"Lendület" Research Center for Educational and Network Studies (RECENS)

Text network analysis and visualization of Hungarian, communist-era political reports

Attila Gulyás, Martina K. Szabó, István Boros Jr., Gergő Havadi

Text2Story Workshop

Grenoble, France

2018.03.26.

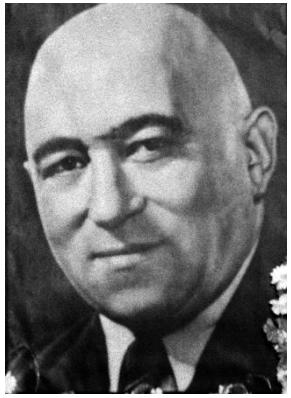
European Research Council
Established by the European Commission
Supporting top researchers from anywhere in the world

About the research project

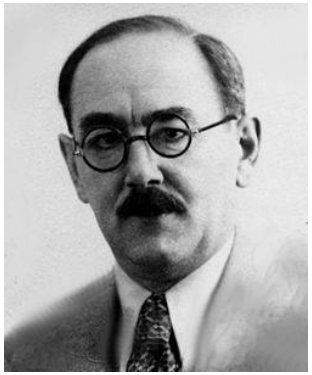
- **Main goals**
  - The Rákosi-era’s (1948-1956) power network
  - Via social network analysis (SNA): The latent and manifested hierarchical historic network and its dynamics are being investigated by the very basis of proper names co-mentions
  - Text network analysis (TNA)
  - Sentiment analysis on historical corpora

Historical background II.

- Hungary was on the losing side of WWII and had became under soviet occupation
- The soviet-copied state was led by Mátyás Rákosi
- A total dictatoriate was built, with planned economy
- The members of the previous elites of aristocracy were forced to leave their goods or face imprisonment
- The continuous upkeeping of combat readiness
- The fake state of workers
- Between 1948 and 1953 the Rákosi system was in its full strength, between 1953 and 1956 was still continuously working with some modifications
- In 1955-56 an easing had begun led by Imre Nagy the new PM of Hungary



Mátyás Rákosi (1892-1971)

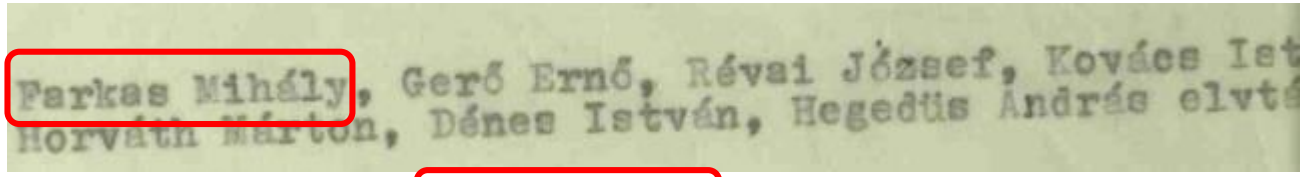


Imre Nagy (1896-1958)

The basic features of the corpus

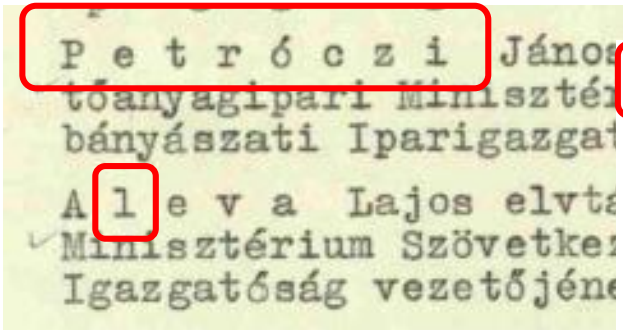
- **The features of the corpus**
  - Typewritten, edited, oftentimes handwritten commented and corrected reports
  - Digitalized documents (JPG), N = approximately 800
  - The repost were converted to .txt formats via Optical Character Recognition (OCR)
  - Heavily damaged documents
- **Repair processes on the corpus**
  - The correction of the typical character errors in the corpus via manual and machine solutions
  - The correction of proper names via manual and machine solutions
  - The identification of the proper names via manual solution

