Temporal bibliographic analysis

Vladimir Batagelj, Daria Maltseva

IMFM Ljubljana, IAM UP Koper and NRU HSE Moscow

SDA workshop:

Advances in data science for big and complex data

University Paris-Dauphine, January 10-11, 2019

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Daria Maltseva: d_malceva@mail.ru

Current version of slides (January 12, 2019 at 06 : 07): slides PDF

WoS

To the Web of Science (WoS), Clarivate Analytics's multidisciplinary databases of bibliographic information, we put the query

`"social network*"`

Additionally, all the articles from the following journals were collected:

`Social Networks, Network Science,`
`Social Network Analysis and Mining,`
`Journal of Complex Networks`

Other network-related journals are **not considered** in WoS:

`Computational Social Networks, Applied Network Science,`
`Online Social Networks and Media, Connections,`
`Journal of Social Structure`

We limited the search to the Web of Science Core Collection because for other data bases from WoS the CR-fields (containing citation information) can not be exported. The first data set was collected in 2007, second – in June, 2018.

WoS record

```
PT J
AU GRANOVET.MS
TI STRENGTH OF WEAK TIES
SO AMERICAN JOURNAL OF SOCIOLOGY
LA English
DT Article
C1 JOHNS HOPKINS UNIV, BALTIMORE, MD 21218 USA.
CR BARNES JA, 1969, SOCIAL NETWORKS URBA
   BECKER MH, 1970, AM SOCIOL REV, V35, P267
   BERSCHEID E, 1969, INTERPERSONAL ATTRAC
   BOISSEVAIN J, 1968, MAN, V3, P542
   BOTT E, 1957, FAMILY SOCIAL NETWOR
NR 61
TC 2156
PU UNIV CHICAGO PRESS
PI CHICAGO
PA 5720 S WOODLAWN AVE, CHICAGO, IL 60637
SN 0002-9602
J9 AMER J SOCIOL
JI Am. J. Sociol.
PY 1973
VL 78
IS 6
BP 1360
EP 1380
PG 21
SC Sociology
GA P7726
UT ISI:A1973P772600003
ER
SK IP
```

Converting WoS data into networks

We applied the program WoS2Pajek 1.5 to the collected data.

The following networks were constructed:

1. the authorship network *WA* on works × authors (from the field AU),
2. the journalship network *WJ* on works × journals (from the field CR or J9),
3. the keywordship network *WK* on works × keywords (from the field ID or DE or TI),
4. the citation network *Cite* on works (from the field CR).

We obtained also the following node properties:

1. the partition *year* of works by publication year,
2. the *DC* partition distinguishing between works with complete description (DC=1) and the cited only works (DC=0),
3. the vector of number of pages *NP*.

WoS

We call a *terminal* node a node without a description in the collected data set – it appears only in the WoS CR field as a reference.

We additionally collected on WoS and Google data for terminal nodes with large indegree in the citation network – highly cited works without description in the collected data set.

If a description of a node was not available in WoS we **manually** constructed a corresponding description **without** CR data (using RIS bibliographic format and converting it to WoS).

As the data set of 2007 was already completed, we made this additional search only for works 2008-* in July 2018.

Sizes of Original cleaned and Reduced networks

	# nodes (sum)	# nodes 1	# nodes 2	# arcs
CiteN	1,297,133			2,753,633
CiteR	**70,792**			398,199
WAn	1,693,104	1,297,133	395,971	1,442,240
WAr	163,803	**70,792**	**93,011**	215,901
WKn	1,329,542	1,297,133	32,409	1,167,670
WKr	103,201	**70,792**	**32,409**	1,167,666
WJn	1,366,279	1,297,133	69,146	720,044
WJr	79,735	**70,792**	**8,943**	61,741

An important property of all these networks is that they share as the first node set the same set of works (papers, reports, books, etc.) – they are *linked*.

