Network analysis approach to the analysis of event sequences

Vladimir Batagelj, Anuška Ferligoj, Oštro Center
IMFM, UP IAM, FDV, Oštro

IFCS, San Jose, Costa Rica, 2024

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Anuška Ferligoj: anuska.ferligoj@fdv.uni-lj.si
Current version of slides (July 15, 2024 at 10:17): slides PDF
https://github.com/bavla/TQ/tree/master/trajectories/Feb24

Oštro, Zvezoskop, data

Oštro is the Center for Investigative Journalism in the Adriatic region and cultivates investigative and data journalism.Colleagues from

Oštro have developed the *Zvezoskop* project, which allows the public to independently research the connections between current ministers, state secretaries, and members of parliament, and their educational and career paths. Based on their CVs, the careers of 160 current politicians are included in Zvezoskop.

Raw data


```
,person_id,person_name,part_of_cv,show_in_network,start_day,start_month,
start_year,institution_si,institution_standardized_si,institution_department_si,
affiliation_type_si,position_si,end_day,end_month,end_year,notes_institution_si,notes_position_s
...
2241,112,Matej Tonin,izobraževanje,True,,9,1998,gimnazija in srednja šola rudolfa maistra kamnik
2242,112,Matej Tonin,izobraževanje,True,,10,2002,ul fakulteta za družbene vede,,,,dodiplomski št
2243,112,Matej Tonin,izobraževanje,True,,10,2007,ul fakulteta za družbene vede,,,,magistrski štu
2244,112,Matej Tonin,delovne izkušnje,True,16/4/2007,,2007,državni zbor,,poslanska skupina NSi,j
2245,112,Matej Tonin,strankarska pozicija,True,,,2006,nsi,,občinski odbor Kamnik,,predsednik,,,2
2246,112,Matej Tonin,strankarska pozicija,True,10/9/2010,,2010,nsi,,,,kanididat za župana v obči
2247,112,Matej Tonin,delovne izkušnje,True,29/11/2006,,2006,občina kamnik,,,politik,občinski sve
2248,112,Matej Tonin,strankarska pozicija,True,,,2007,nsi,,Mlada Slovenija,,podpredsednik,,,2009
2249,112,Matej Tonin,delovne izkušnje,True,3/11/2008,,2008,vizija agencija za komuniciranje mate
2250,112,Matej Tonin,strankarska pozicija,True,,,2010,nsi,,,,podpredsednik,,,2012,,
2251,112,Matej Tonin,delovne izkušnje,False,21/12/2011,,2011,državni zbor,,,politik,poslanec,13/
2253,112,Matej Tonin,delovne izkušnje,True,22/6/2018,,2018,državni zbor,,,politik,predsednik,23/
2254,112,Matej Tonin,delovne izkušnje,False,13/5/2022,,2022,državni zbor,,,politik,poslanec,,,21
2255,112,Matej Tonin,strankarska pozicija,True,22/8/2008,,2008,nsi,,,,kandidat za poslanca,21/9/
2256,112,Matej Tonin,delovne izkušnje,True,13/3/2020,,2020,ministrstvo za obrambo,,,politik,mini
2257,112,Matej Tonin,delovne izkušnje,True,13/3/2020,,2020,vlada,,,politik,podpredsednik,1/6/202
2258,112,Matej Tonin,strankarska pozicija,True,1/2/2018,,2018,nsi,,,,v d predsednika ,20/4/2018,
2259,112,Matej Tonin,strankarska pozicija,False,21/4/2018,,2018,nsi,,,,predsednik,,,2100,,
2260,112,Matej Tonin,svetovalne in nadzorne funkcije etc.,True,,,2015,inštitut dr janeza evangel
2261,112,Matej Tonin,prostočasne aktivnosti,True,27/11/2006,,2006,športno društvo tuhinj,,,,pred
2262,112,Matej Tonin,strankarska pozicija,False,,12,2001,nsi,,,,član,,,2100,,
2263,112,Matej Tonin,strankarska pozicija,False,1/2/2018,,2018,nsi,,izvršilni odbor,,predsednik,
...
```

Reduced data

A person X's CV, $CV_X = (e_1, e_2, \ldots, e_{k_x})$, consists of a sequence of events e_i. For an event $e_i = (s_i, f_i, R_i, S_i, \ldots)$ we at least know its start date s_i, its end (finish) date f_i, the type R_i of the event, the state (location) S_i of the event, and maybe something more.

The cleaned data table was first reduced to variables considered in the analysis.

- R – row (event) number (index)
- ID – `person_name`: person's name
- s – `start_day|month|year`: start date
- f – `end_day|month|year`: finish date
- S – `part_of_cv`: type of activity R_i (work, education, free time, political party, counseling and supervision)
- T – `institution_si` S_i : location (institution, party, organization, etc.)

