Russian Twitter Trolls

Social Networks and Recommendation Systems - Project 12

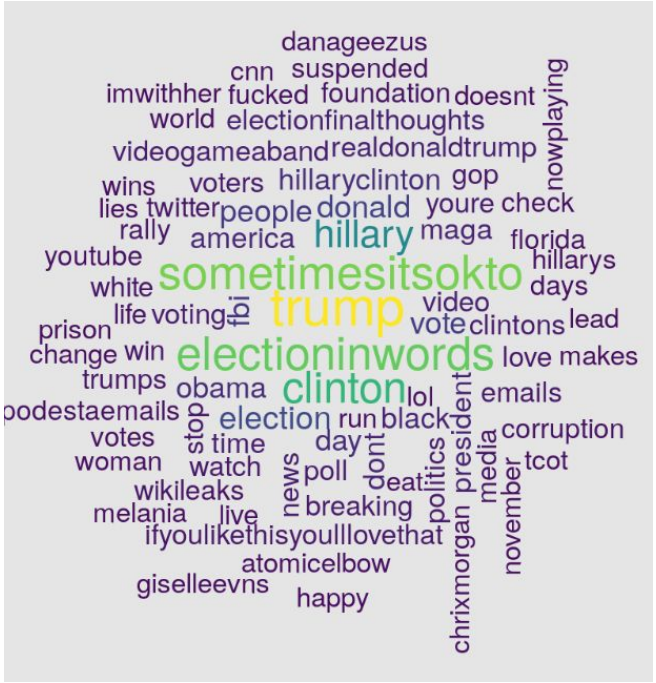
Elżbieta Jowik, Mateusz Wójcik

Agenda

1. Network Generation with Named Entity Recognition
2. Network Visualisation
3. Network Analysis
    a. statistics and visualisation
    b. degree distribution
    c. power law test
    d. PageRank
    e. communities clustering
4. Top Twitter Trolls

Russian Twitter Trolls Dataset

- It consists of two dataframes – *users.csv* and *tweets.csv*
- **203 482 tweets** with **16 columns** of data, for example retweet count, favorite count, text, hashtags
- **453 users** with **14 columns** of data, for example name, number of followers, verification status, favorites count



The most frequent words before the US election.
Source: https://www.kaggle.com/code/jonathanbouchet/beware-of-trolls

Named Entity Extraction

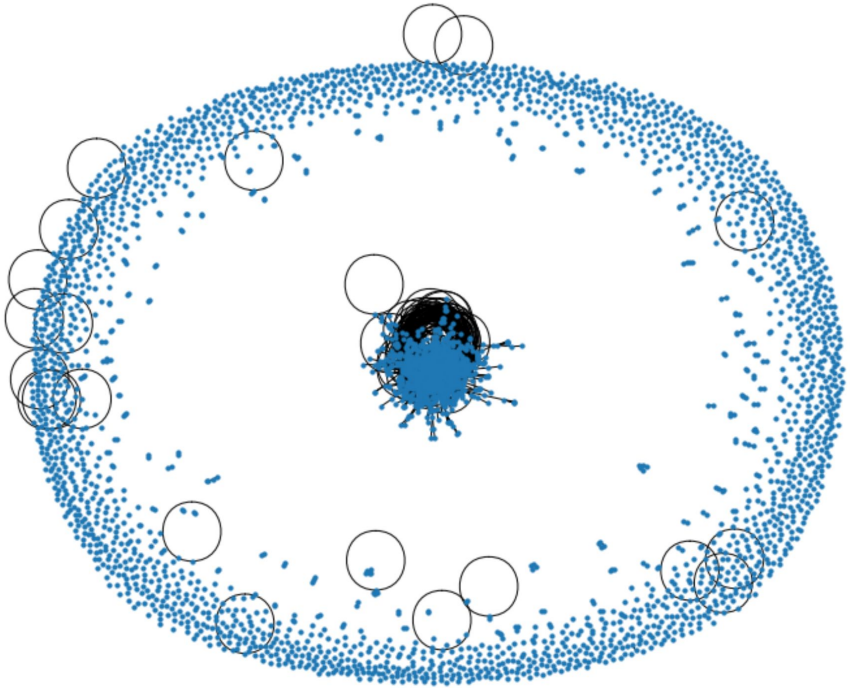
- Named entity extraction task aims to extract phrases from plain text that correpond to entities.
- To find Location entities mentioned in the tweets, we applied Polyglot's entity extraction functionality.
- Polyglot employs link structure from Wikipedia and Freebase → it is a language agnostic technique for entity extraction.
- Polyglot entity extraction works by taking a piece of text (tweets in our case), and annotating (or tagging) the text where it recognizes named entities.