Cite net

The most cited works - indegree

i	freq	id	i	freq	id
1	5348	**WASSERMA_S(1994):**	31	734	*NEWMAN_M(2001)98:404
2	4471	**GRANOVET_M(1973)78:1360**	32	719	*NEWMAN_M(2010):
3	2906	*WATTS_D(1998)393:440	33	701	PORTES_A(1998)24:1
4	2614	*BARABASI_A(1999)286:509	34	687	BLEI_D(2003)3:993
5	2561	**FREEMAN_L(1979)1:215**	35	670	**BURT_R(2004)110:349**
6	2447	BOYD_D(2007)13:210	36	654	HANSEN_M(1999)44:82
7	2429	MCPHERSO_M(2001)27:415	37	639	PALLA_G(2005)435:814
8	2330	**BURT_R(1992):**	38	634	*CLAUSET_A(2004)70:066111
9	1886	**COLEMAN_J(1988)94:95**	39	629	*BONACICH_P(1987)92:1170
10	1572	*NEWMAN_M(2003)45:167	40	628	ERDOS_P(1959)6:290
11	1520	*GIRVAN_M(2002)99:7821	41	628	UZZI_B(1997)42:35
12	1510	**PUTNAM_R(2000):**	42	628	**ROGERS_E(2003):**
13	1285	*ALBERT_R(2002)74:47	43	613	**PUTNAM_R(1993):**
14	1240	**GRANOVET_M(1985)91:481**	44	593	BERKMAN_L(1979)109:186
15	1192	**SCOTT_J(2000):**	45	583	**ZACHARY_W(1977)33:452**
16	1171	**EVERETT_M(2002):**	46	572	**BORGATTI_S(2009)323:892**
17	1166	NEWMAN_M(2004)69:026113	47	569	*NEWMAN_M(2001)64:025102
18	1093	**COLEMAN_J(1990):**	48	565	**BURT_R(2005):**
19	1058	STEINFIE_C(2007)12:1143	49	561	ADLER_P(2002)27:17
20	1034	FORTUNAT_S(2010)486:75	50	559	**CHRISTAK_N(2008)358:2249**
21	999	**BORGATTI_S(2002):**	51	555	**ROGERS_E(1995):**
22	945	**CHRISTAK_N(2007)357:370**	52	554	MILGRAM_S(1967)1:61
23	867	**FREEMAN_L(1977)40:35**	53	553	BARON_R(1986)51:1173
24	854	**HANNEMAN_R(2005):**	54	550	**GRANOVET_M(1978)83:1420**
25	800	**LIN_N(2001):**	55	539	**FISCHER_C(1982):**
26	757	KAPLAN_A(2010)53:59	56	537	BRIN_S(1998)30:107
27	756	*BLONDEL_V(2008):P10008	57	524	**MARSDEN_P(1990)16:435**
28	742	NAHAPIET_J(1998)23:242	58	523	KEMP_D(2003):137
29	740	FORNELL_C(1981)18:39	59	523	KLEINBER_J(1999)46:604
30	740	*NEWMAN_M(2006)103:8577	60	517	*BOCCALET_S(2006)424:175

Labels ending with : represent books

Temporal networks

A *temporal network* $\mathcal{N}_T = (\mathcal{V}, \mathcal{L}, \mathcal{T}, \mathcal{P}, \mathcal{W})$ is obtained by attaching the *time*, $\mathcal{T}$, to an ordinary network where $\mathcal{T}$ is a set of *time points*, $t \in \mathcal{T}$.

In a temporal network, nodes $v \in \mathcal{V}$ and links $l \in \mathcal{L}$ are not necessarily present or active in all time points. Let $T(v)$, $T \in \mathcal{P}$, be the *activity set* of time points for node v and $T(l)$, $T \in \mathcal{W}$, the activity set of time points for link l.

Besides the presence/absence of nodes and links also their properties can change through time.

Temporal quantities

We introduce a notion of a *temporal quantity*

$$a(t) = \begin{cases} a'(t) & t \in T_a \\ \maltese & t \in \mathcal{T} \setminus T_a \end{cases}$$

where T_a is the *activity time set* of a and $a'(t)$ is the value of a in an instant $t \in T_a$, and ⌘ denotes the value *undefined*.

We assume that the values of temporal quantities belong to a set A which is a *semiring* $(A, +, \cdot, 0, 1)$ for binary operations $+$ and $\cdot$.

We can extend both operations to the set $A_⌘ = A \cup \{⌘\}$ by requiring that for all $a \in A_⌘$ it holds

$$a + ⌘ = ⌘ + a = a \quad \text{and} \quad a \cdot ⌘ = ⌘ \cdot a = ⌘.$$

The structure $(A_⌘, +, \cdot, ⌘, 1)$ is also a semiring.

