International laboratory for Applied Network Research

MASNA: Applied Statistics with Network Analysis

Moscow 2023, July 4

Analysis and visualization of complex networks in Pajek

14th Summer School 'Methods and tools for social network analysis'

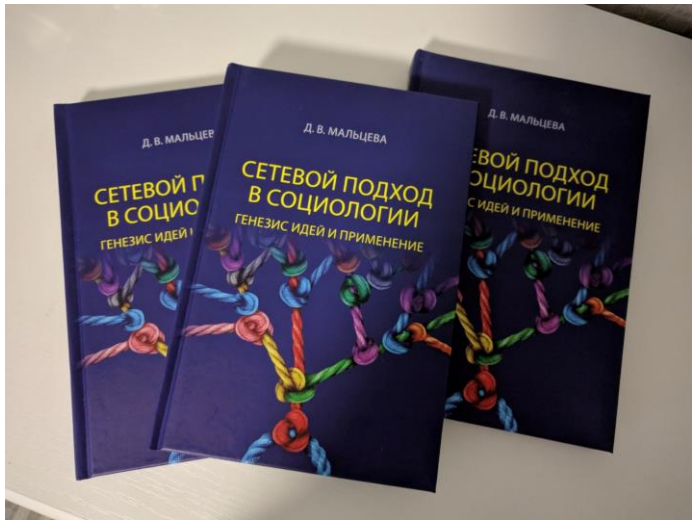
Daria Maltseva, Ph.D.

About me

- Head of the International Laboratory for Applied Network Research
- Leading Research Fellow, Visiting Lecturer
- Ph.D. | Candidate of Sciences in Sociology
- Degree in Sociology, Russian State University for the Humanities
- Young Faculty Support Program (Group of Young Academic Professionals), Category "New Researchers" (2018-2019)
- Grant of the Russian Scientific Foundation “Collaboration patterns in the Russian sociological community: the structure of scientific schools and their growth potential” 2021-2023
- Have a 2.5-year-old son

E-mail: dmaltseva@hse.ru

Expertise



Сетевой подход как феномен социологической теории

версия для печати

Мальцева Д. В.
Кандидат социологических наук, зам. заведующего, Международная лаборатория прикладного сетевого анализа НИУ «Высшая школа экономики», Москва, Россия d_malceva@mail.ru

DOI: 10.7868/S0132162518040013
ID статьи на сайте журнала: 7135

Рубрика: Методология и методы социологических исследований

Ссылка при цитировании:

***Мальцева Д. В.* Сетевой подход как феномен социологической теории // Социологические исследования. 2018. № 4. С. 3-14.**
DOI: 10.7868/S0132162518040013

Текст статьи.

Аннотация

Представлен сравнительный анализ трех направлений сетевого подхода в социологии – анализа социальных сетей, реляционной социологии и акторно-сетевой теории. Основаниями для анализа выступают время и контекст появления, основные теоретические положения, методология и методы эмпирических исследований. Делаются выводы о сходстве (до определенной степени) направлений реляционной социологии и акторно-сетевой теории и их отличия от анализа социальных сетей. Каждое из направлений автономно и занимает определенное место в структуре социологического знания. Делается вывод о некорректности представления «сетевого подхода» как единого теоретического блока, объединения его направлений под унифицирующими названиями «сетевая теория», «сетевая парадигма» и др. Подчеркнуто, что это собирательное название для разных теорий и подходов, оперирующих понятием сети в разных смыслах.

Springer Link

Published: 19 April 2022

Collaboration between authors in the field of social network analysis

Daria Maltseva & Vladimir Batagelj

Scientometrics (2022) | Cite this article
267 Accesses | **2** Altmetric | Metrics

Abstract

This paper presents a study of authors writing articles in the field of SNA and groups the
means of bibliographic network analysis. The dataset consists of works from the Web of
Science database obtained by searching for "social network*", works highly cited in the f
works published in the flagship SNA journals, and written by the most prolific authors
(70,000+ publications and 93,000+ authors), up to and including 2018. Using a two-mo
network linking publications with authors, we constructed and analysed different types

Springer Link

Published: 30 August 2019

Social network analysis as a field of invasions: bibliographic approach to study SNA development

Daria Maltseva & Vladimir Batagelj

Scientometrics **121**, 1085–1128 (2019) | Cite this article
994 Accesses | **10** Citations | **1** Altmetric | Metrics

Abstract

In this paper, the results of a study on the development of social network analysis (SNA) and its evolution over time, using the analysis of bibliographic networks are presented. The dataset consists of articles from the Web of Science Clarivate Analytics database obtained by

Springer Link

Published: 25 January 2020

Towards a systematic description of the field using keywords analysis: main topics in social networks

Daria Maltseva & Vladimir Batagelj

Scientometrics **123**, 357–382 (2020) | Cite this article
806 Accesses | **8** Citations | **1** Altmetric | Metrics

Abstract

This paper presents the results of the analysis of keywords used in Social Network Analysis (SNA) articles included in the WoS database and main SNA journals, from 1970 to 2018.

Springer Link

Published: 25 February 2021

Journals publishing social network analysis

Daria Maltseva & Vladimir Batagelj

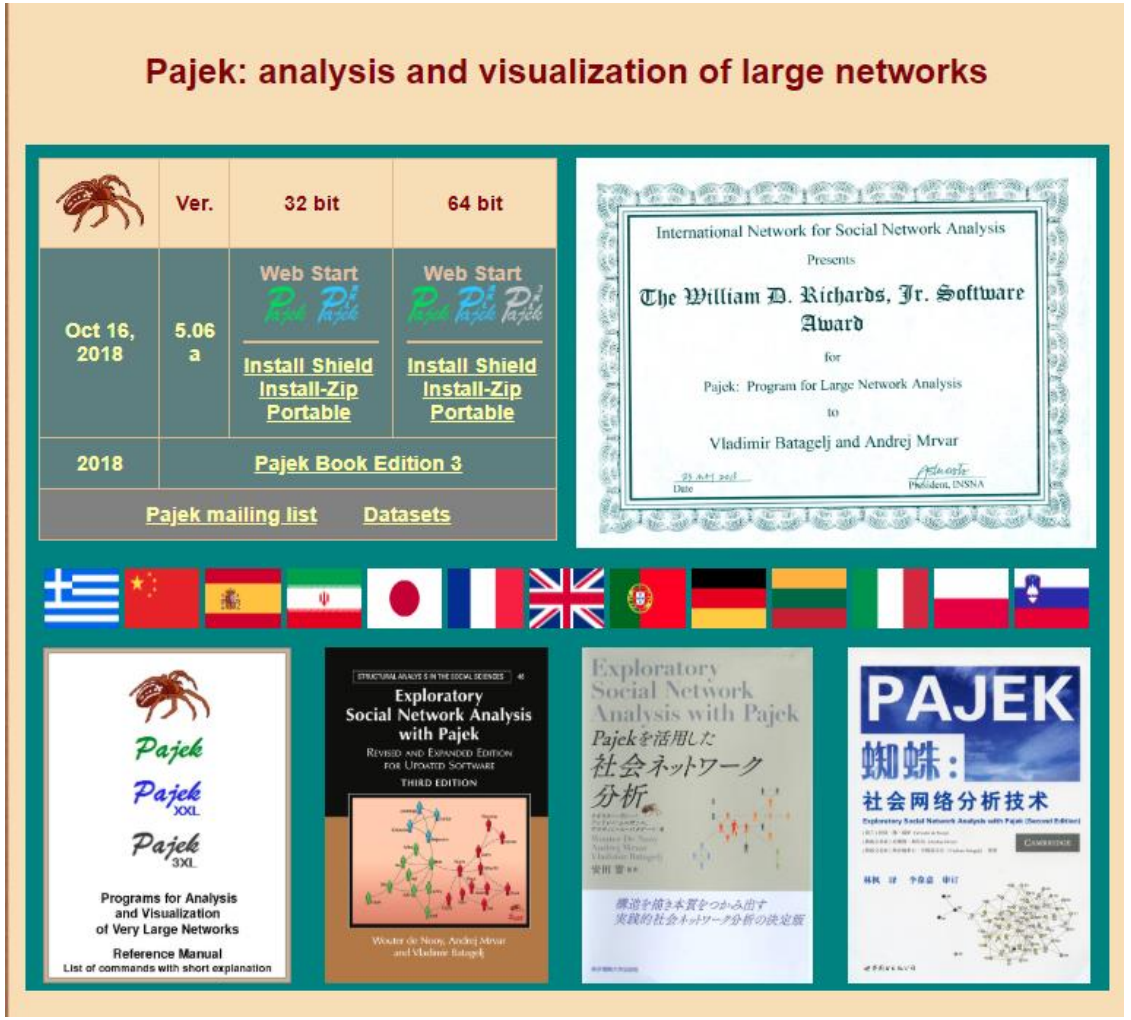
Scientometrics **126**, 3593–3620 (2021) | Cite this article
585 Accesses | **3** Citations | **1** Altmetric | Metrics