Co-presence network

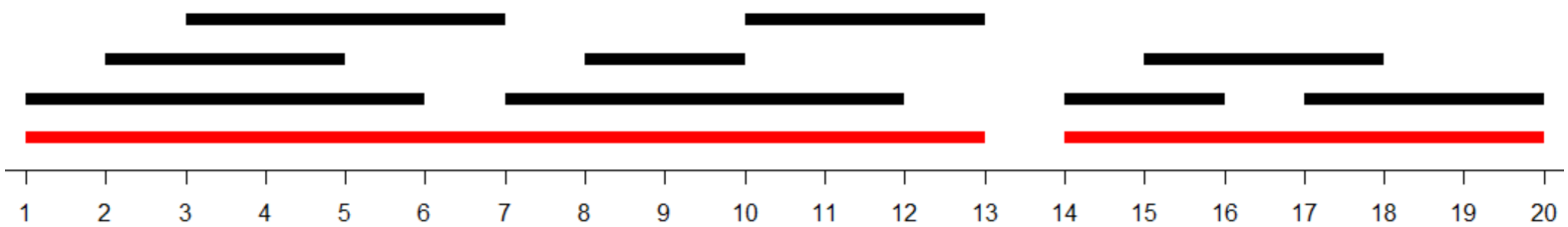
We decided to base our analysis on the corresponding *co-presence network* $N = (V, L, w, t)$ – a weighted (number of days of co-presence on a link) multi-relational (5 areas) temporal network. The set of nodes V consists of 160 politicians.

Co-presence network

There is a *link* (edge) $\ell = (u : v; R)$ of type (relation) R between persons $u, v \in V$ iff there exist events $e_u = (s_u, f_u, R_u, S_u)$ and $e_v = (s_v, f_v, R_v, S_v)$ such that $R_u = R_v = R$ and $S_u = S_v$ and $[s, f] = [s_u, f_u] \cap [s_v, f_v] \neq \emptyset$. The weight of the link $\ell \in L$ is $w(\ell) = f - s$, that is the length of the corresponding time interval $[s, f]$. The traditional sequence analysis deals mainly with the analysis of states (3). The co-presence network enables us to analyze groups of persons using network analysis methods such as cuts, cores, islands, network clustering and blockmodeling (1). The proposed approaches will be illustrated on the dataset Zvezoskop of Slovenian politicians (2).

Co-presence network

In preliminary experiments, we found that parallel events of the same kind can greatly increase the similarity between persons. One way to reduce this influence is to count parallel events in the same institution between two persons only once.



All time intervals $\{[s_i, f_i] : i \in 1 : k\}$ of edges between nodes u and v of kind S we replace with a single edge with weight

$$w(u, v; S) = |\bigcup_{i=1}^{k} [s_i, f_i]|$$

Data – co-presence


```
   R           ID          s          f         S                    T
  64   Igor Papič 1973-09-01 1981-08-01 education          oš v kočevju
1116 Robert Golob 1979-09-01 1983-08-01 education       gim nova gorica
  65   Igor Papič 1981-09-01 1985-06-01 education      vegova ljubljana
1068 Robert Golob 1983-10-01 1989-09-29 education                 ul fe
  66   Igor Papič 1986-10-01 1992-01-01 education                 ul fe
1070 Robert Golob 1989-11-15 1994-11-14      work                 ul fe
1069 Robert Golob 1992-10-01 1994-10-17 education                 ul fe
  67   Igor Papič 1992-10-01 1995-01-01 education                 ul fe
1115 Robert Golob 1994-01-01 1994-01-01 education tu georgia atlanta us
  70   Igor Papič 1994-01-01 1996-01-01      work             siemens de
1071 Robert Golob 1994-11-15 1995-10-31      work                 ul fe
  69   Igor Papič 1993-03-01 1996-01-01      work                 ul fe
  68   Igor Papič 1995-10-01 1998-01-01 education                 ul fe
1072 Robert Golob 1995-11-01 1997-09-30      work                 ul fe
1073 Robert Golob 1997-10-01 2001-06-30      work                 ul fe
1078 Robert Golob 1998-11-01 1999-08-31      work   min za gospodarstvo
  71   Igor Papič 1999-01-01 2007-01-01      work                 ul fe
  96   Igor Papič 2000-01-01 2002-06-30      work                 ul fe
  94   Igor Papič 2001-01-01 2002-12-31      work                 ul fe
...
                u            v          s          f       rel    d   Ru   Rv
1   Robert Golob   Igor Papič 1986-10-01 1989-09-29 education 1094 1068   66
2   Robert Golob   Igor Papič 1993-03-01 1994-11-14      work  623 1070   69
3   Robert Golob   Igor Papič 1992-10-01 1994-10-17 education  746 1069   67
4     Igor Papič Robert Golob 1994-11-15 1995-10-31      work  350   69 1071
5     Igor Papič Robert Golob 1995-11-01 1996-01-01      work   61   69 1072
7   Robert Golob   Igor Papič 1999-01-01 2001-06-30      work  911 1073   71
10  Robert Golob   Igor Papič 2000-01-01 2001-06-30      work  546 1073   96
12  Robert Golob   Igor Papič 2001-01-01 2001-06-30      work  180 1073   94
...
          u            v          s          f    d       rel   test
1 Igor Papič Robert Golob 1993-03-01 1996-01-01 1037      work  ul fe
2 Igor Papič Robert Golob 1999-01-01 2021-01-01 8037      work  ul fe
3 Igor Papič Robert Golob 1992-10-01 1994-10-17  747 education  ul fe
```