Operations with temporal quantities

Let $A_{⌘}(\mathcal{T})$ denote the set of all temporal quantities over $A_{⌘}$ in time $\mathcal{T}$. To extend the operations to networks and their matrices we first define the *sum* (parallel links) $a + b$ as

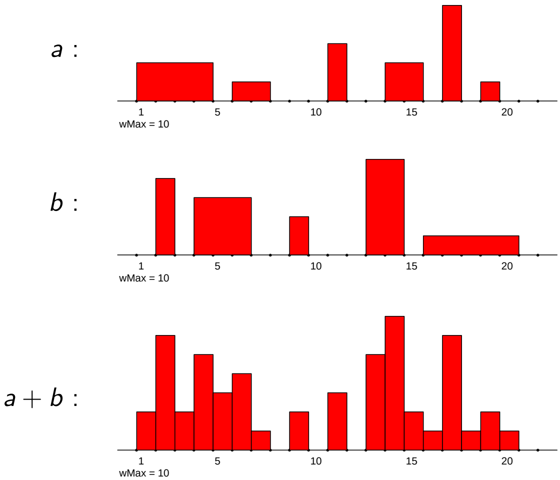
$$(a + b)(t) = a(t) + b(t) \quad \text{and} \quad T_{a+b} = T_a \cup T_b.$$

The *product* (sequential links) $a \cdot b$ is defined as

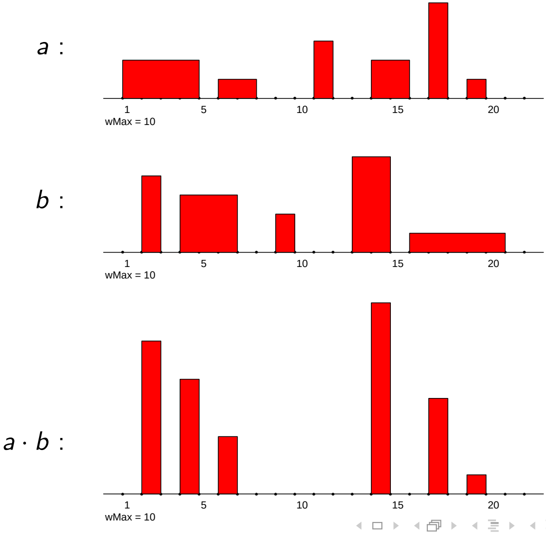
$$(a \cdot b)(t) = a(t) \cdot b(t) \quad \text{and} \quad T_{a \cdot b} = T_a \cap T_b.$$

Let us define the temporal quantities $\mathbf{0}$ and $\mathbf{1}$ with requirements $\mathbf{0}(t) = ⌘$ and $\mathbf{1}(t) = 1$ for all $t \in \mathcal{T}$. Again, the structure $(A_{⌘}(\mathcal{T}), +, \cdot, \mathbf{0}, \mathbf{1})$ is a semiring.

Addition of temporal quantities



Multiplication of temporal quantities



Creating temporal networks

Let the binary matrix $\mathbf{A} = [a_{ep}]$ describe a two-mode (or one-mode) network on the set of events E and the set of of participants P:

$$a_{ep} = \begin{cases} 1 & p \text{ participated in the event } e \\ 0 & \text{otherwise} \end{cases}$$

The function $d : E \to \mathcal{T}$ assigns to each event e the date $d(e)$ when it happened. $\mathcal{T} = [\mathit{first}, \mathit{last}] \subset \mathbb{N}$. Using these data we can construct two temporal affiliation matrices:

- **instantaneous Ai** $= [ai_{ep}]$, where

$$ai_{ep} = \begin{cases} [(d(e), d(e)+1, 1)] & a_{ep} = 1 \\ [\,] & \text{otherwise} \end{cases}$$

- **cumulative Ac** $= [ac_{ep}]$, where

$$ac_{ep} = \begin{cases} [(d(e), \mathit{last}+1, 1)] & a_{ep} = 1 \\ [\,] & \text{otherwise} \end{cases}$$