Abstract

This paper presents the analysis of journals publishing articles on social network analysis (SNA). The dataset consists of articles from the Web of Science database obtained by searching for "social network*", works intensively cited, written by the most prominent

Pajek

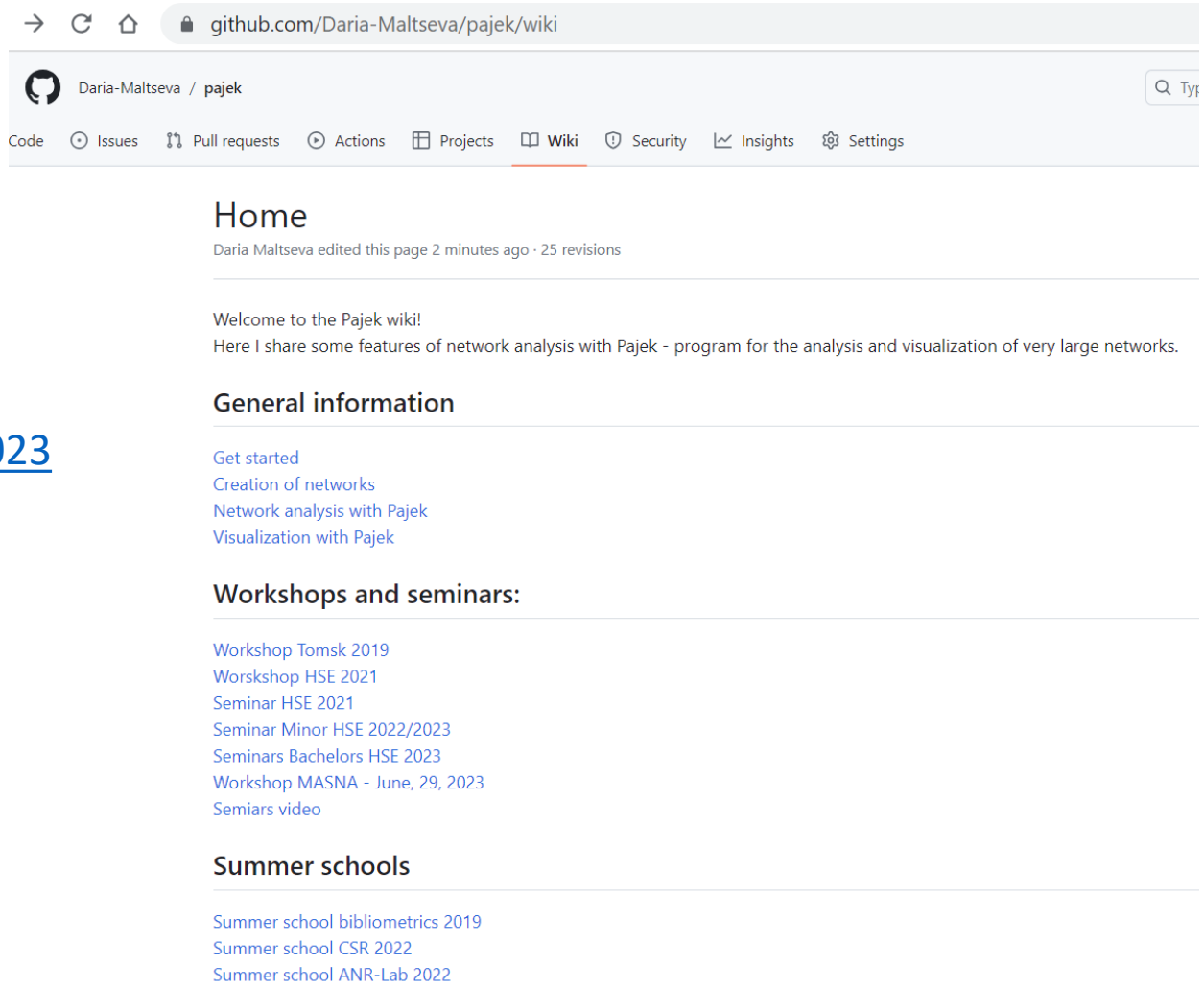
- Pajek is a program, for Windows, for analysis and visualization of large networks having some thousands or even millions of vertices. The latest version of Pajek is freely available, for noncommercial use.
- Link: http://mrvar.fdv.uni-lj.si/pajek/
- Creators: Andrej Mrvar and Vladimir Batagelj



Pajek

Pajek materials:
https://github.com/Daria-Maltseva/pajek/wiki

Summer school ANR-Lab 2022
https://github.com/Daria-Maltseva/pajek/wiki/14ss2023



Workshop plan

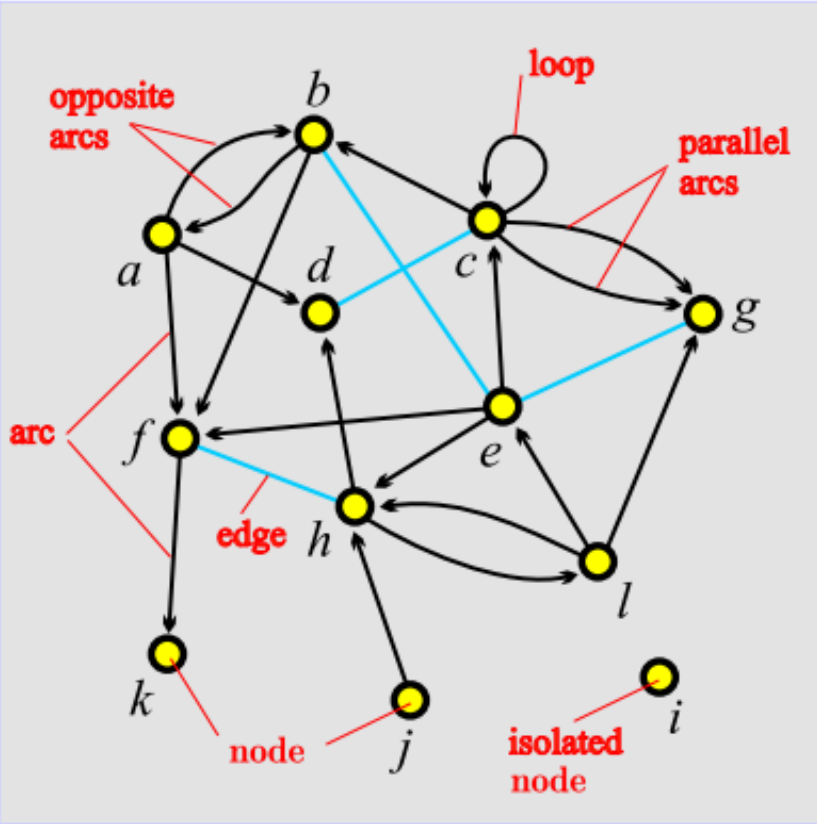
1. Overview of Pajek main capabilities observing:

- Basic operations in Pajek:
  - Creation of networks
  - Calculation of the basic metrics for individual nodes and whole networks
  - Networks visualization

- Advanced operations in Pajek:
  - Two-mode networks and multiplication
  - Acyclic networks
  - Temporal networks
  - Blockmodeling

2. Using Pajek in practice: Studying SNA development with bibliometric network analysis

Network definition



Visualization by: Vladimir Batagelj

The **network** is based on two sets - a set of **nodes** (vertices) representing the selected units of analysis and a set of **lines** (links) representing the connections between the units of analysis, which together form a **graph**.

The line can be directed (**arc**) or undirected (**edge**).

Nodes and lines can contain additional data - characteristics / **attributes** (name, type, value) - that can be measured or calculated.