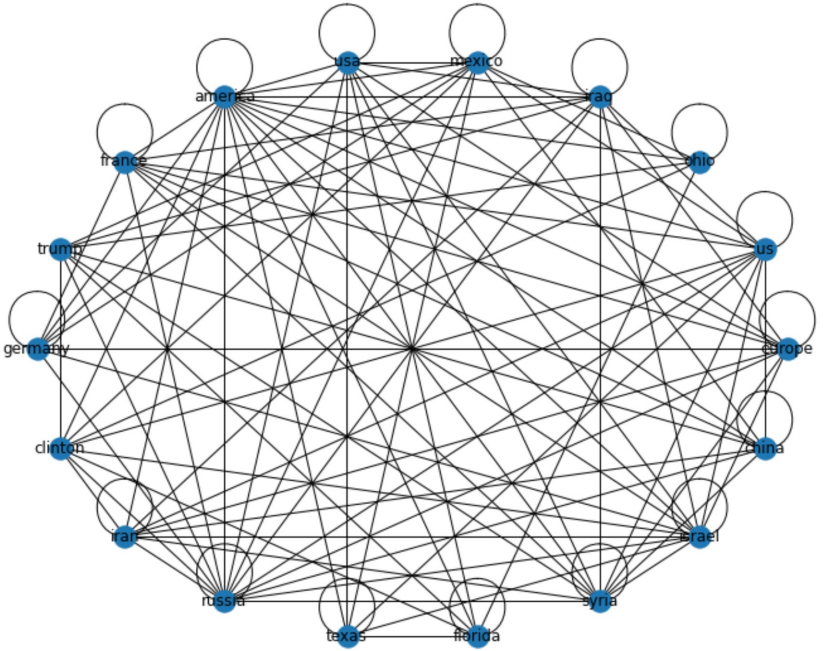
Steps for Network Generation

- We looped through all tweets. For each iteration, we extracted entities that are tagged as I-LOC
- We cleaned the entities one by one:
  - first, we transformed entities to lowercase and stripped the blanks in the for loop
  - then, using FuzzyWuzzy that implements Levenshtein distance to calculate the similarities between words, we trimmed the entities down into a more standard naming convention
  - for the last but not least step before creating the network graph, we got rid of the empty lists of entities that were generated by articles who didn't have any I-LOC entity types
- We generated the graph, in which nodes correspond to entities