Typical character errors

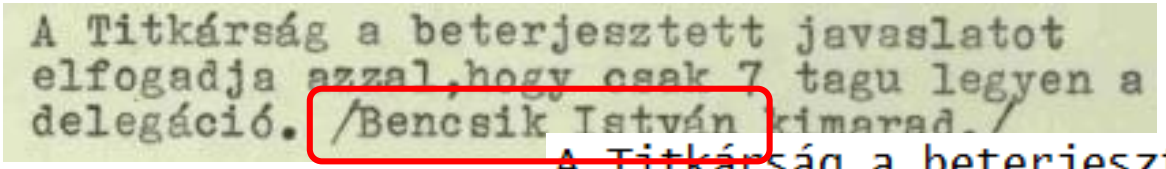

Parkas Mihály, Gerő Ernő, Révai József, Kovács Ist
Horváth Márton, Dénes István, Hegedüs András elvt

Parkae Yihály, Gerti Ernő, 2évai ficizeef, Yovele Ietván,

Horváth Várton, Dénes Ietvvn, Hegedtte fldrie elvtársak,


P e t r ó c z i Jáno
bányászati Iparigazgat
A l e v a Lajos elvta
Minisztérium Szövetkez
Igazgatóság vezetőjéne

P e t r á c z i János elvtársat az Épi-
tőanyagipari Minisztérium Kő- és Kavics-
bányászati Iparigazgatósága vezetőjének,
A l e v a Lajos elvtársat a Helyi Ipari
Minisztérium Szövetkezeti és Kisipari
Igazgatóság vezetőjének,


A Titkárság a beterjesztett javaslatot
elfogadja azzal,hogy csak 7 tagu legyen a
delegáció. /Bencsik István kimarad./

A Titkárság a beterjesztett javaslatot
elfogadja azzal,hogy csak 7 tagu legyen a
delegáció. /Bencsik István kimarad./

The test corpus

- **The feature of the test corpus**
  - 6 pieces of randomly chosen Secretariat reports
  - All of the documents were fully corrected from character errors
  - The structure of the reports was carefully kept
  - Now the reports are in tidy-text format
- **The aims of our test corpus**
  - Methodological demonstration
  - The explanation of the interpretation

The fundamentals of TNA

- **TNA**
  - Paranyushkin (2011): **Everything is representable as a network, even texts!**
  - Texts as networks:
    - The nodes of an adjacency matrix are text units, mostly words
    - The connection is defined when two words co-occurre within a specified range of text
    - The collocation of the words are presented as connection vectors

Methodological background I.

- **Main RQs**
  - The structure of the text, what is it like?
  - Which are the central concepts, how are they connected
  - Paranyushkin (2011): ***meaning circulation***
- **Fundamental background**
  - Bag of words
  - The whole vocabulary of the text
  - Considering the specifics of the Hungarian language no stemming nor lemmatization was applied
    - highly inflective and agglutinative
    - affixes are directly connected to the words
    - stemming would remove the contextual functions of the words,
    - as well as the individual's
  - N-grams are not applied

Methodological background II.

- **Term adjacency matrix**
  - **WORDij** (Danowski, 2013)
  - Co-occurrences within a range of document, paragraph, sentence or Δx words
    - **In a sentence within the range of three words**
    - Undirected networks
  - **Filters**
    - Term frequency: 2
    - Frequency of co-occurrences: 2
    - Stop list of conjunctions

Vizualization

- The tool for visualization: **Gephi** (Bastian et al., 2009) **https://gephi.org/**
- **Force Atlas** layout (Jacomy, 2009)
- Animation
  - Optimized for small and medium sized graphs

Sullivan (2002): Bourdieu had interpreted the society as an entity existing in „space”

Therefore, a text network drawn by co-occurrences (co-mentions) is a possible map of the text

Network metrics in the case of text networks I.

- **Average Degree**
  - The average number of nodes adjacent to a given node.
- **Average Path Length**
  - The average of geodesic distances one has to travel in order to connect two nodes in the network.
- **Diameter**
  - The longest geodesic distance in the network. The diameter is the maximum number of possible steps on a geodesic distance between two nodes.
- **Density**
  - Describes how dense the network is by comparing the number of existing connections to the number of all the possible connections in a fully connected network with the same number of nodes.

Network metrics in the case of text networks II.

- **Degree centrality (nodal degree)**
  - The number of nodes adjacent to a given node. Degree centrality describes the centrality of a node by the means of how well connected it is.
- **Betweenness centrality**
  - The number of times one has to touch a specific node on a geodesic distance in order to connect two randomly selected node in the network.
- **Modularity**
  - Modularity algorithm identifies the communities within the network. Nodes ordered in the same community have more connections than it would be expected on the basis of chance in a random network with the same amount of nodes and density. The coefficient of modularity equals to the number of edges within a group of nodes minus the number of edges of the group of nodes in the random network.