Network = Graph + Data

Different types of networks

Network data	Network types
Directed vs. Undirected links	Networks with directional / non-directional / mixed links
One vs. several types of links	Multiple relations (multiplex) - networks with several types of links
One vs. several types of nodes	2-mode - bimodal networks
	Multilevel - networks with several types of links and nodes
Single static vs. dynamic (panel) monitoring	Temporal, dynamic networks
Unique links (genealogical)	Specialized networks (p-graphs, Petri net)

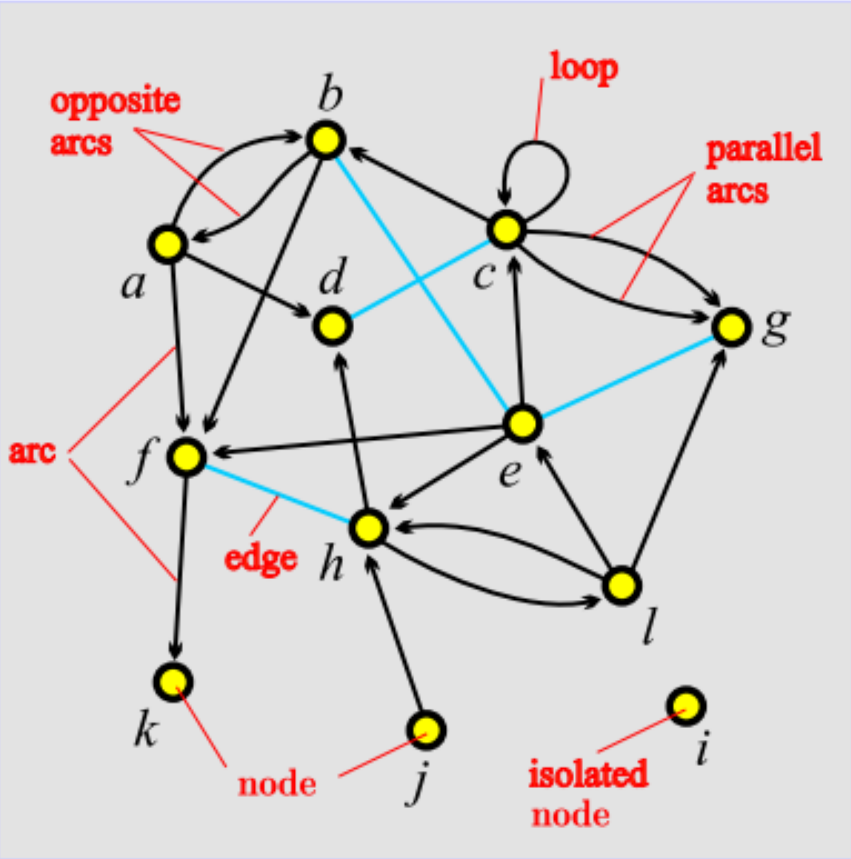
Different types of networks

Network data	Network types
Directed vs. Undirected links	**Networks with directional / non-directional / mixed links**
One vs. several types of links	Multiple relations (multiplex) - networks with several types of links
One vs. several types of nodes	**2-mode - bimodal networks**
	Multilevel - networks with several types of links and nodes
Single static vs. dynamic (panel) monitoring	Temporal, dynamic networks
Unique links (genealogical)	Specialized networks (p-graphs, Petri net)

1. Overview of Pajek main capabilities:

- Basic operations in Pajek:
  - Creation of networks
  - Calculation of the basic metrics for individual nodes and whole networks
  - Networks visualization

Network data format: set of nodes and links

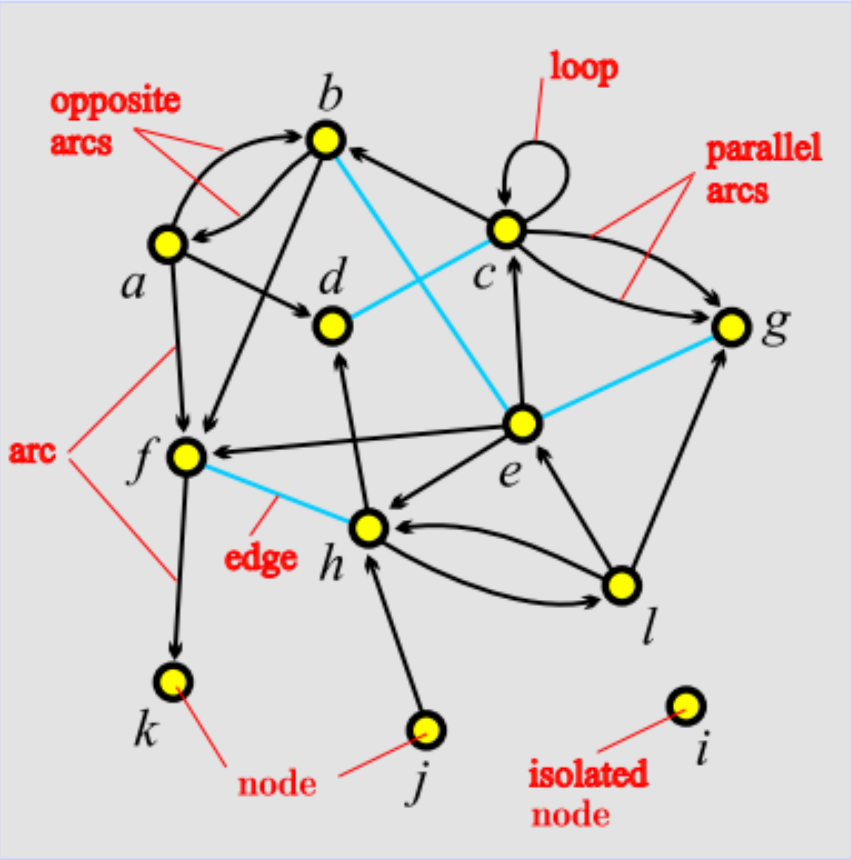


$$
\begin{aligned}
\mathcal{V} &= \{a, b, c, d, e, f, g, h, i, j, k, l\} \\
\mathcal{A} &= \{(a,b),(a,d),(a,f),(b,a), \\
&\quad (b,f),(c,b),(c,c),(c,g)_1, \\
&\quad (c,g)_2,(e,c),(e,f),(e,h), \\
&\quad (f,k),(h,d),(h,l),(j,h), \\
&\quad (l,e),(l,g),(l,h)\} \\
\mathcal{E} &= \{(b\colon e),(c\colon d),(e\colon g),(f\colon h)\} \\
\mathcal{G} &= (\mathcal{V}, \mathcal{A}, \mathcal{E}) \\
\mathcal{L} &= \mathcal{A} \cup \mathcal{E}
\end{aligned}
$$

```
*Vertices  12
  1 "a"     0.1020   0.3226
  2 "b"     0.2860   0.0876
  3 "c"     0.5322   0.2304
  4 "d"     0.3259   0.3917
  5 "e"     0.5543   0.4770
  6 "f"     0.1552   0.6406
  7 "g"     0.8293   0.3249
  8 "h"     0.4479   0.6866
  9 "i"     0.8204   0.8203
 10 "j"     0.4789   0.9055
 11 "k"     0.1175   0.9032
 12 "l"     0.7095   0.6475
*Arcs
  1  2
  2  1
  1  4
  1  6
  2  6
  3  2
  3  3
  3  7
  3  7
  5  3
  5  6
  5  8
  6 11
  8  4
 10  8
 12  5
 12  7
  8 12
 12  8
*Edges
  2  5
  3  4
  5  7
  6  8
```

Link weight can be added (third column)

Network data format: links of nodes with “neighbors”

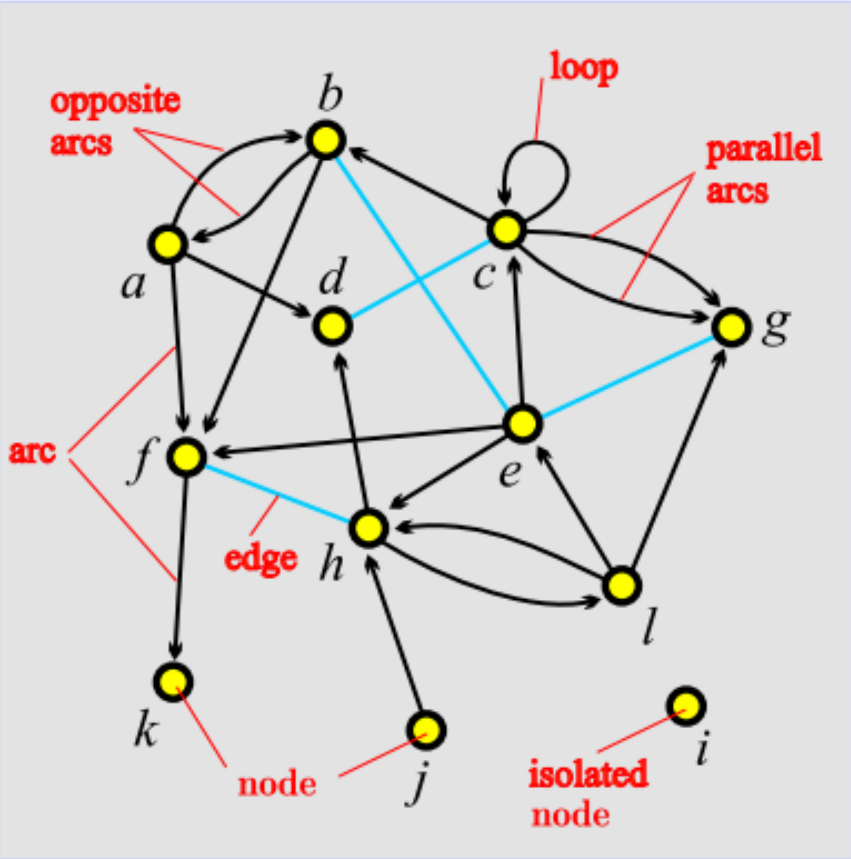


$$N_A(a) = \{b, d, f\}$$
$$N_A(b) = \{a, f\}$$
$$N_A(c) = \{b, c, g, g\}$$
$$N_A(e) = \{c, f, h\}$$
$$N_A(f) = \{k\}$$
$$N_A(h) = \{d, l\}$$
$$N_A(j) = \{h\}$$
$$N_A(l) = \{e, g, h\}$$