Temporal networks in Nets

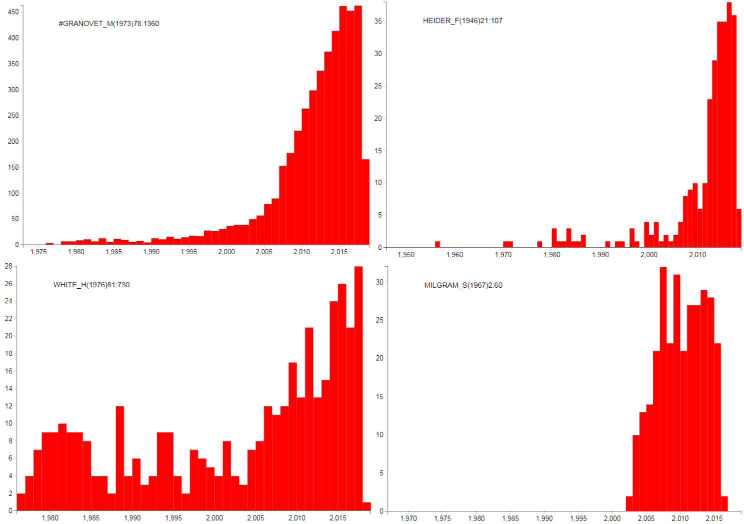
```
>>> net = wdir+"/WAins.json"
>>> WAi = N.loadNetJSON(net)
>>> I = WAi.Index()
>>> I["ERDOS_P(1959)6:290"]
776
>>> WAi._nodes[776]
[{}, {}, {71246: [1091], 89670: [214562]},
{'mode': 1, 'lab': 'ERDOS_P(1959)6:290', 'act': [[1959, 2019, 1]]}]
>>> WAi._nodes[71246]
[{}, {776: [1091], 11539: [25213], 11540: [25214], 15565: [33820],
42898: [129167]}, {},
{'mode': 2, 'lab': 'ERDOS_P', 'act': [[1894, 2019, 1]]}]
>>> WAi._nodes[89670]
[{}, {776: [214562], 15565: [215596]}, {},
{'mode': 2, 'lab': 'RENYI_A', 'act': [[1894, 2019, 1]]}]
>>> WAi._links[1091]
[776, 71246, True, None, {'tq': [[1959, 1960, 1]]}]
>>> WAi._links[214562]
[776, 89670, True, None, {'tq': [[1959, 1960, 1]]}]
```

Temporal citation networks

```
gdir = 'C:/Users/batagelj/work/Python/graph/Nets'
wdir = "C:/Users/batagelj/work/Python/WoS/SocNet/2018/Time/cite"
cdir = "C:/Users/batagelj/work/Python/WoS/SocNet/2018/Time/work/chart"
import sys, os, re, datetime, json
sys.path = [gdir]+sys.path; os.chdir(wdir)
from TQ import *
from Nets import Network as N
Ci = N.loadNetJSON(wdir+"/CiteIns.json")
Cu = N.loadNetJSON(wdir+"/CiteCum.json")
L = [ "#WASSERMA_S(1994):", "#GRANOVET_M(1973)78:1360", "WATTS_D(1998)393:440",
      "BARABASI_A(1999)286:509", "FREEMAN_L(1979)1:215", "#BOYD_D(2007)13:210",
      "MCPHERSO_M(2001)27:415", "BURT_R(1992):" ]
I = Ci.Index()
----
>>> I["#WASSERMA_S(1994):"]
33
>>> TCin = [ (u, Ci.TQnetInDeg(I[u])) for u in L ]
>>> TCuin = [ (u, Cu.TQnetInDeg(I[u])) for u in L ]
>>> TCin[0]
('#WASSERMA_S(1994):', [(1994, 1995, 2), (1995, 1996, 7), (1996, 1997, 12), (1997, 1998, 23),
 (1998, 1999, 26), (1999, 2000, 41), (2000, 2001, 31), (2001, 2002, 54), (2002, 2003, 38),
 (2003, 2004, 64), (2004, 2005, 76), (2005, 2006, 85), (2006, 2007, 103), (2007, 2008, 220),
 (2008, 2009, 238), (2009, 2010, 313), (2010, 2011, 364), (2011, 2012, 424), (2012, 2013, 459),
 (2013, 2014, 438), (2014, 2015, 504), (2015, 2016, 579), (2016, 2017, 552), (2017, 2018, 528),
 (2018, 2019, 167)])
>>> TCuin[0]
('#WASSERMA_S(1994):', [(1994, 1995, 2), (1995, 1996, 9), (1996, 1997, 21), (1997, 1998, 44),
 (1998, 1999, 70), (1999, 2000, 111), (2000, 2001, 142), (2001, 2002, 196), (2002, 2003, 234),
 (2003, 2004, 298), (2004, 2005, 374), (2005, 2006, 459), (2006, 2007, 562), (2007, 2008, 782),
 (2008, 2009, 1020), (2009, 2010, 1333), (2010, 2011, 1697), (2011, 2012, 2121),
 (2012, 2013, 2580), (2013, 2014, 3018), (2014, 2015, 3522), (2015, 2016, 4101),
 (2016, 2017, 4653), (2017, 2018, 5181), (2018, 2019, 5348)])
>>> w = 800; h = 500
>>> p=0; (Tmin,Tmax,tt,TQmax) = TQ.TqSummary(TCin[p][1]); tit = TCin[p][0]
>>> N.TQshow(TCin[p][1],cdir,TQmax,Tmin,Tmax,w,h,tit,fill='red')
```

