considering all authors from $\mathcal{E}$ and applied hierarchical clustering, Ward's method to it. Again, because of incomplete data for authors with $\mathbf{n}_E = 2$, their dissimilarities are relative to $\mathcal{E}$. The obtained dendrogram (clustering tree) is given in Figure 6.

8 Final remarks

In the paper we presented some possible approaches to analysis of large networks and applied them to the Erdős collaboration graph. Because of incomplete data for authors with $\mathbf{n}_E = 2$ the results are valid only for authors with $\mathbf{n}_E = 1$, or they should be interpreted for each group separately.

For better interpretation of the obtained results and for further analyses additional information about authors (year of birth, subjects of interest, geographic location, nationality, $\ldots$) and papers connecting them (number of papers, list of MR categories, $\ldots$) would be needed.

The (truncated) Erdős collaboration graph in Pajek format and some files with results are available at URL:

`http://vlado.fmf.uni-lj.si/pub/networks/doc/erdos/`

This paper is an elaborated version of the talk presented at the *Fourth Slovene International Conference in Graph Theory*, June 28 – July 2, 1999, Bled, Slovenia.

Acknowledgment: This work was supported by the Ministry of Science and Technology of Slovenia, Project J1-8532.

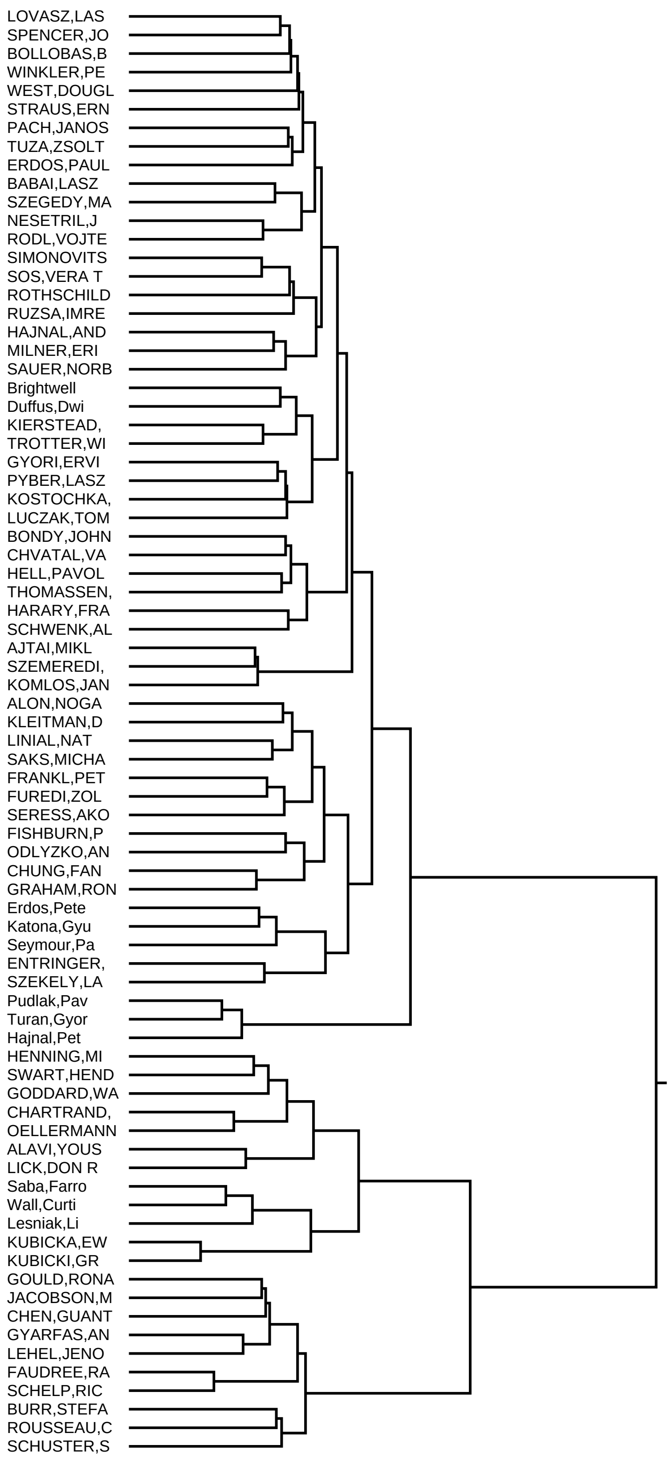

Figure 6: $\{8, 9, 10\}$-core clustering.

References

Batagelj, V., 1997. Notes on Blockmodeling. *Social Networks* **19**, 143-155.

Batagelj, V., Bren, M., 1995. Comparing Resemblance Measures. *Journal of Classification* 12(1), 73-90.

Batagelj, V., Ferligoj, A., and Doreian, P., 1998. Fitting Pre-Specified Blockmodels, in *Data Science, Classification, and Related Methods*, Eds., C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, and Y. Baba, Springer-Verlag, Tokyo, 199-206.

Batagelj, V., Mrvar, A., 1998. Pajek – A Program for Large Network Analysis. *Connections* **21 (2)**, 47-57.

Batagelj, V., Mrvar, A., Zaveršnik, M., 1999. Partitioning Approach to Visualization of Large Networks. To appear in *Proceedings of Graph Drawing '99*. LNCS, Springer.

Gordon, A.D., 1981. *Classification*. Chapman and Hall, London.

Grossman, J.W., Ion, P.D.F., 1995. On a Portion of the Well-Known Collaboration Graph. Proc. 26th Southeastern Inter. Conf. on Combinatorics, Graph Theory and Computing (Boca Raton, FL, 1995), *Congr. Numer.* **108**, 129-131.

Grossman, J.W., 1996. The Erdős Number Project. `http://www.oakland.edu/~grossman/erdoshp.html`

Seidman, S.B., 1983. Network Structure And Minimum Degree. *Social Networks* **5**, 269-287.