$$N_E(e) = \{b, g\}$$
$$N_E(c) = \{d\}$$
$$N_E(f) = \{h\}$$

```
*Vertices  12
  1 "a"    0.1020   0.3226
  2 "b"    0.2860   0.0876
  3 "c"    0.5322   0.2304
  4 "d"    0.3259   0.3917
  5 "e"    0.5543   0.4770
  6 "f"    0.1552   0.6406
  7 "g"    0.8293   0.3249
  8 "h"    0.4479   0.6866
  9 "i"    0.8204   0.8203
 10 "j"    0.4789   0.9055
 11 "k"    0.1175   0.9032
 12 "l"    0.7095   0.6475
*Arcslist
  1  2  4  6
  2  1  6
  3  2  3  7  7
  5  3  6  8
  6 11
  8  4 12
 10  8
 12  5  7  8
*Edgeslist
  2  5
  3  4
  5  7
  6  8
```

Network data format: adjacency matrix



	a	b	c	d	e	f	g	h	i	j	k	l
a	0	1	0	1	0	1	0	0	0	0	0	0
b	1	0	0	0	1	1	0	0	0	0	0	0
c	0	1	1	1	0	0	2	0	0	0	0	0
d	0	0	1	0	0	0	0	0	0	0	0	0
e	0	1	1	0	0	1	1	1	0	0	0	0
f	0	0	0	0	0	0	0	1	0	0	1	0
g	0	0	0	0	1	0	0	0	0	0	0	0
h	0	0	0	1	0	1	0	0	0	0	0	1
i	0	0	0	0	0	0	0	0	0	0	0	0
j	0	0	0	0	0	0	0	1	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0
l	0	0	0	0	1	0	1	1	0	0	0	0

Difficulty with distinguishing between displaying two reciprocal links (a, b) and non-directional links (b, e)

```
*Vertices  12
  1 "a"     0.1020   0.3226
  2 "b"     0.2860   0.0876
  3 "c"     0.5322   0.2304
  4 "d"     0.3259   0.3917
  5 "e"     0.5543   0.4770
  6 "f"     0.1552   0.6406
  7 "g"     0.8293   0.3249
  8 "h"     0.4479   0.6866
  9 "i"     0.8204   0.8203
 10 "j"     0.4789   0.9055
 11 "k"     0.1175   0.9032
 12 "l"     0.7095   0.6475
*Matrix
 0 1 0 1 0 1 0 0 0 0 0 0
 1 0 0 0 1 1 0 0 0 0 0 0
 0 1 1 1 0 0 2 0 0 0 0 0
 0 0 1 0 0 0 0 0 0 0 0 0
 0 1 1 0 0 1 1 1 0 0 0 0
 0 0 0 0 0 0 0 1 0 0 1 0
 0 0 0 0 1 0 0 0 0 0 0 0
 0 0 0 1 0 1 0 0 0 0 0 1
 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 1 0 1 1 0 0 0 0
```

Network data format: adjacency matrix

Binary matrix - no connection weight, only the presence / absence of the link itself

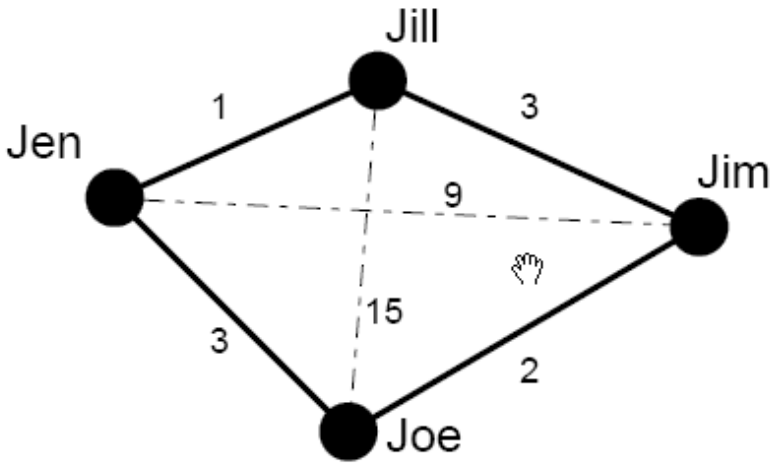
Friendship

	Jim	Jill	Jen	Joe
Jim	-	1	0	1
Jill	1	-	1	0
Jen	0	1	-	1
Joe	1	0	1	-

Weighted matrix - presence of link weight

Proximity

	Jim	Jill	Jen	Joe
Jim	-	3	9	2
Jill	3	-	1	15
Jen	9	1	-	3
Joe	2	15	3	-



Additional information

- **clu** – clustering: partition of nodes – nominal or ordinal data about nodes (node belongs to the cluster/group)
- **vec** – vector: numeric data about nodes (the property has value on node);
- **per** – permutation: ordering of nodes (node is at the certain position)

When collecting the network data consider to provide as much properties as possible.

Wolfe Monkey Data

inter.net

```
*Vertices  20
  1 "m01"
  2 "m02"
  3 "m03"
  4 "m04"
  5 "m05"
  6 "f06"
  7 "f07"
  8 "f08"
  9 "f09"
 10 "f10"
 11 "f11"
 12 "f12"
 13 "f13"
 14 "f14"
 15 "f15"
 16 "f16"
 17 "f17"
 18 "f18"
 19 "f19"
 20 "f20"
*Edges
  1   2   2
  1   3  10
  1   4   4
  -   -   -
```

inter.net

```
  1   6   5
  1   7   9
  1   8   7
  1   9   4
  1  10   3
  1  11   3
  1  12   7
  1  13   3
  1  14   2
  1  15   5
  1  16   1
  1  17   4
  1  18   1
  2   3   5
  2   4   1
  2   5   3
  2   6   1
  2   7   4
  2   8   2
  2   9   6
  2  10   2
  2  11   5
  2  12   4
  2  13   3
  2  14   2

  ...
```

sex.clu

```
*vertices 20
 1
 1
 1
 1
 1
 2
 2
 2
 2
 2
 2
 2
 2
 2
 2
 2
 2
 2
 2
```

age.vec

```
*vertices 20
 15
 10
 10
  8
  7
 15
  5
 11
  8
  9
 16
 10
 14
  5
  7
 11
  7
  5
 15
  4
```

rank.per

```
*vertices 20
  1
  2
  3
  4
  5
 10
 11
  6
 12
  9
  7
  8
 18
 19
 20
 13
 14
 15
 16
 17
```