Data

Co-presence network was constructed for the time before the elections on April 24, 2022.

	relation	apr22
1	work	1767
2	education	470
3	leisure activities	6
4	party	1086
5	counseling & supervision	10
	total	3339
	nodes	154

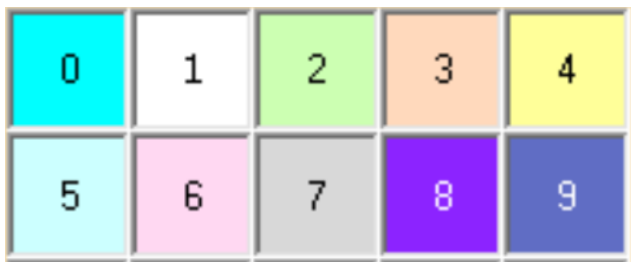
The link values in the network are in the interval $[1, 3339]$. Applying the square root we transform them into the interval $[1, 58]$. Another application of the square root gives the interval $[1, 8]$, which is used for the visualization.

The analysis was done using program Pajek.

Partitions

In the data on persons, there are three partitions that describe the characteristics of an individual person: party (party membership), position (role in the country), and gender.

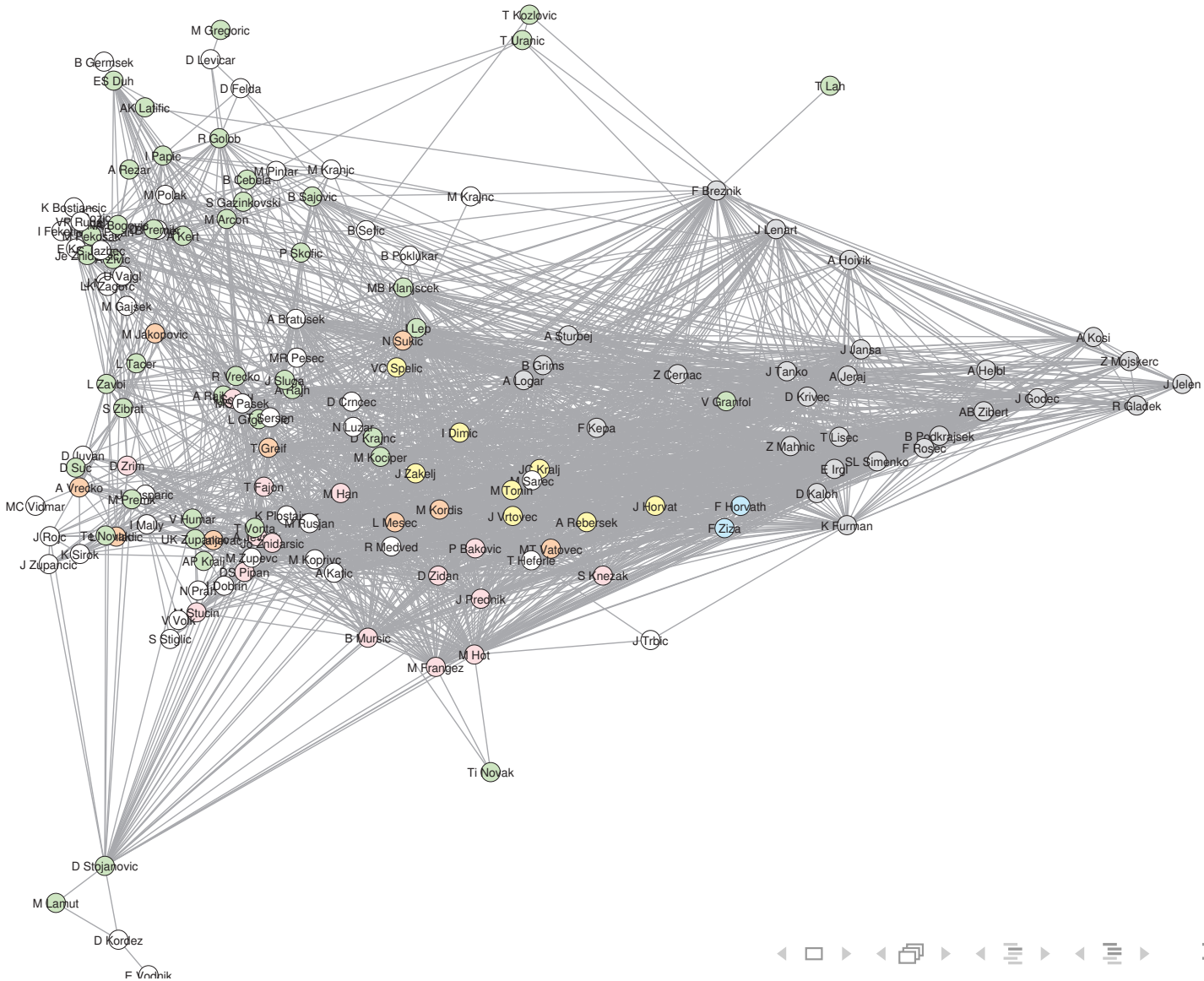
The party membership is displayed as the node color. The following bright colors denote political parties:



1: Undefined; 2: Freedom (Svoboda) Movement; 3: Left; 4: NSi; 5: IT+HU; 6: SD; 7: SDS.

Their roles are presented by the following symbols: ♟ – member of parliament, ♞ – secretary, ♜ – minister, ♛ – prime minister; and their gendre by colors: blue – male, red – female.

Co-presence network



Skeletons

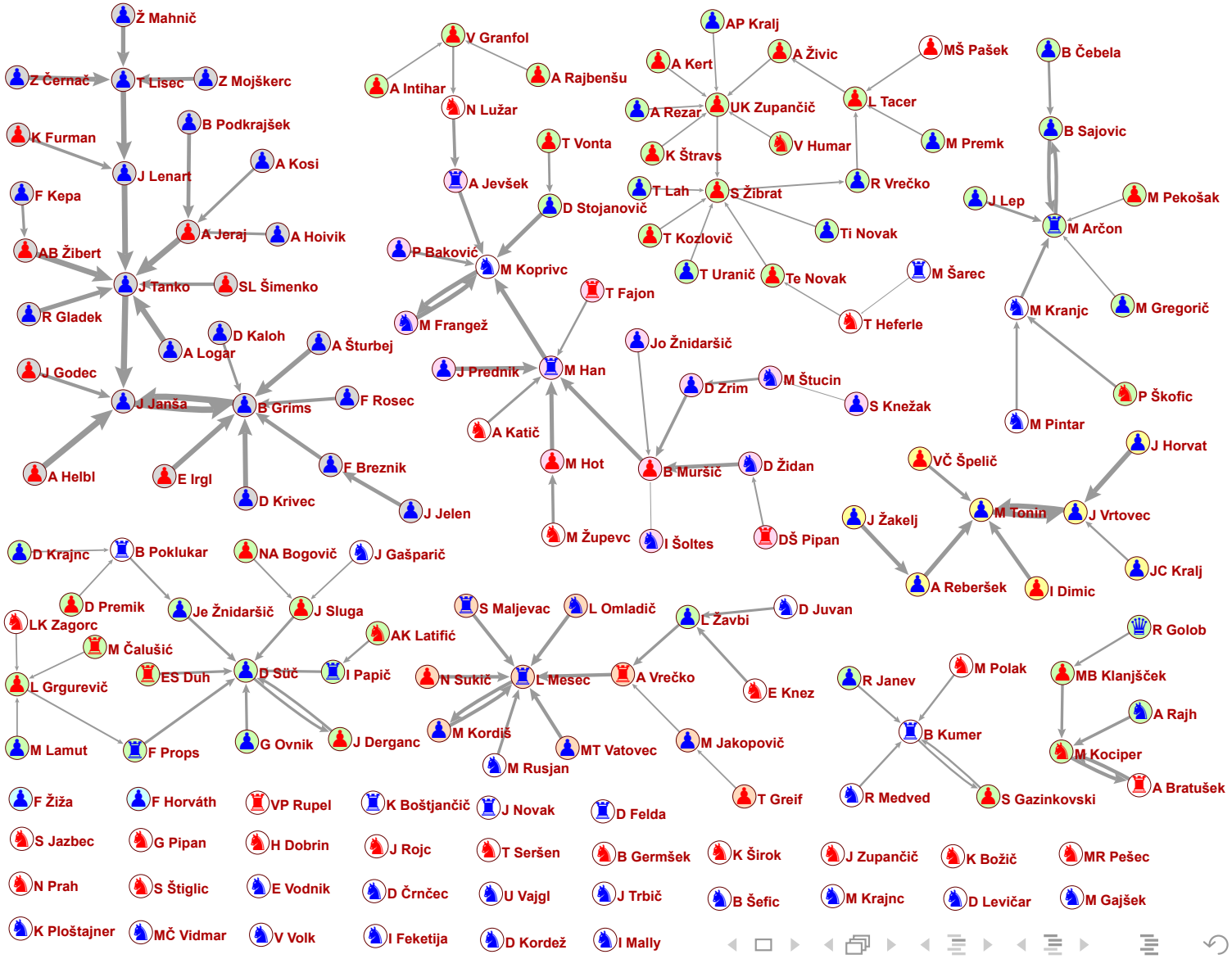
For a quick inspection of the structure of weighted networks, we use various skeletons, in which less important elements (nodes, links) are discarded. An example of a skeleton is a link-cut or a node-cut at selected level. Another example is a *k-neighbors* network, in which we keep only the k most important neighbors (highest weights) in each node.