Degree and betweennes centrality distribution

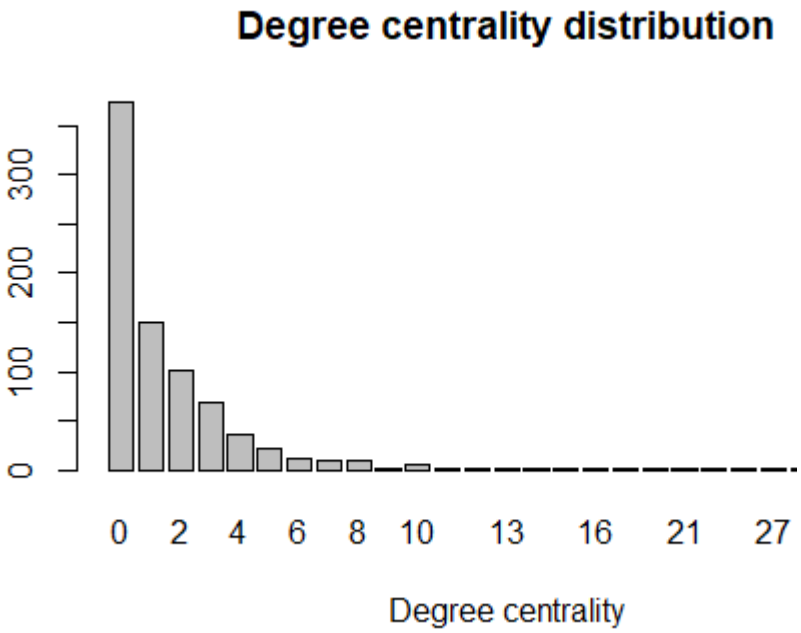
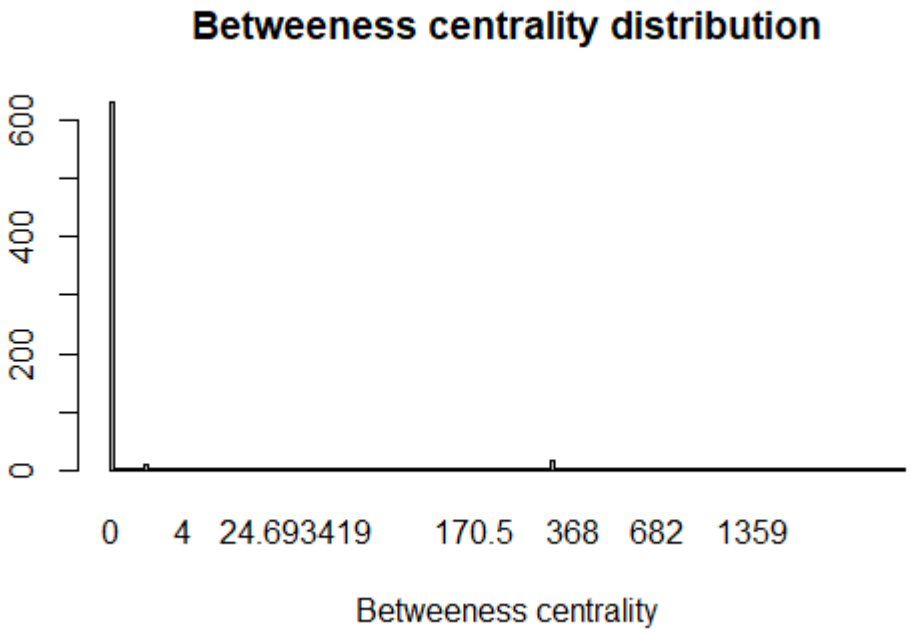