0 is not allowed as node number

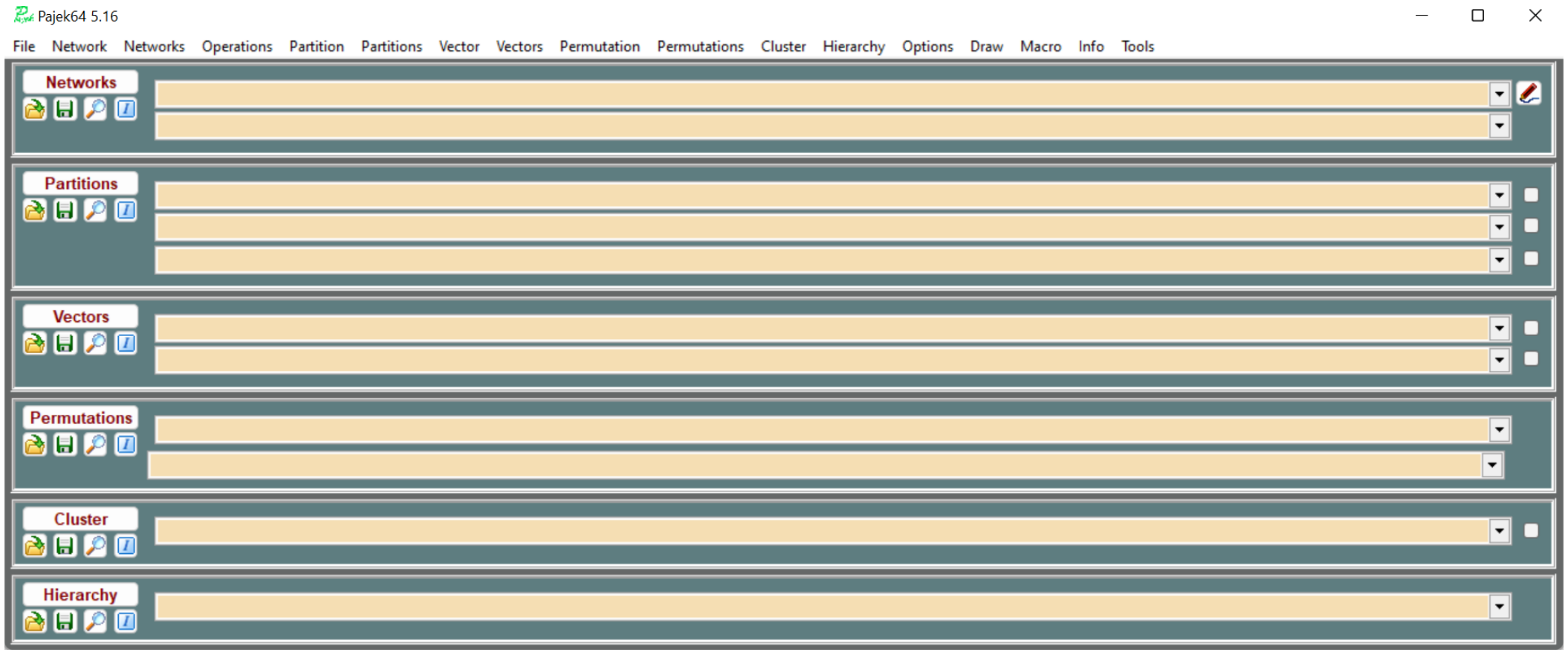
Pajek

.net

.clu

.vec

.per



Link to the workshop materials:
https://github.com/Daria-Maltseva/pajek/wiki/14ss2023

1. Overview of Pajek main capabilities observing:

- Advanced operations in Pajek:
  - Two-mode networks and multiplication
  - Acyclic networks
  - Temporal networks
  - Blockmodeling

1. Overview of Pajek main capabilities observing:

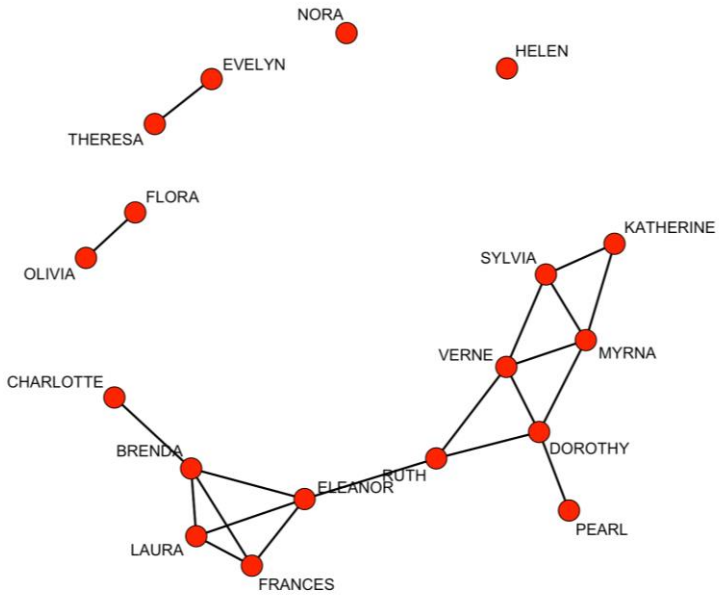
- Advanced operations in Pajek:
  - **Two-mode networks and multiplication**
  - **Acyclic networks**
  - Temporal networks
  - Blockmodeling

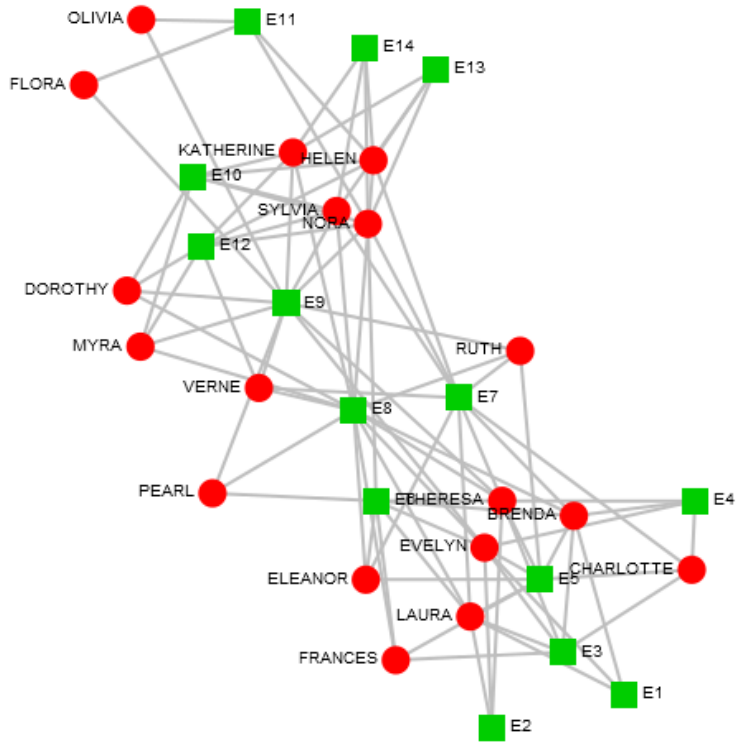
Network data format: other network types

Two-mode network:

It is necessary to distinguish between nodes belonging to mode 1 and mode 2:

```
*Vertices 5 2
1 "Author 1"
2 "Author 2"
3 "Book1"
4 "Book2"
5 "Article"
*Edges
1 3
1 5
2 4
2 5
```





Network data format: other network types

Temporal network:

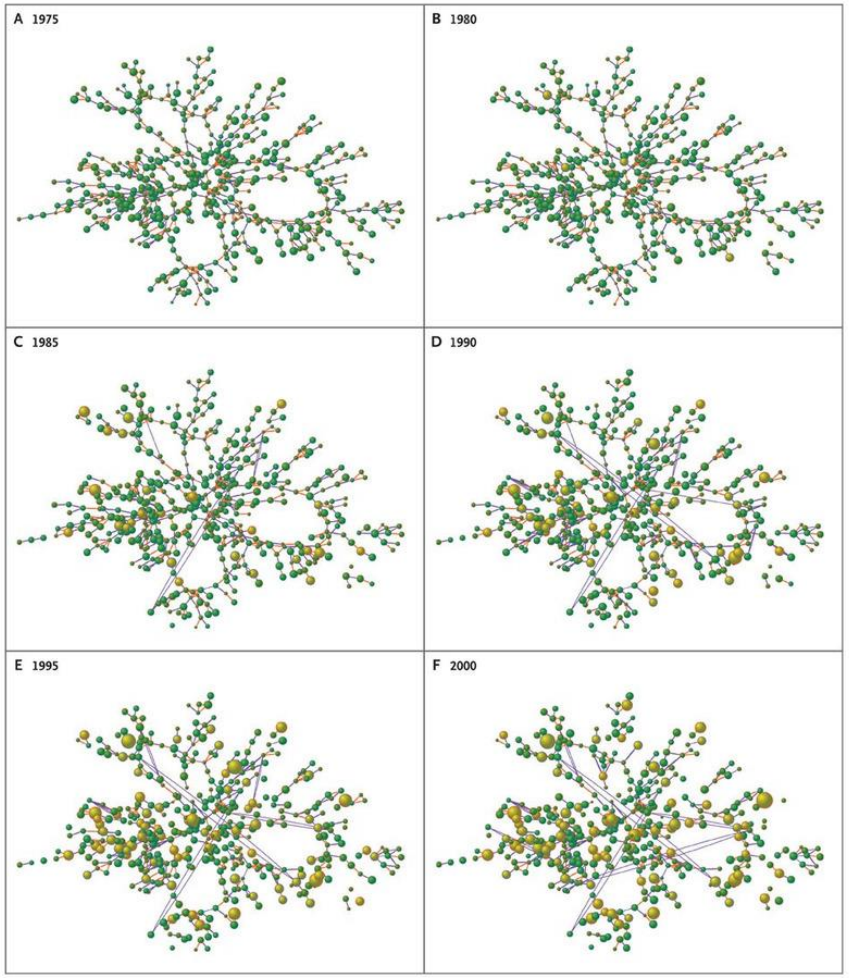
2 approaches:

- Set of networks for each time period
- Fixing the existence of a node or connection in a certain period:

```
*Vertices 3
1 "a" [5-10,12-14]
2 "b" [1-3,7]
3 "e" [4-*]
*Edges
1 2 1 [7]
1 3 1 [6-8]
```

Network (input) file formats should provide a means to express different types of networks. All important data for analysis should be recorded.

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England journal of medicine, 357(4), 370-379.
Видео



1. Overview of Pajek main capabilities observing:

- Advanced operations in Pajek:
  - **Two-mode networks and multiplication**
  - **Acyclic networks**
  - Temporal networks
  - Blockmodeling

Bibliometric network analysis

Bibliometrics is the application of mathematical and statistical methods to the study of books, periodicals, and other publications.