Most often we use 1-neighbors because of their tree like structure.

1-neighbors for party links February 2024

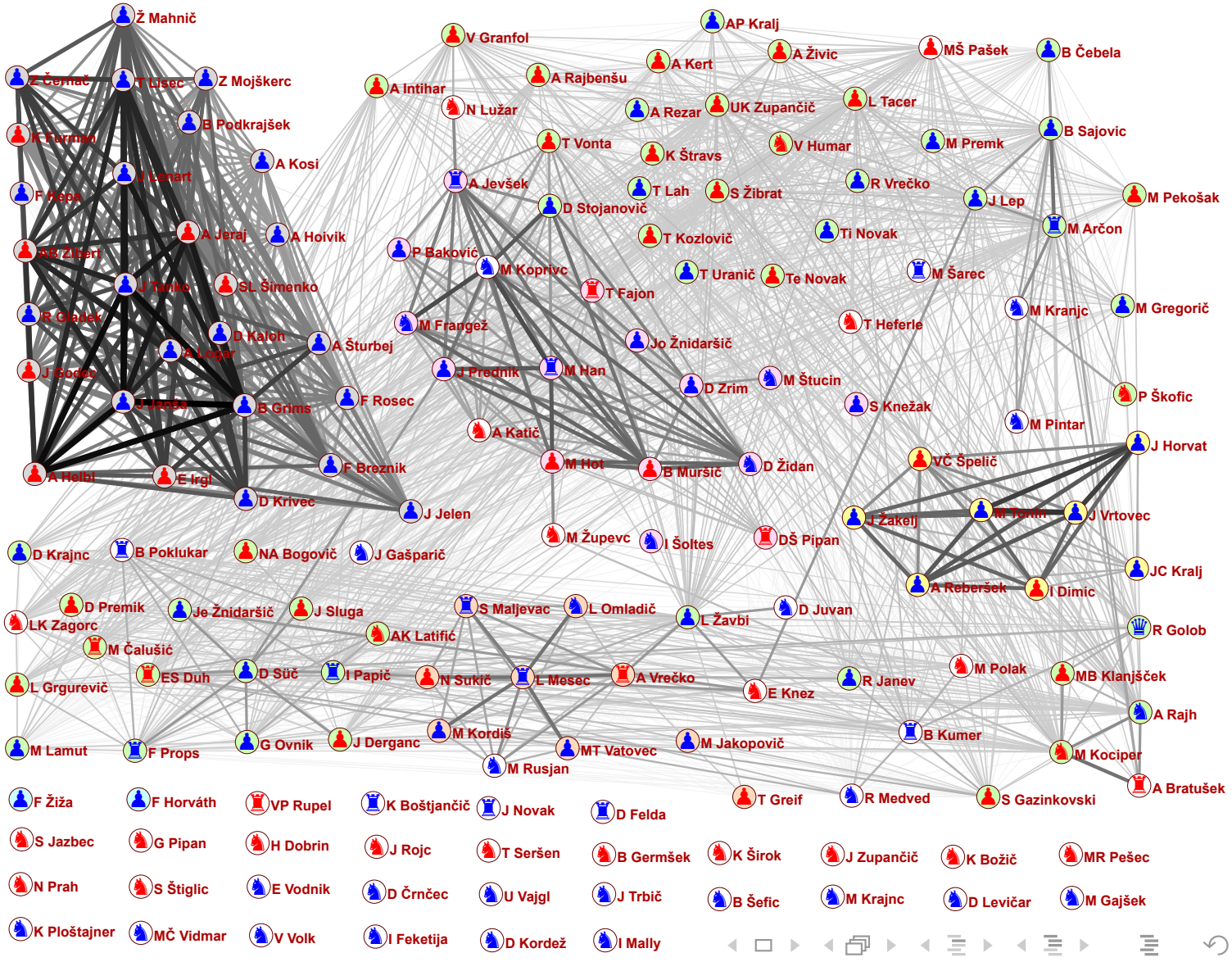
Whole network for party links February 2024 on the 1-neighbors nodes layout

1-neighbors / Parties February 2024 – relation Party



Oštro

Parties February 2024 – relation Party



Blockmodeling

The co-presence network is a weighted one. Therefore, we decided to use indirect blockmodeling approach, where