Figure 1. Degree centrality distribution

Figure 2. Betweeness centrality distribution



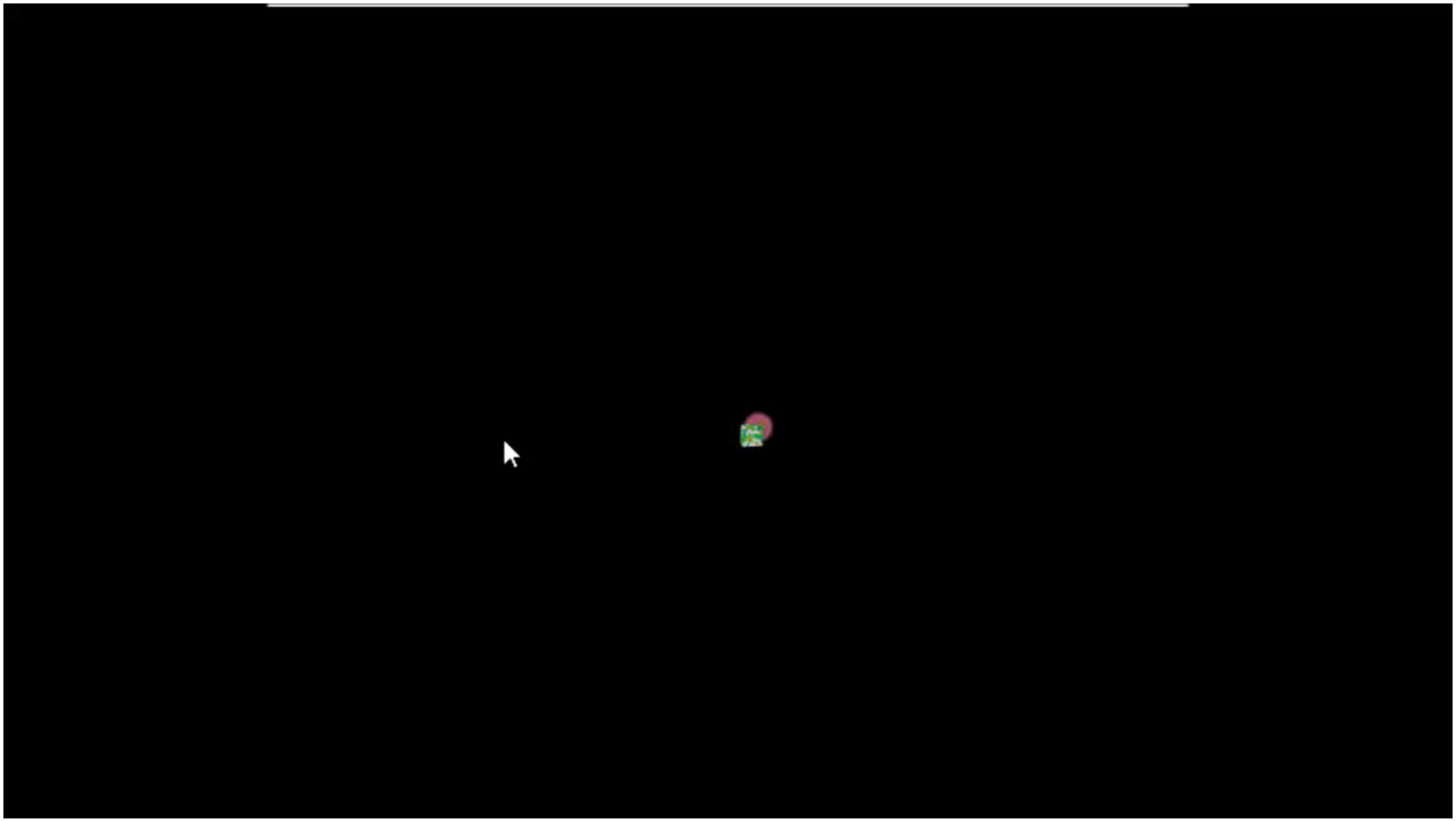
Degree and betweennes centrality distribution

- The scale of the degree centrality is from 0 up to 104
- Numerous word-pairs mentioned together with a notable high frequency
- The longest distance is 10
- Some combination of words shorten the paths
- The modularity of this network is 0.65, suggests that the clusters identified are not random clusters, the nodes therein are more interconnected than those in a random graph
- In total, there are 420 clusters in the network, 9 of them make up more than 2% of the whole network
- Those nodes that are not connected to any others considered as individual clusters
- The density is far from complety with 0.2%