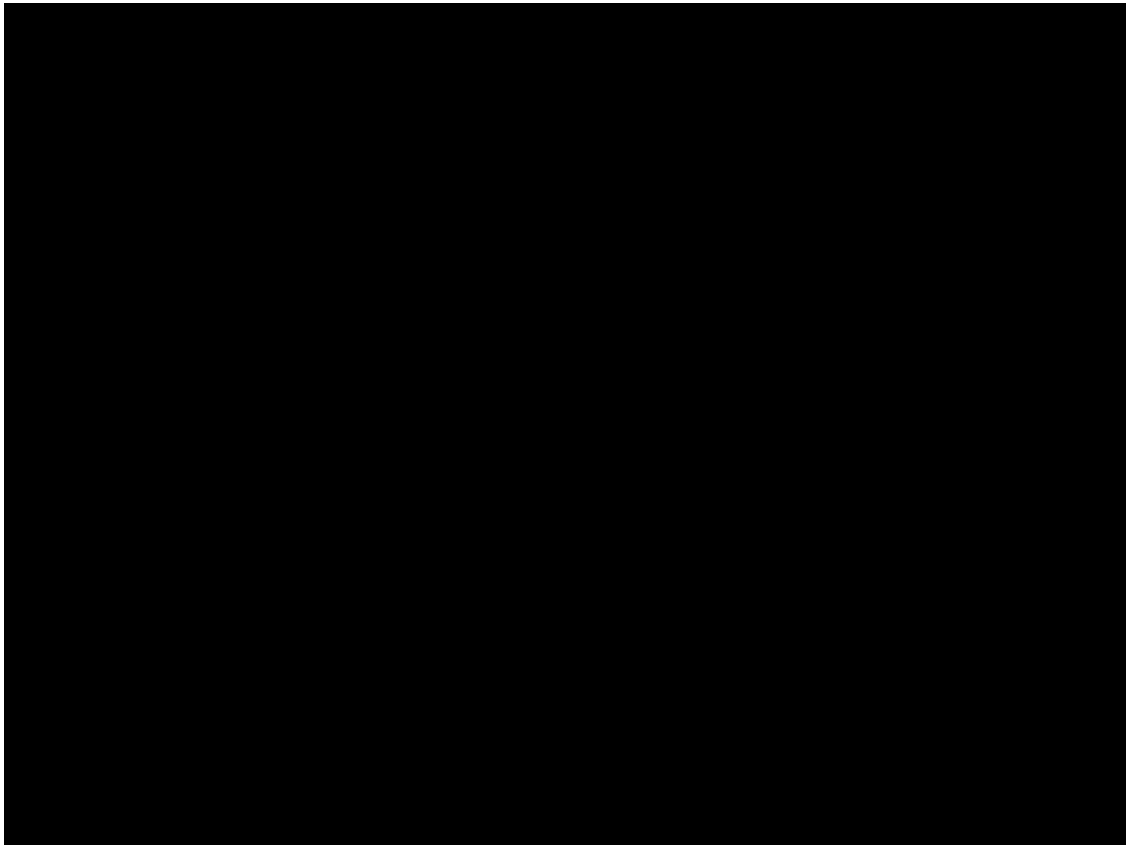
Network Visualisation

The basic visualisation of the whole network. A huge hub and a lot of small disconnected subgraphs can be noticed