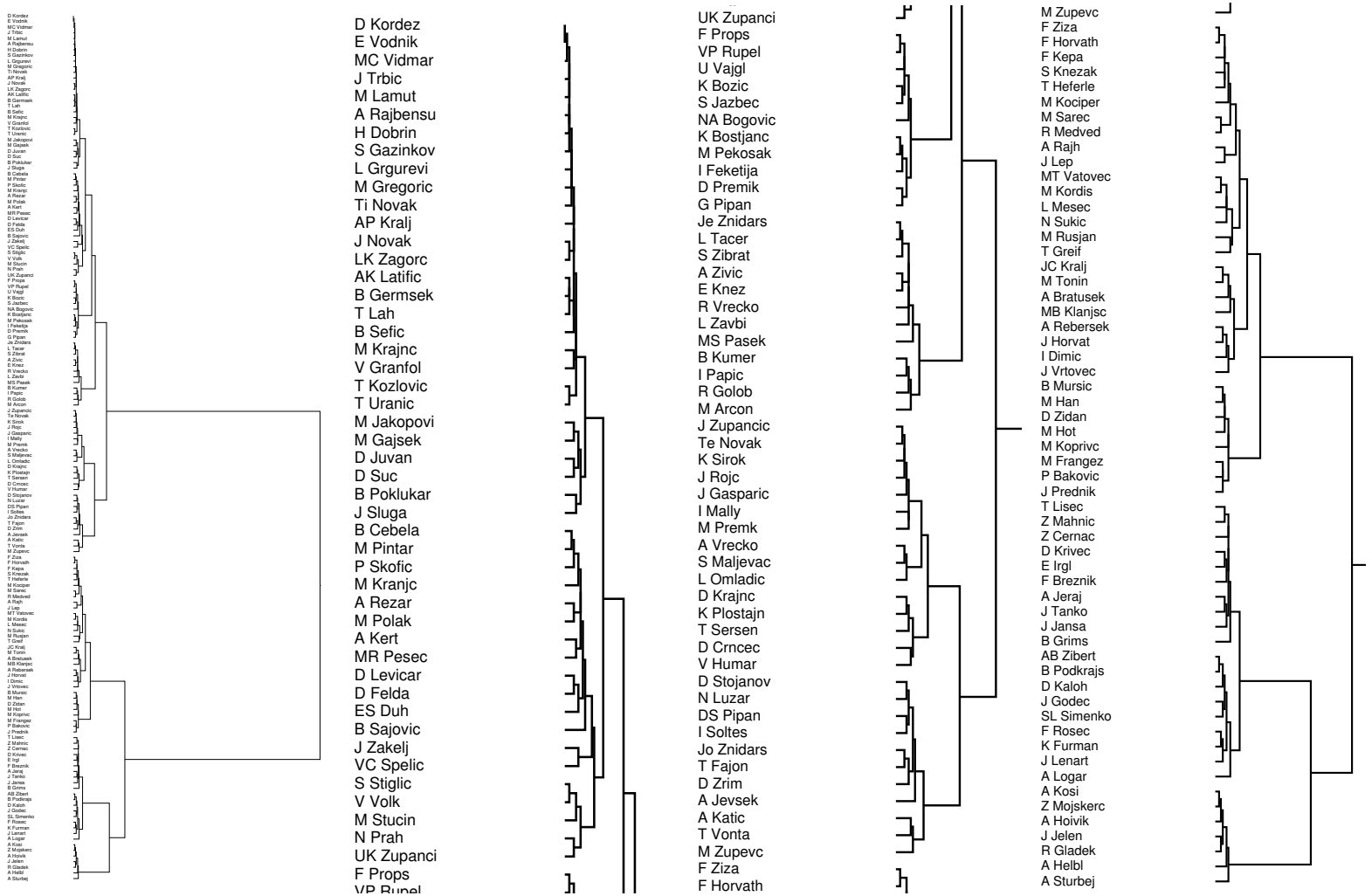
- dissimilarity between nodes was calculated by corrected eucliden distance

$$D[u,v] = \sqrt{(w[u,v]-w[v,u])^2 + (w[u,u]-w[v,v])^2 + \sum_{t:t\neq u, t\neq v} (w[u,t]-w[v,t])^2}$$

- Ward hierarchical agglomerative method was used as a clustering procedure

The obtained dendrogram is presented in the next slide.