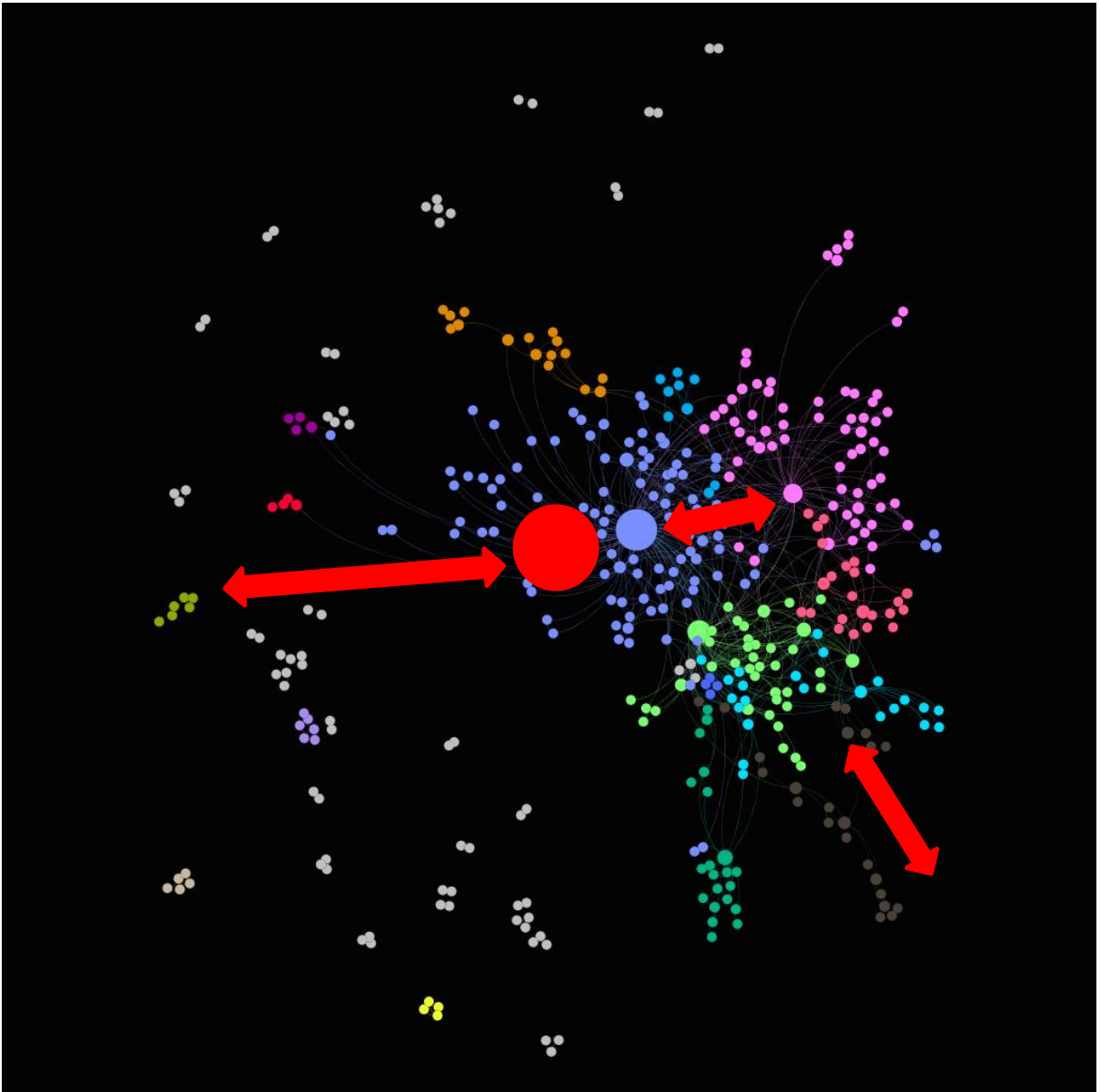
Nodes	806
Edges	783
Average degree centrality	1,94
Average path length	3,63
Diameter	10
Louvain Modularity	0,65
Number of communities	420
Density	0,002

Table 1. Networks statistics of the test corpus

Force Atlas



Force Atlas in details



- **Gravity** scales the distribution of the nodes and pulls every node towards the center of the screen
- **Attraction strength** is the amount of strength the pairs attract each other
- **Attraction distribution** the distribution of attraction between the connections which distributes the hubs

Figure 3. The TNA visualization of the test corpus

Interpretation of the clusters

- Porper names, affexes, personal orders (13,4%)
- Formalities, the partitions of the bureaucracy (5%)
- Nominations, placements (8,2%)
- Az államapparátus és karhatalom elemei (2,9%)
- Press and cultural events, international relations (2,6%)