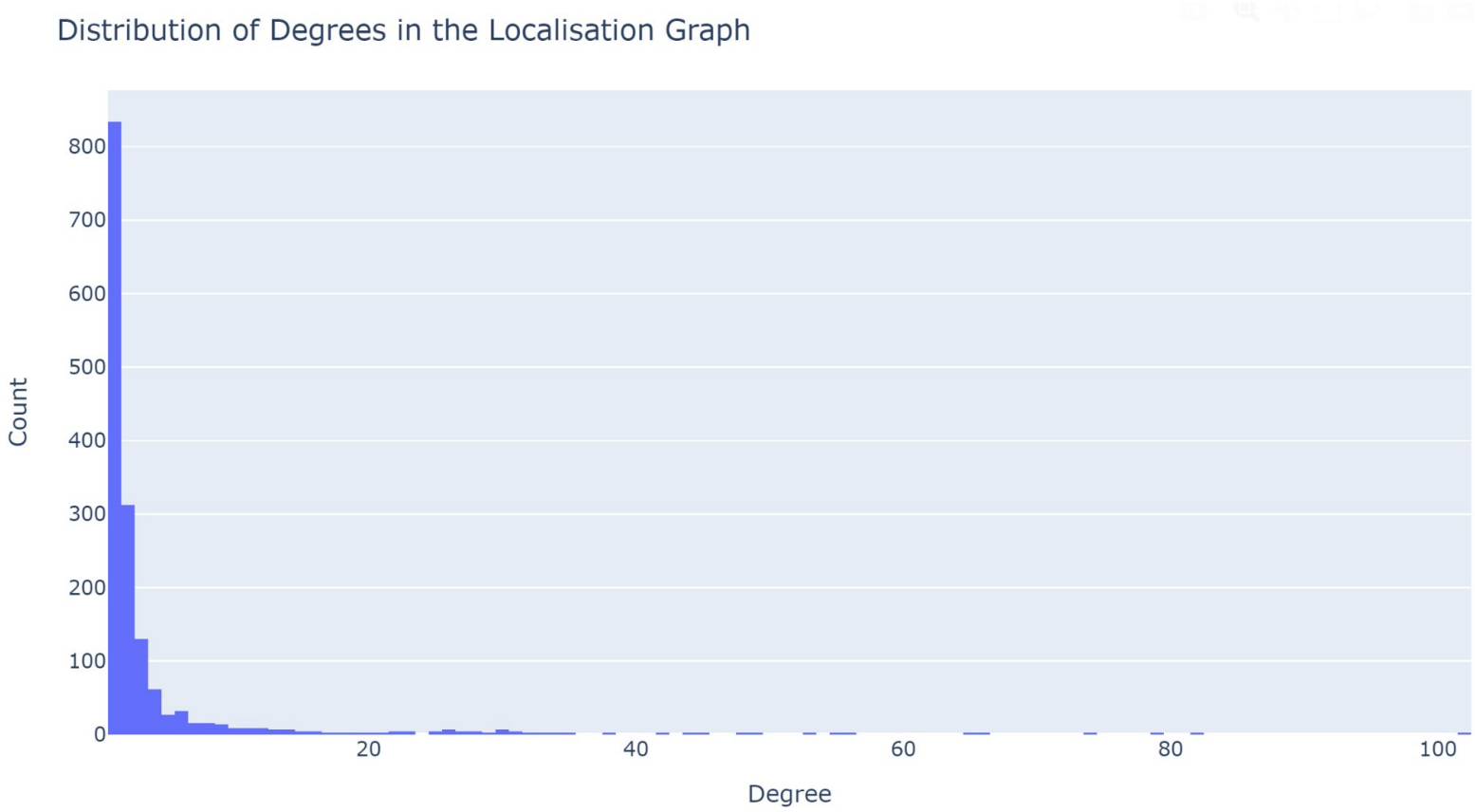


A subgraph generated by the nodes with the degree of at least 36



Interactive visualisation of the whole network

Network Analysis

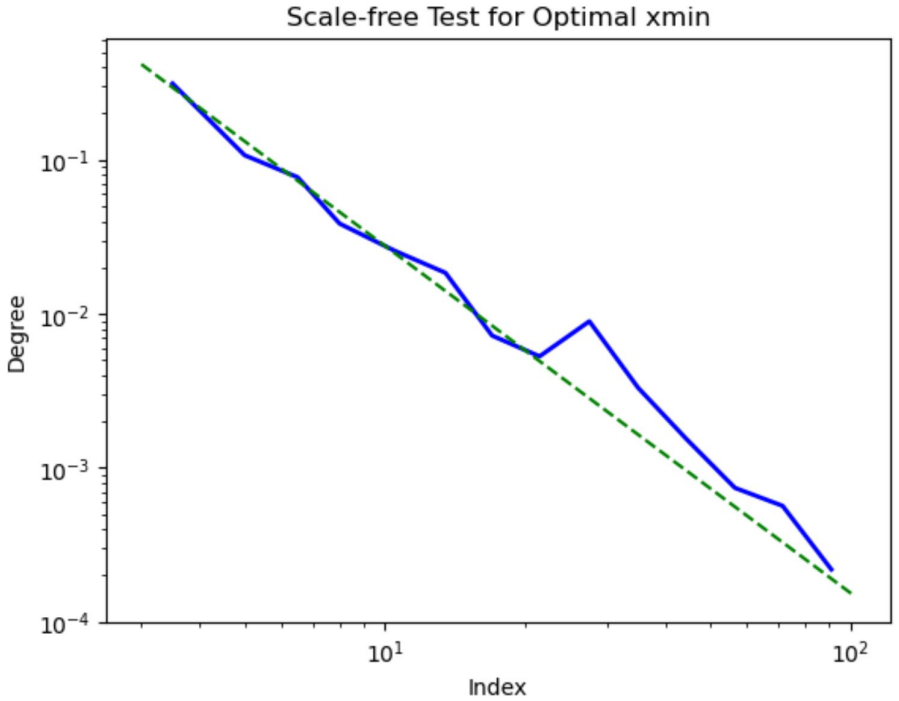


Degree distribution for the whole network. We can see a few nodes of a high degree in the network.

Is the Network Scale-Free?

We used *powerlaw* library available in Python to fit the power law model with the degree sequence of the graph.

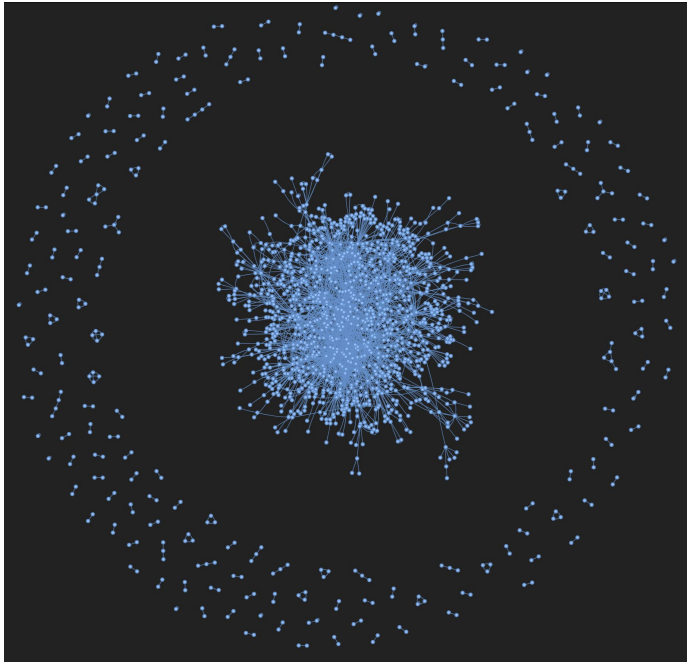
The minimum degree value was optimised automatically by the model. The locally linear trend can be noticed on the graph in comparison with the power distribution.



Connected Components

The main hub in the network, which is the maximum connected component of the graph, consists of **1183 nodes**. On the other hand, the second maximum component consists of just **5 nodes**.

In the further analysis, we considered the main hub by removing the remaining components.



PageRank - Top 10

1. *america* – **0.0129**
2. *russia* – 0.0102
3. *florida* – 0.0096
4. *syria* – 0.0089
5. *israel* – 0.0077
6. *iran* – 0.0075
7. *ohio* – 0.0071
8. *texas* – 0.007
9. *us* – 0.0064
10. *iraq* – 0.0062

Based on degrees and PageRank results, we noticed that a vast number of tweets concern America or particular states of the United States.