Dendrogram

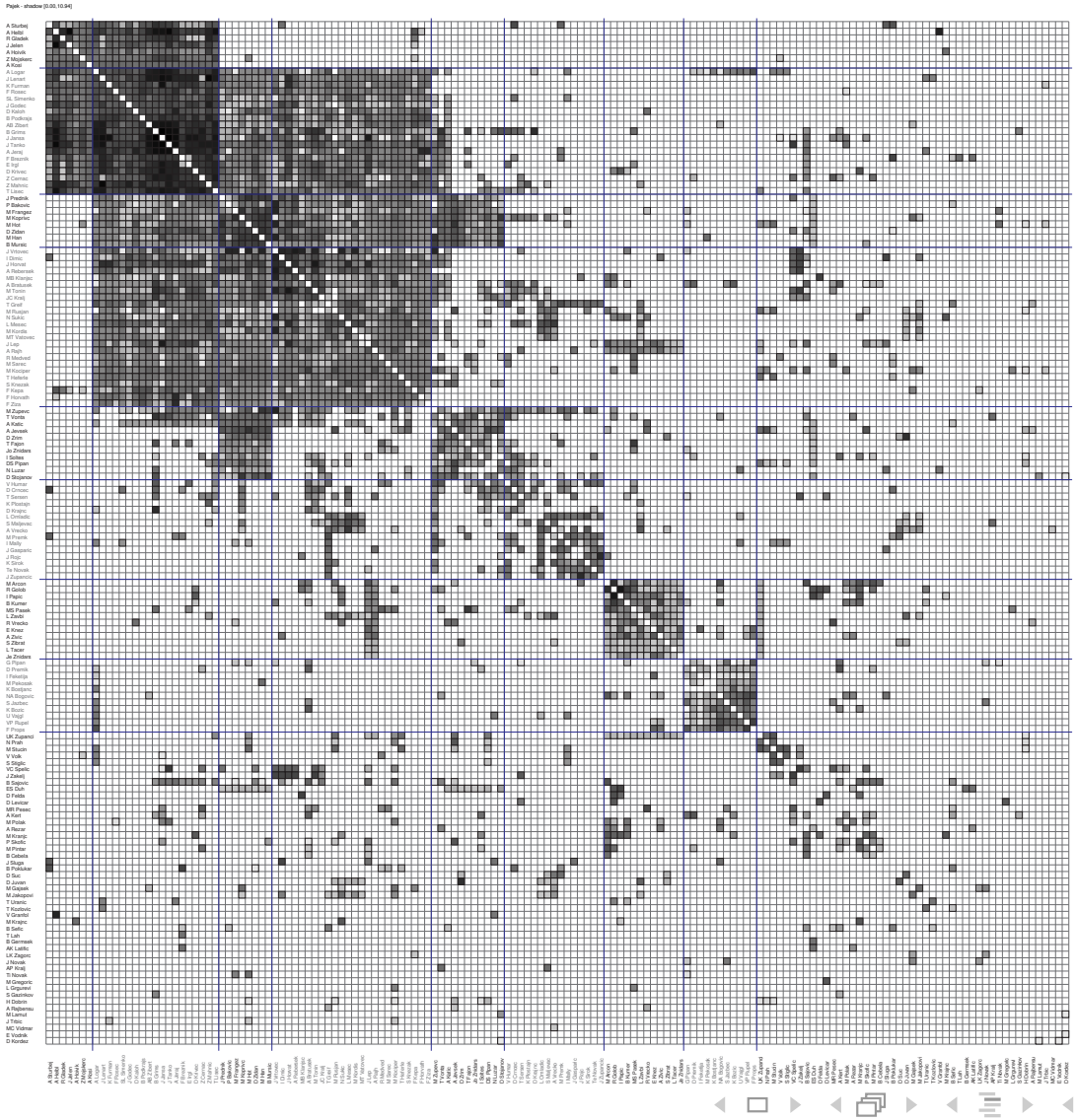


According to the dendrogram 9 clusters of politicians were determined.