Types of bibliometric network analysis

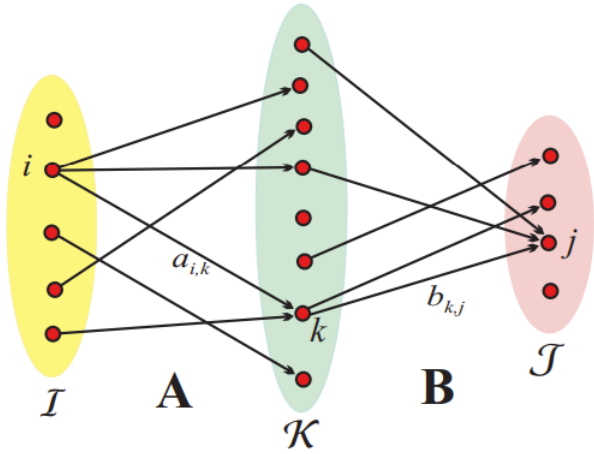
Based on citation:
- Citation
- Co-citation
- Bibliographic coupling

between works, journals, authors, teams of different levels

Based on co-authorship:
- Co-authorship
- Collaboration (normalized data)

Based on co-presence of other units of analysis:
- Keyword co-occurence

To create most networks, the **network multiplication** procedure is used:

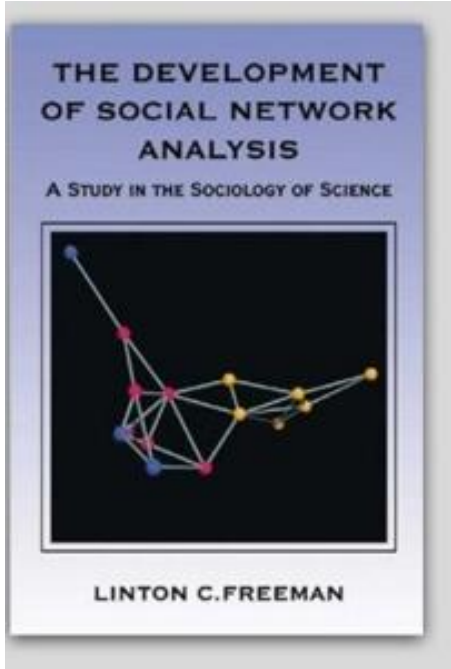


$$c_{i,j} = \sum_{k \in N_A(i) \cap N_B^-(j)} a_{i,k} \cdot b_{k,j}$$

If all weights in networks $\mathcal{N}_A$ and $\mathcal{N}_B$ are equal to 1 the value of $c_{i,j}$ counts the number of ways we can go from $i \in \mathcal{I}$ to $j \in \mathcal{J}$ passing through $\mathcal{K}$.

2. Using Pajek in practice:
Studying SNA development with bibliometric network analysis

Studying the current state of the SNA field



Linton Freeman – the following periods in SNA development:

- mid 19th century – 1920s – Prehistory
- 1930–1940s – "Birth and Death"
- 1940-1960s – "Dark Ages"
- 1970s – "Growing up"
- 1990s - the institutionalization of SNA
- 2000s - the "invasion" of physicists, Network Science

Freeman, L. (2004). The development of social network analysis. A Study in the Sociology of Science, 1(687), 159-167

Watts, D. J. (2004). The "new" science of networks. Annual review of sociology, 243-270.

The current state of the SNA field

2010s: New "invasions" into SNA - by behavioral biologists

Data:

- 70,792 records obtained for the query "social network*" in WoS
- 1, 297, 133 of works (hits and cited only)
- 93, 011 authors
- 32, 409 keywords
- 8, 943 journals

Analysis of one-mode and two-mode derived networks

Maltseva D., Batagelj V. Social network analysis as a field of invasions: bibliographic approach to study SNA development , Scientometrics. 2019. Vol. 121. No. 2. P. 1085-1128.

Main path analysis of network of citations between works *Cite*

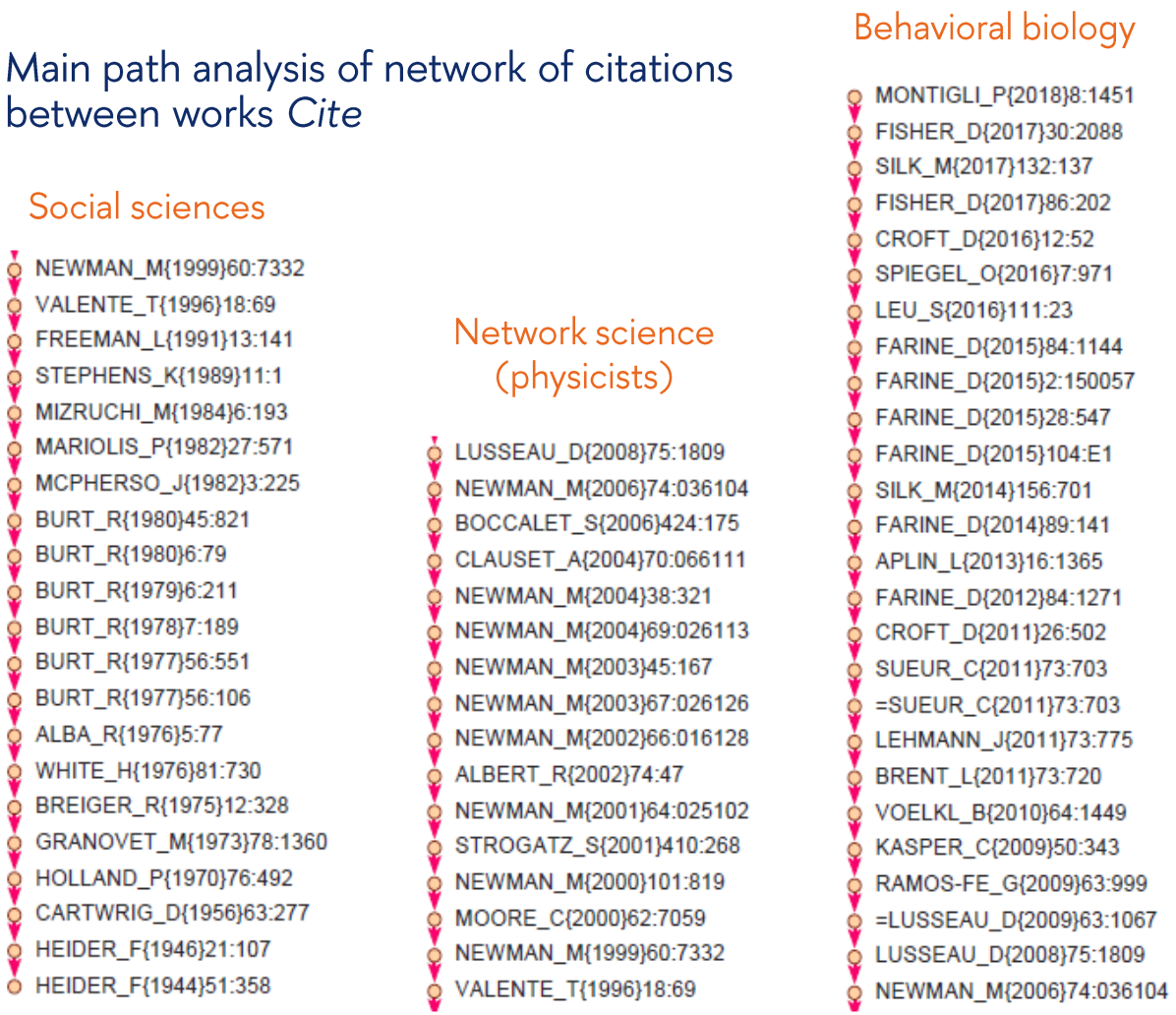


Figure 6: SPC net: Main path by fragments – sociology, physics, biology
(2nd and 3rd parts starts with two works from the previous group)

The current state of the SNA field

Number of publications for hits and cited only works per year

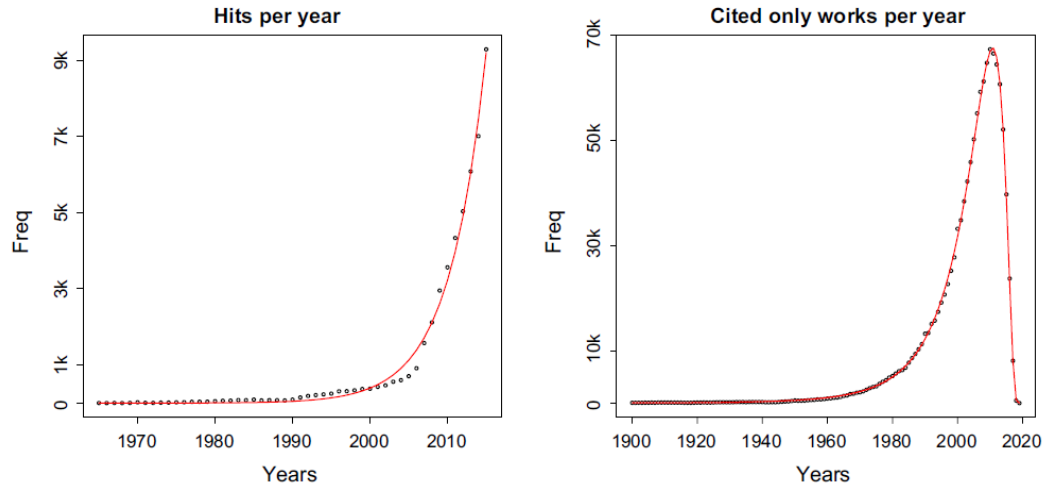


Fig. 2 CiteN: Distribution of hits (left) and terminal works (right) by years

Maltseva D., Batagelj V. Social network analysis as a field of invasions: bibliographic approach to study SNA development // Scientometrics. 2019. Vol. 121. No. 2. P. 1085-1128.