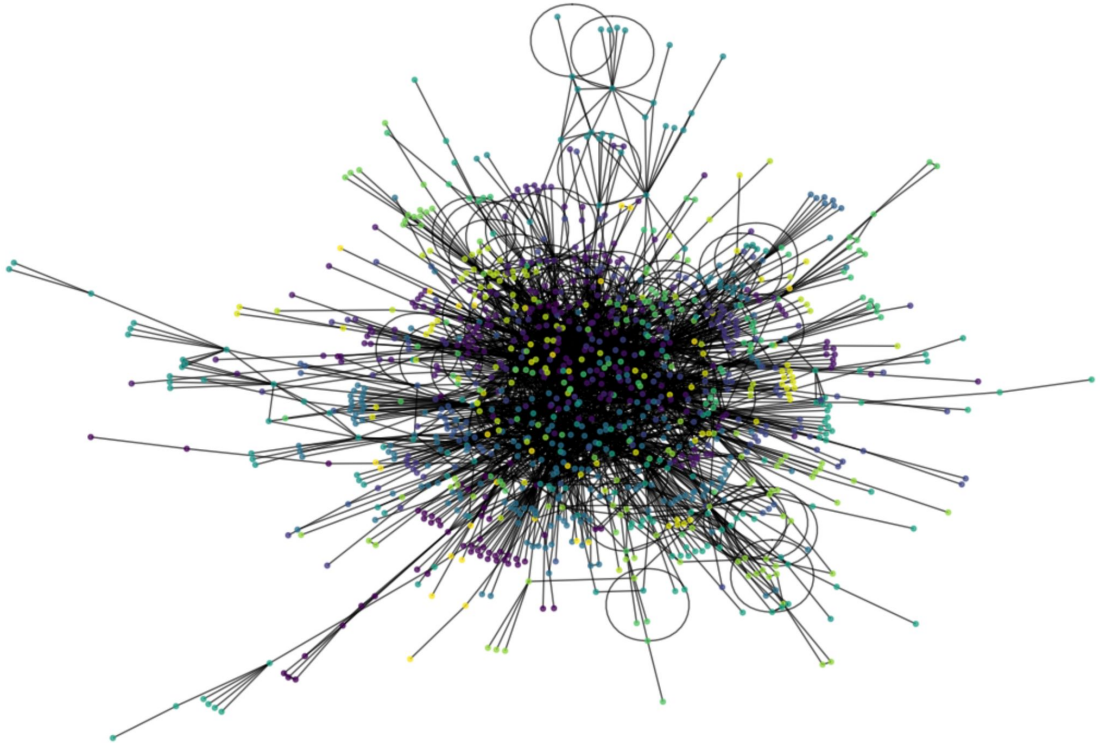
In the next stages of the project, we tried to cluster a community containing *america* entity to find Twitter trolls that mention this location the most. How should we punish them?