Temporal citation network

temporal indegrees



Authors Collaboration

Network Co

$$\mathbf{Co} = \mathbf{WA}^T * \mathbf{WA}$$

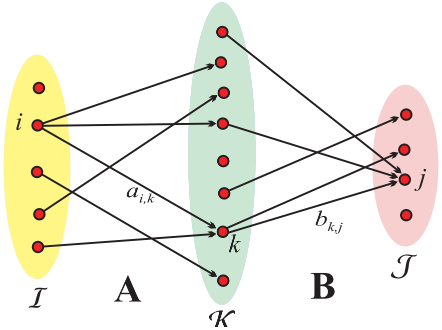
The weight of the edges between the nodes *i* and *j* is equal to total number of works author i and j wrote together.

The values of loops in **Co** are equal to the total number of works that each author have (which is also equal to the indegree values of the **WA** network).

The proposed approach has some *limitations*, such as the overrating of the contribution of works with many authors. To make the contribution of each work equal we have to use the normalized version $n(\mathbf{WA})$ of the network (matrix) (*fractional approach*)

$$n(\mathbf{WA})[w, a] = \frac{\mathbf{WA}[w, a]}{\max(1, \mathrm{outdeg}[w])}$$

Multiplication of networks



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

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 $\mathcal{K}$.

Derived networks

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

The weight of the arc from the node a to the node k is equal to the number of works in which the author a used the keyword k.

$$\mathbf{CiteJ} = (\mathbf{WJ})^T * \mathbf{Cite} * \mathbf{WJ}$$

the value of weight of the element [u,v] is equal to the **number of citations** from journal i to journal j.

$$\mathbf{CiteAn} = (\mathbf{WA})^T * n(\mathbf{Cite}) * \mathbf{WA}$$

The value of element CiteAn[u,v] is equal to the number of **fractional contribution** of citations from works coauthored by u to works coauthored by v.

etc.

Multiplication of co-occurence networks

Instantaneous

Instantaneous **A** on $P \times A$ and **B** on $P \times B$. $\mathbf{C} = \mathbf{A}^T.\mathbf{B}$ on $A \times B$.

$$c_{ij}(t) = \sum_{p \in P} a_{pi}(t)^T \cdot b_{pj}(t)$$

$a_{pi} = [(d_{pi}, d_{pi}+1, v_{pi})]$ and $b_{pj} = [(d_{pj}, d_{pj}+1, v_{pj})]$
for $t = d$ we get

$$c_{ij} = [(d, d+1, \sum_{p \in P : d_{pi} = d_{pj} = d} v_{pi}.v_{pj})]_{d \in \mathcal{T}}$$

for $v_{pi} = v_{pj} = 1$ we finally get

$$v_{ij}(d) = |\{p \in P : d_{pi} = d_{pj} = d\}|$$

For binary temporal two-mode networks **A** and **B** the value $v_{ij}(d)$ of the product $\mathbf{A}^T.\mathbf{B}$ is equal to the number of different members of P with which both i and j have contact in the instant d.