Top-25 most cited papers

Table 2 CiteN: The most cited works—indegree

i	Freq	id
1	5348	**WASSERMA_S(1994):**
2	4471	**GRANOVET_M(1973)78:1360**
3	2906	*WATTS_D(1998)393:440
4	2614	*BARABÁSI_A(1999)286:509
5	2561	**FREEMAN_L(1979)1:215**
6	2447	BOYD_D(2007)13:210
7	2429	**MCPHERSO_M(2001)27:415**
8	2330	**BURT_R(1992):**
9	1886	**COLEMAN_J(1988)94:95**
10	1572	*NEWMAN_M(2003)45:167
11	1520	*GIRVAN_M(2002)99:7821
12	1510	**PUTNAM_R(2000):**
13	1285	*ALBERT_R(2002)74:47
14	1240	**GRANOVET_M(1985)91:481**
15	1192	**SCOTT_J(2000):**
16	1171	**EVERETT_M(2002):**
17	1166	*NEWMAN_M(2004)69:026113
18	1093	**COLEMAN_J(1990):**
19	1058	STEINFIE_C(2007)12:1143
20	1034	*FORTUNAT_S(2010)486:75
21	999	**BORGATTI_S(2002):**
22	945	**CHRISTAK_N(2007)357:370**
23	867	**FREEMAN_L(1977)40:35**
24	854	**HANNEMAN_R(2005):**
25	800	**LIN_N(2001):**

The current state of the SNA field

(Sub)network of citations between journals

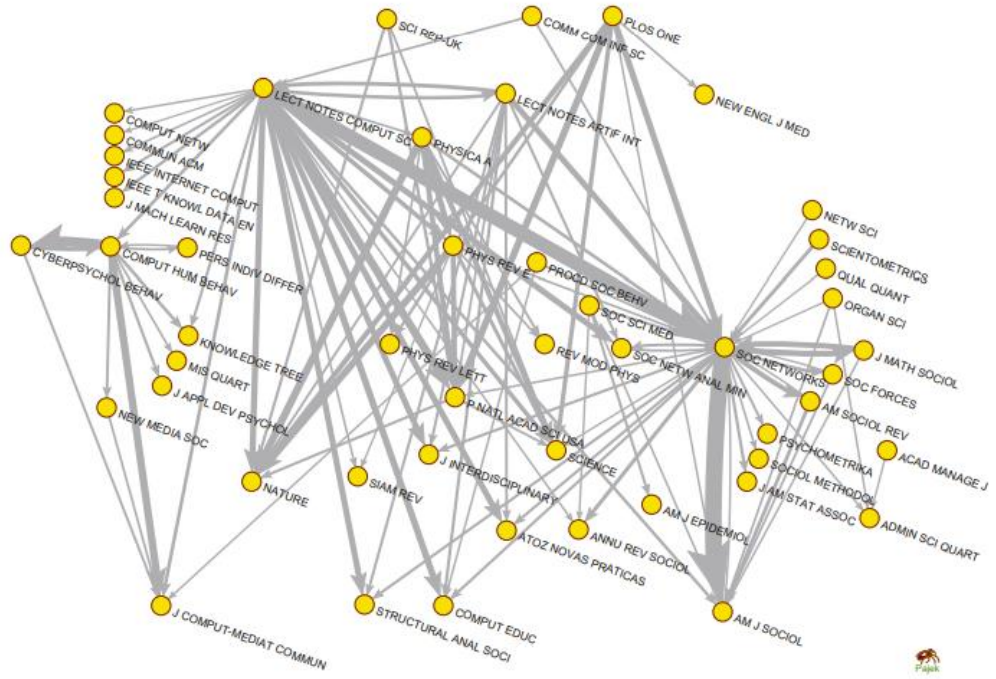


Fig. 11 Citations among journals—main island

Maltseva D., Batagelj V. Journals Publishing Social Network Analysis // Scientometrics. 2021. Vol. 126. No. 4. P. 3593-3620.