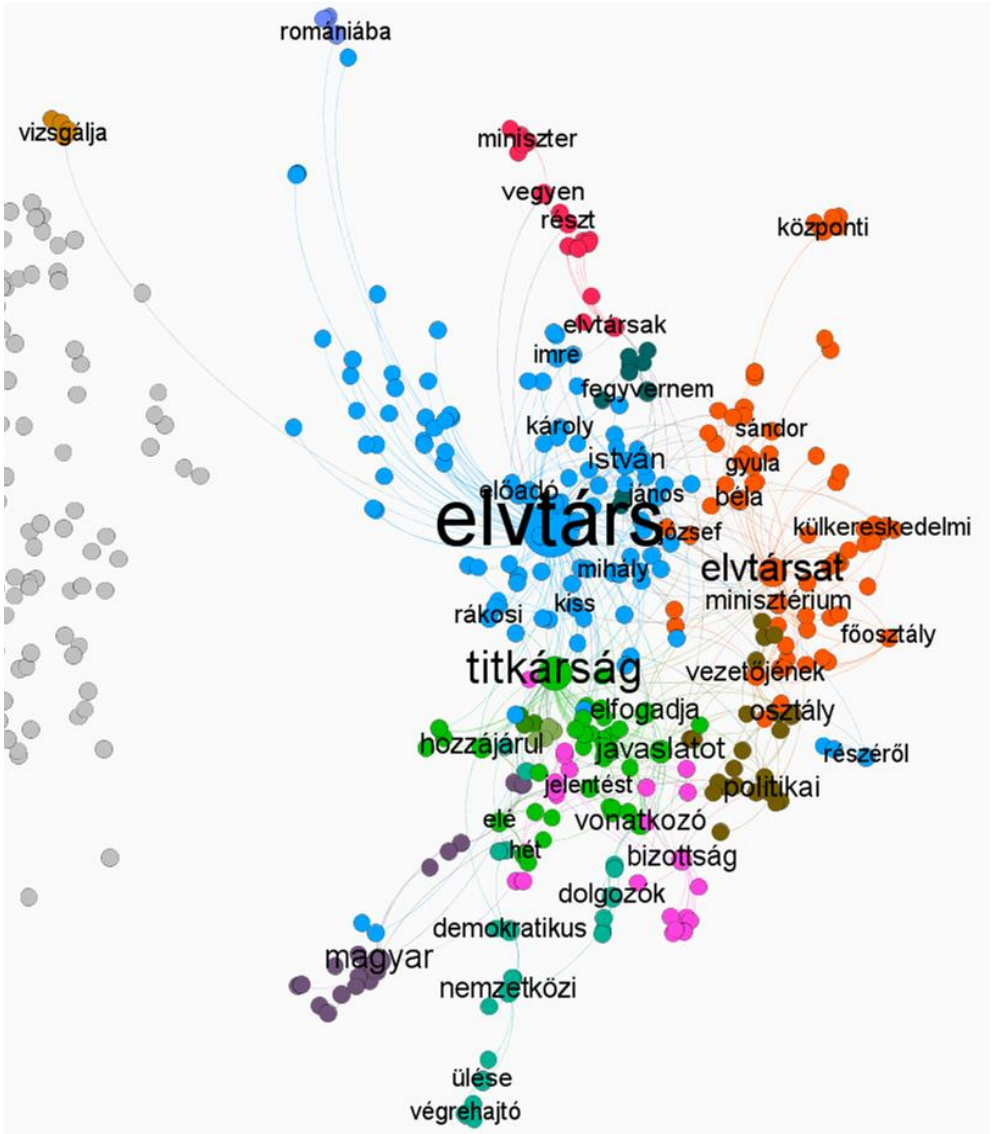


Figure 4. The topic cummunities

The challenges of sentiment analysis on historical corpora

- The special aspects of historical sentiments, such as „comrade” = +1
- Some terms that usually present a negative opinion in Hungarian language are neutral in the reports
- Sentiments are rare in the source documents
- **Conclusions and future research**
- What should be the size of the window of co-occurrences
- Stop list optimization
- Sentiment lexicon optimization for historical corpora

Thank you for your attention!

szabo.martina@tk.mta.hu
havadi.gergo@tk.mta.hu
gulyas.attila@tk.mta.hu
boros.istvan@tk.mta.hu