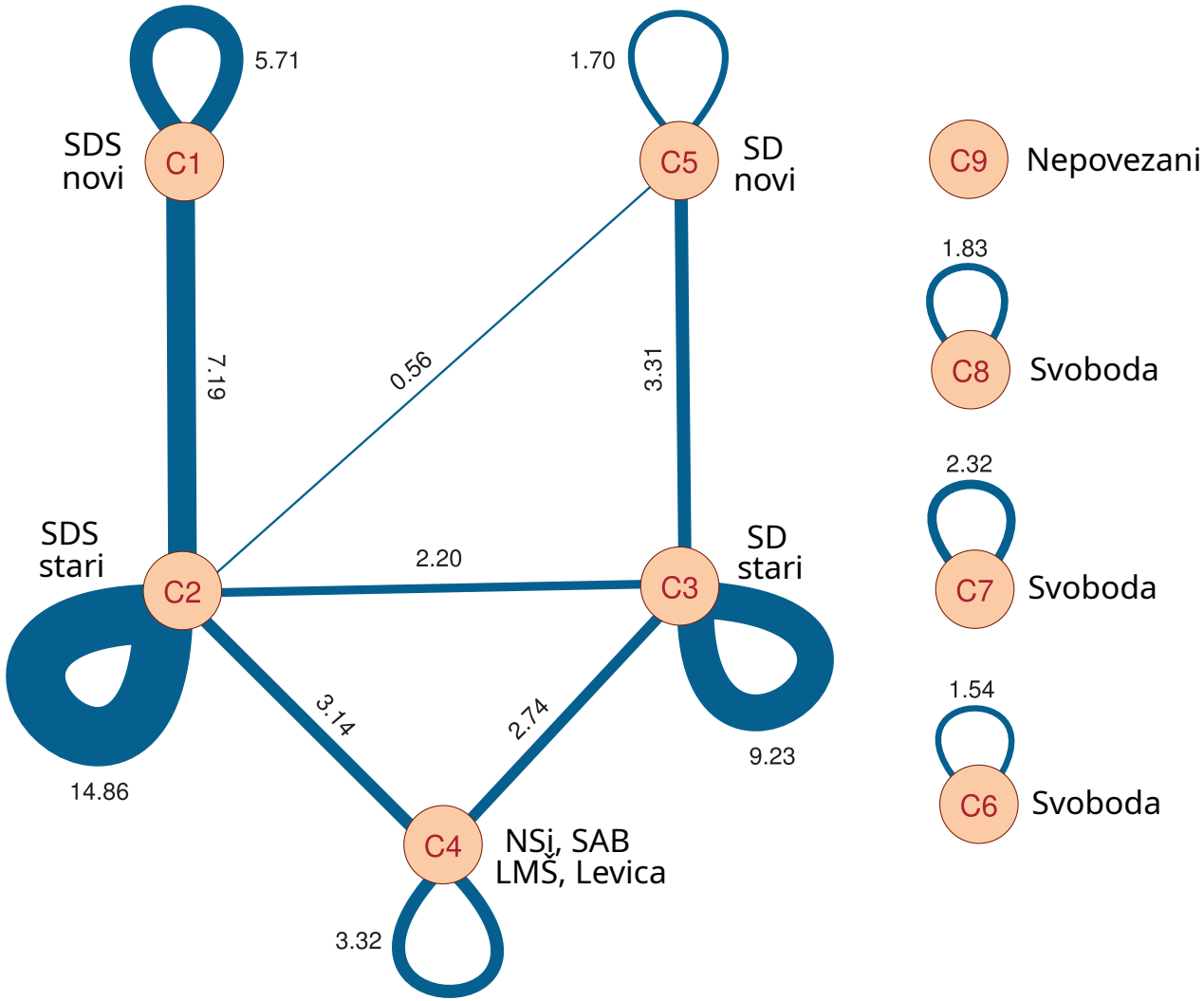
Partition

1. **'New' SDS MPs** (7)
2. **'Old' SDS MPs** (19)
3. **'Old' SD MPs** (8)
4. **MPs not members of SDS or SD** (24)
5. **'New' SD MPs** (11)
6. **Members of Gibanje Svoboda (Freedom) and Levica (Left)** (15)
7. **Members of Gibanje Svoboda (Freedom)** (12)
8. **Members of Gibanje Svoboda (Freedom)** (11)
9. **Isolated political office holders** (47)

Reordered network matrix / $\sqrt{w}$



Blockmodel (weights in years)



Some observations

- The strongest co-presence is among members of the SDS party (among older members, almost 15 years on average). The members of the second oldest SD party (almost 10 years among the seniors) are also quite connected. The greatest co-presence is expected between the two SDS clusters (C1 and C2) (over seven years on average) and between the SD clusters (C3 and C5) with an average of just over three years.
- The fourth group (C4) consists of politicians who have been active for a long time, but are not members of SDS or SD. They are expected to be co-present together with older SDS members and older SD members (on average around three years).
- In the connected part of the blockmodel, politicians have been active for a longer time, including members of NSi and former SAB and LMŠ, but without members of Svoboda.
- Connections within and between the first five groups are mainly due to co-presence in the previous National Assemblies and the state administration.

. . . Some observations

- The members of Svoboda mainly consist of two clusters (C7, C8) and, together with Levica, the cluster C6. These clusters are not co-present with each other, their members are co-present only within their own cluster.
- As expected, politicians within the three Svoboda clusters were co-present before the last parliamentary elections, each of them also includes ministers and their secretaries (e.g. the seventh most connected Svoboda cluster includes the Prime Minister and the persons who were co-present with him in the municipality of Nova Gorica, at the Faculty of Electrical Engineering, and at the company GENi; in the cluster C8 there are members of Svoboda who, on average, have collaborated for a little less than 2 years and work at the ministries of finance and public administration).
- The quite large cluster C9, which has negligible mutual co-presence, consists of politicians from different parties.
- The co-presence in education is interesting (Faculty of Electrical Engineering, Faculty of Social Sciences, Faculty of Economics, and Law) and play a strong role.

Conclusions

1. Traditional sequence analysis concentrate attention mainly to the states. Our network approach deals primary with units.
2. Apply traditional sequence analysis to Zvezoskop data.
   Extend the dataset at least to public officials from the previous government.
   Kinship!? Additional personal attributes. Homophily $uE_a v \equiv a(u) = a(v)$.
3. Find other interesting and well documented datasets.

References I

[1] Batagelj, V., Doreian, P., Ferligoj, A. and Kejžar, N.: Understanding large temporal networks and spatial networks. (Vol. 2). John Wiley & Sons (2014)

[2] Oštro center: Zvezoskop. https://www.zvezoskop.si/en/ (accessed March 31, 2024)

[3] Ritschard, G., Studer, M.: Sequence analysis and related approaches: Innovative methods and applications. Springer Nature (2018)