(Sub)network of bibliographic coupling between journals

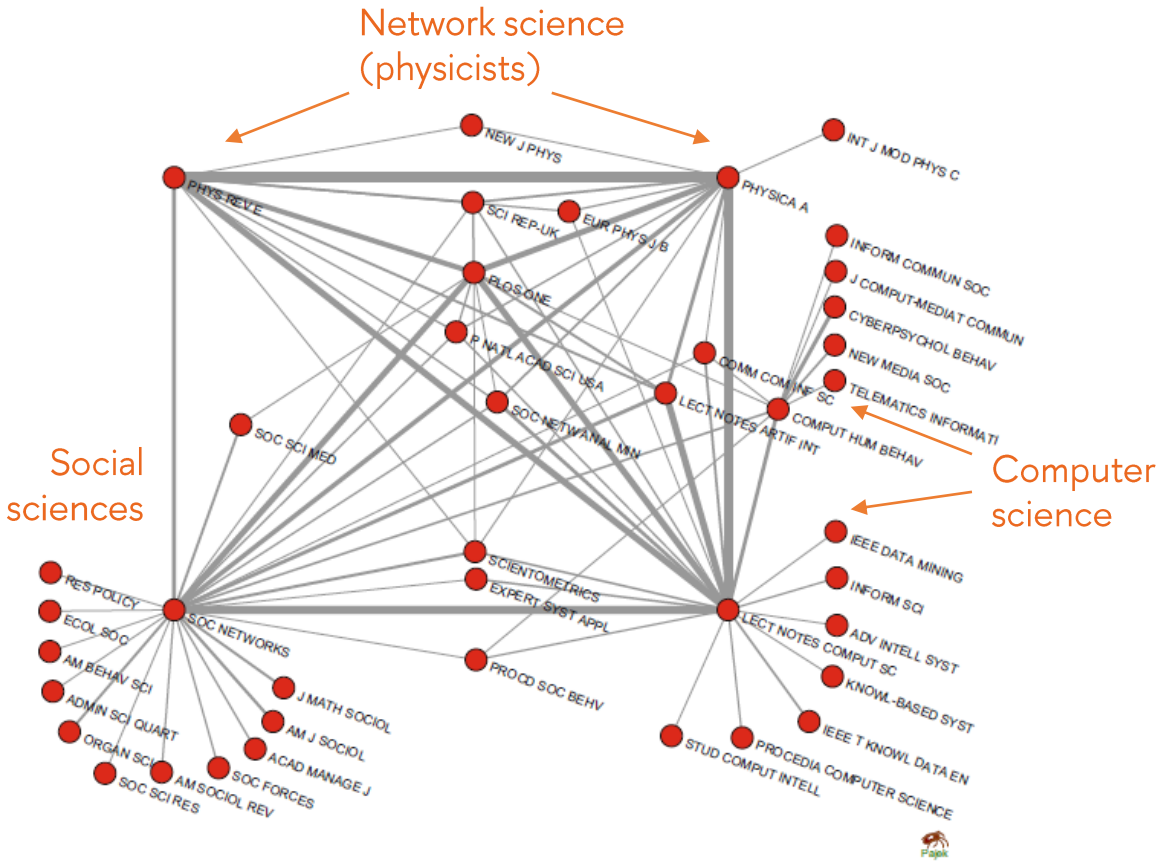


Fig. 12 Cj network: journals

The current state of the SNA field

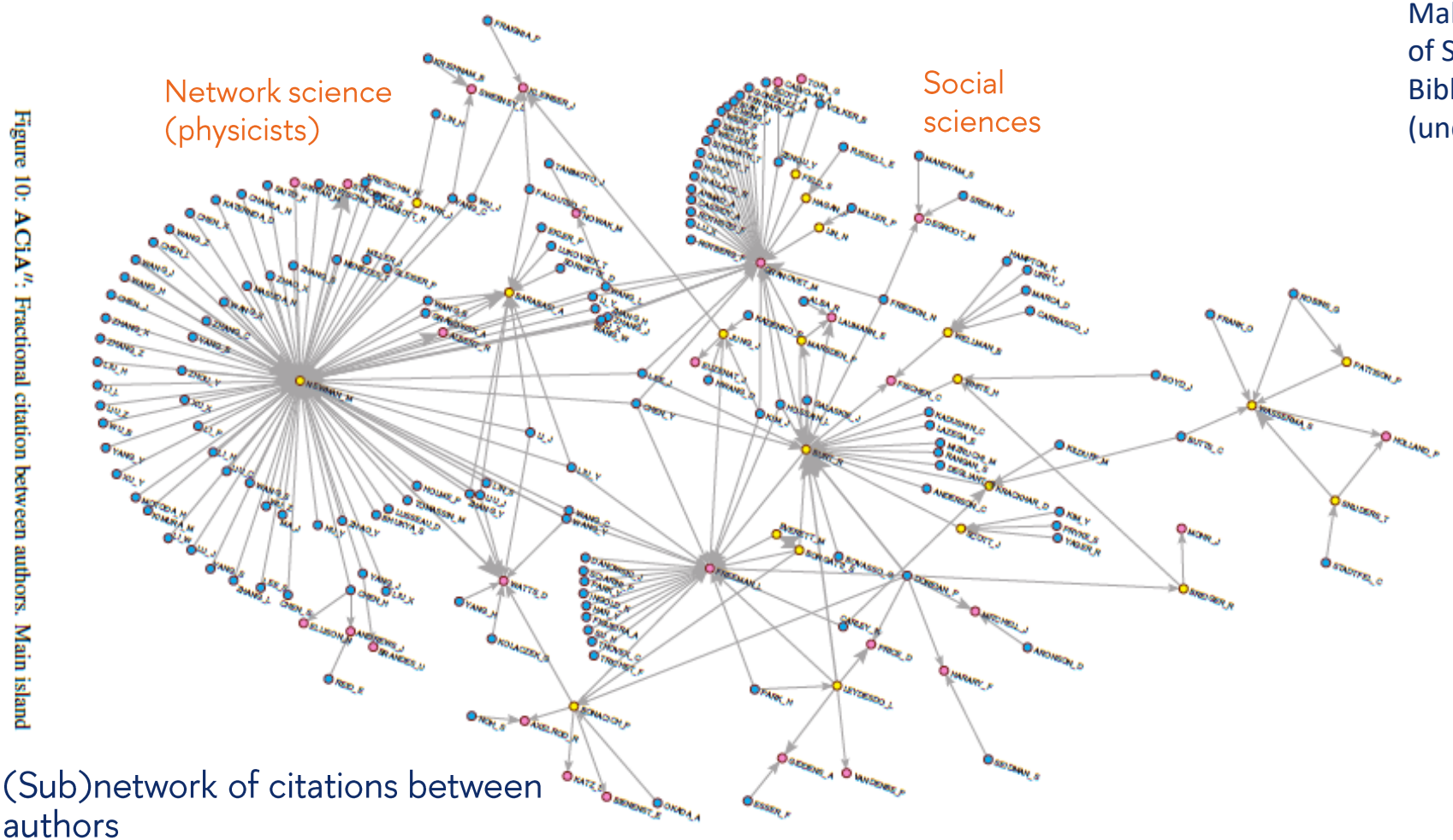


Figure 10: ACiA": Fractional citation between authors. Main island

(Sub)network of citations between authors

Maltseva D., Batagelj V. Authors in the Field of Social Network Analysis: Citation and Bibliographic Coupling. Scientometrics (under review)

The current state of the SNA field

(Sub)network of bibliographic coupling between authors

Network science (physicists)

Social sciences

Social sciences

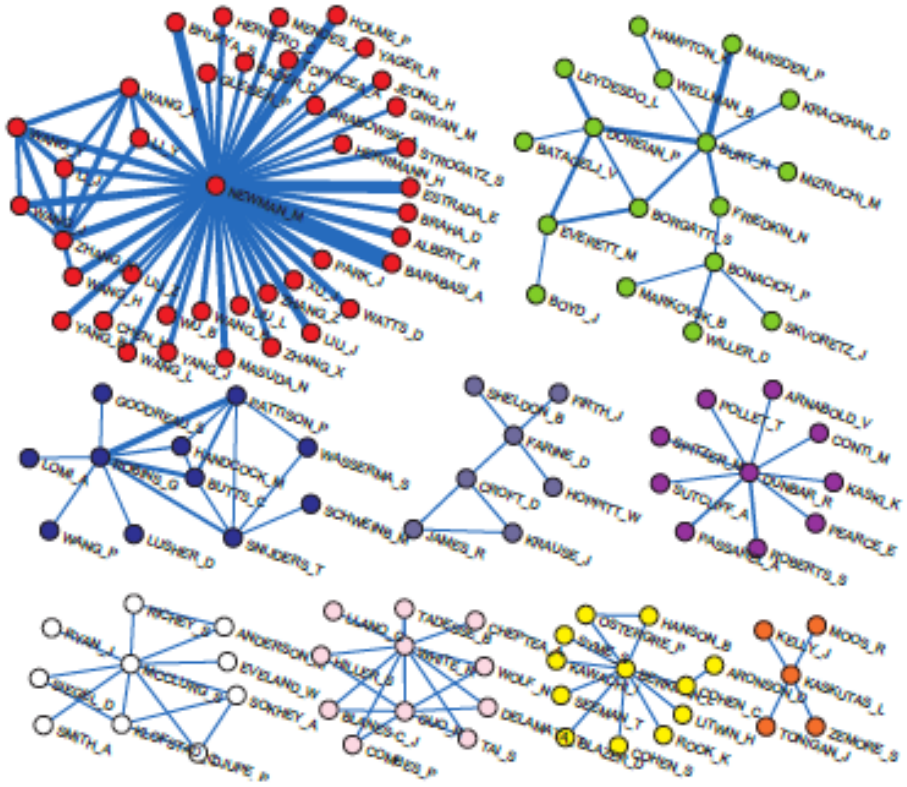


Figure 12: Jaccard network **ACoj**: General islands

Maltseva D., Batagelj V. Authors in the Field of Social Network Analysis: Citation and Bibliographic Coupling. Scientometrics (under review)

The current state of the SNA field: some conclusions

Is the field still growing?
Starting from its institutionalization in the 1980–90s, SNA has grown significantly, both in terms of the number of publications and the number of disciplines involved in the research using the SNA approach. The number of publications shows exponential growth, and on average it doubles every 3 years.

What are the disciplines involved?
The analysis confirmed the previous studies on SNA development using citation network analysis. Up to the middle of 1990s the most "important" works belong to social sciences, and starting from the 2000s the field experienced the "invasion of the physicists" leading to the establishment of Network science. To our surprise, from the 2010s both groups experience the "invasion" of scientists from a completely different field—animal SNA. This does not mean that either social scientists or physicists are no longer presented in the field—it means that the new group is more active both in the number of publications and citations of each other.

What are the most important and influential works?
Despite the "invasions", the most cited works still belong to the social scientists—with Wasserman, Faust, and Granovetter on the top. Other highly cited works are from social scientists, physicists, and computer scientists. The works of physicists are cited by the newly established group of the animal SNA.

Studying Pajek at MASNA

Vladimir Batagelj

- Introduction to Network Analysis
- Advanced Network analysis methods

The structure of the curriculum of the Master's programme "Applied Statistics with Network Analysis", 2023/2024

	1 module	2 module	3 module	4 module
1st year	— Introduction to Statistics — Programming in R and Python	— Probability Theory — Introduction to Network Analysis	— Data Mining	
1st year	— Contemporary Data Analysis: Methodology and Methods of Interdisciplinary Research		— Contemporary Decision Sciences: an Integrated Perspective	
1st year		— Applied Linear Models		
1st year			— Machine Learning — Exploratory Data Analysis	— Advanced Network Analysis Methods — Categorical Data Analysis — Multivariate Data Analysis
1st year			— Nonparametric Theory and Data Analysis	
1st year	— Term Paper — Research Seminars "Computational Social and Network Sciences" or "Application of Network Theory to Business Analytics and Social Networks"			
2nd year		— Methods of Statistical Consulting		
2nd year	— Stuctural Equation Modeling			
2nd year	— Unstructured Data Analysis — Time Series — Social Network Analysis with R		— Multilevel Models — Bayesian Data Analysis — Stochastic Models	
2nd year	— Statistical Methods of Social Network Analysis			
2nd year	— Research Seminar "Working with Network Data" — Project Seminar and working with projects			
2nd year			— Work Experience Internship	— Graduation Master Thesis

Key: Compulsory Courses | Electives | Practical Work (Compulsory)