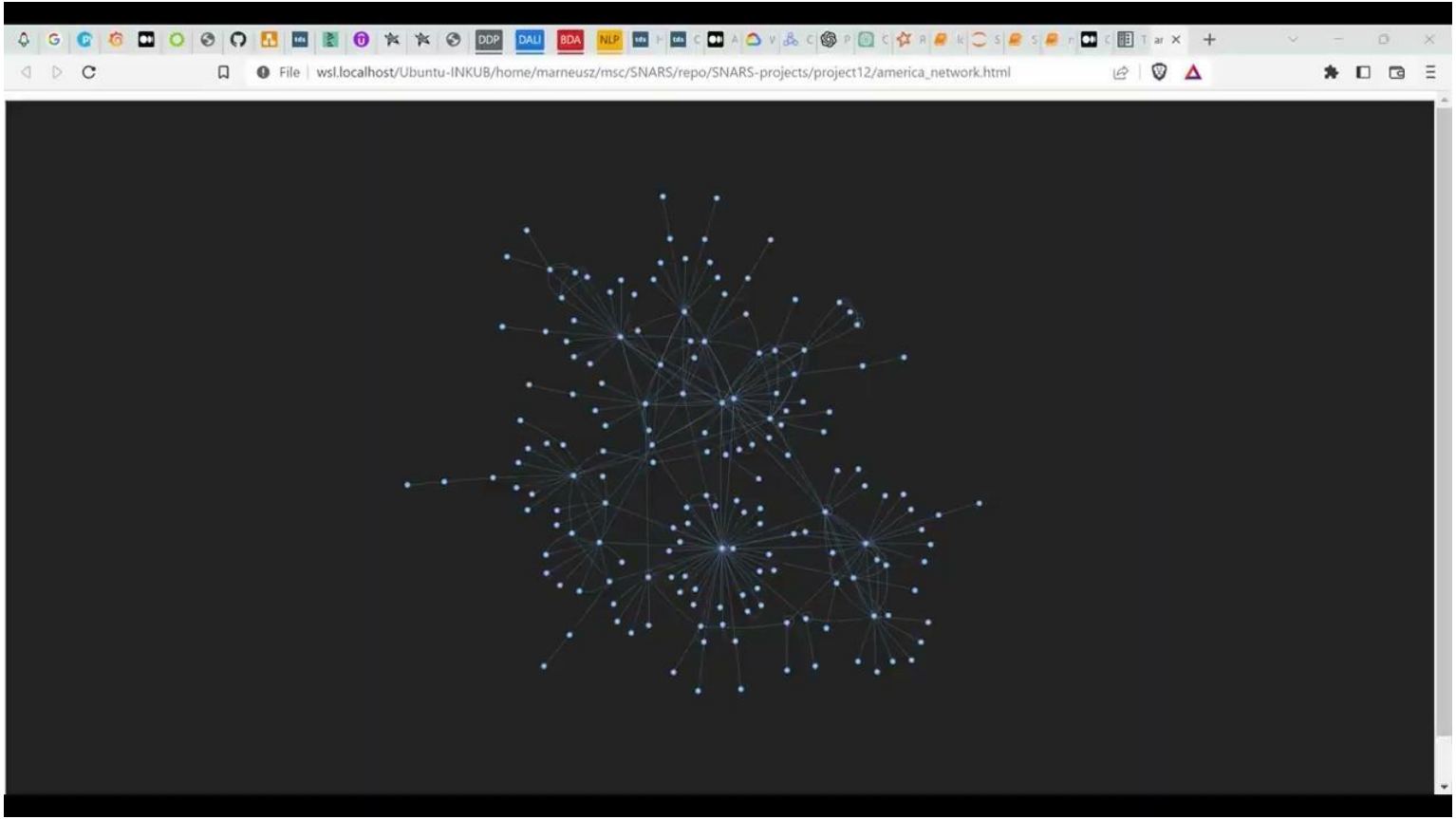
Communities Clustering

Two different community clustering algorithms were used:

- Greedy Modularity Communities
  It found **28 communities** with a high variance of the number of nodes – from 3 to 288 nodes**.**
- Louvain Communities
  It found **20 communities** with the node count between 4 and 181.

Next, we decided to create a subgraph generated by the nodes contained in the “*america* community”.





Interactive visualisation of the subgraph containing *america* community

Information About the Community

- The entities present in the community were mentioned in **6452 tweets** from the given dataset.
- About **250 Twitter trolls** mentioned the same entities.
- The highest number of tweets (421) were submitted by **Amelie Baldwin**.
- The Twitter trolls' accounts writing about the entities were quite popular – at least 900 of followers.
- None of the users were labeled as verified

No.	Name	Followers Count	Tweets Count
1	Amelie Baldwin	2464	421
2	Susan	2225	278
3	Blake Cline	2035	166
4	Pauline Gunn	908	140
5	Steven Cook	1486	125
6	Jackie Best	2078	125
7	Michelle Arrington	3229	123
8	Dorothie	1893	122
9	Nick Luna	1457	119
10	Garret Simpson	1588	117

References

- https://www.lyonwj.com/blog/entity-extraction-russian-troll-tweets-neo4j
- https://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html
- http://nodeassets.nbcnews.com/russian-twitter-trolls/users.csvMASZ2
- https://neo4j.com/blog/story-behind-russian-twitter-trolls/
- https://towardsdatascience.com/populating-a-network-graph-with-named-entities-fb8e3d2a380f