Cumulative **A** on $P \times A$ and **B** on $P \times B$. $\mathbf{C} = \mathbf{A}^T.\mathbf{B}$ on $A \times B$.

$$c_{ij}(t) = \sum_{p \in P} a_{pi}(t)^T \cdot b_{pj}(t)$$

$a_{pi} = [(d_{pi}, last + 1, v_{pi})]$ and $b_{pj} = [(d_{pj}, last + 1, v_{pj})]$
for $t = d$ we get

$$c_{ij} = [(d, d + 1, \sum_{p \in P : (d_{pi} \leq d) \wedge (d_{pj} \leq d)} v_{pi} . v_{pj})]_{d \in \mathcal{T}}$$

for $v_{pi} = v_{pj} = 1$ we finally get

$$v_{ij}(d) = |\{p \in P : (d_{pi} \leq d) \wedge (d_{pj} \leq d)\}|$$

For binary temporal two-mode networks **A** and **B** the value $v_{ij}(d)$ of the product $\mathbf{A}^T.\mathbf{B}$ is equal to the number of different members of P with which both i and j have contact in all instants up to including the instant d.

Citations among journals

JCJ and JCJn from WJins, WJcum, CiteIns nets

$$\mathbf{JCJ} = (\mathbf{WJins})^T * \mathbf{CiteIns} * \mathbf{WJcum}$$

The value of weight of the element $JCJ[i, j]$ is equal to the **number of citations** per year from journal i to journal j.

$$\mathbf{JCJn} = (\mathbf{WJins})^T * n(\mathbf{CiteIns}) * \mathbf{WJcum}$$

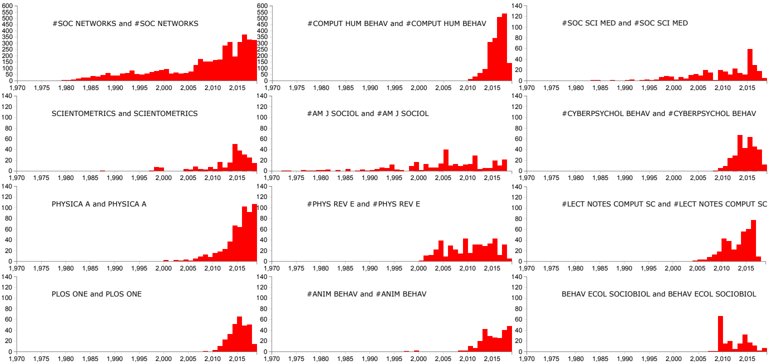
where

$$n(\mathbf{CiteIns})[u, v] = \frac{\mathbf{CiteIns}[u, v]}{\max(1, \mathrm{outdeg}(u))}$$

The value of element $JCJn[i, j]$ is equal to the number of **fractional contribution** of citations per year from journal i to journal j.

Self-citations of journals

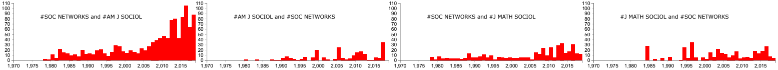
Loops from JCJ network



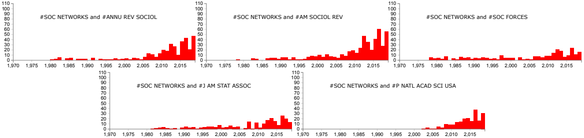
Citations of Social Networks journal

InSum and OutSum of JCJ network without loops

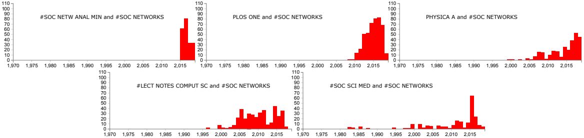
OutSum and InSum:



OutSum:

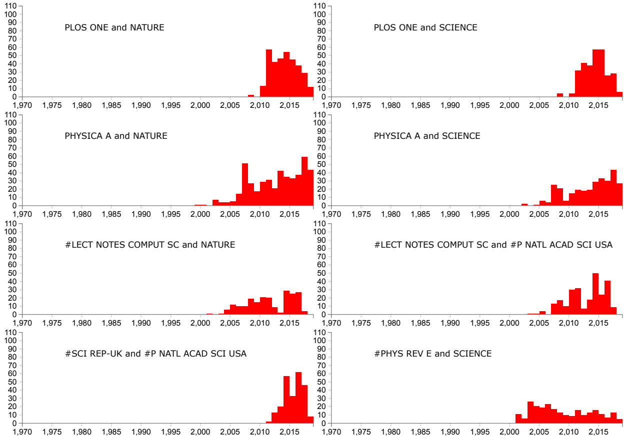


InSum:



Citation of general scientific journals

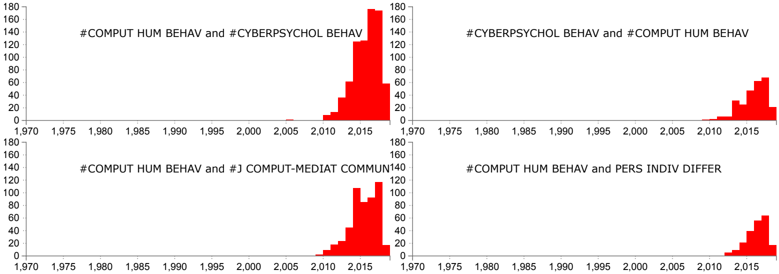
from JCJ network without loops



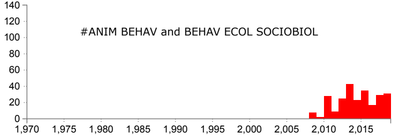
Citations of other journals

from JCJ network without loops

Computer Science:

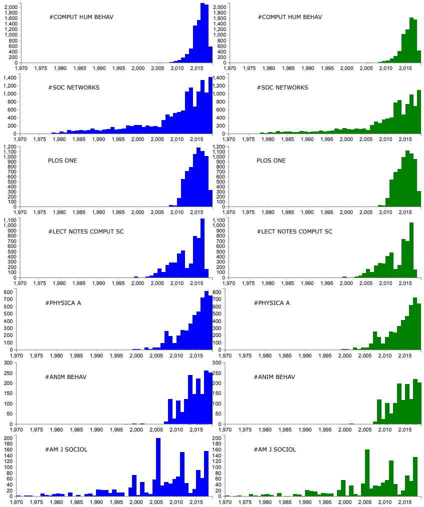


Animal social networks:



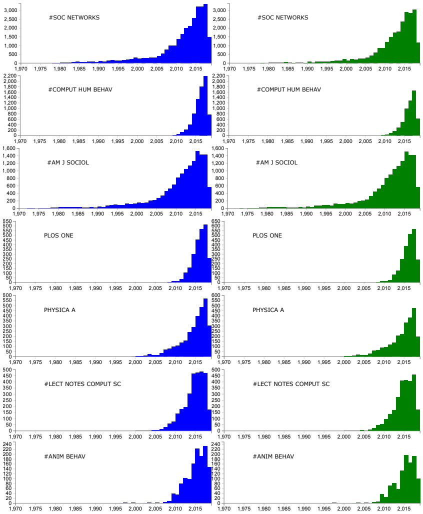
Outgoing citations with/without loops

from JCJn network



Incoming citations with/without loops

from JCJn network



Bibliography I

1. Batagelj, V. (2007) WoS2Pajek. Networks from Web of Science. Version 0.3. Manual. URL: http://vlado.fmf.uni-lj.si/pub/networks/pajek/WoS2Pajek/WoS2Pajek.pdf
2. Batagelj V., Cerinšek M.(2013). On bibliographic networks. Scientometrics. 96 (3), 845-864
3. Batagelj, V., Doreian P., V., Ferligoj, A., Kejzar N. Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution, 2014.
4. Batagelj, V., Ferligoj, A. Squazzoni, F. The emergence of a field: a network analysis of research on peer review. Scientometrics (2017) 113: 503. https://doi.org/10.1007/s11192-017-2522-8
5. Batagelj, V, Cerinšek, M (2013). On bibliographic networks. *Scientometrics* 96(3), 845-864.

Bibliography II

6. Batagelj, V, Doreian, P, Ferligoj, A, Kejžar, N (2014). *Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution*. Wiley.
7. Batagelj, V, Praprotnik, S (2016): An algebraic approach to temporal network analysis based on temporal quantities. *Social Network Analysis and Mining*, 6(1), 1-22.
8. Batagelj, V (2017): **Nets** – a Python package for network analysis. `https://github.com/bavla/Nets`
9. Batagelj, V (2018): Python Packages for Networks. In: Alhajj R., Rokne J. (eds) *Encyclopedia of Social Network Analysis and Mining*. Springer, New York, NY