Project 7: Studying ISIS Twitter influence with social network analysis from the pro-ISIS fanboy tweet data

1st Niklas Saari
dept. Computer Science and Engineering
University of Oulu
Oulu, Finland
niklas.saari@student.oulu.fi

***Abstract*—This document describes project work for the course Social Network Analysis in Spring 2021. Radical and extreme groups have taken social networks as part of their toolchain to spread propaganda messages, get attention, look support for their actions and recruit more members. One of the most brutal of these groups, ISIS, which is also designated as terrorist organization, has been forerunner on using these social platforms successfully. To possibly prevent future terror events, it would be important to study these social networks on how they are used by these radical groups. ISIS Twitter dataset of around 17 000 tweets was selected as dataset to identify important characters from the network, to construct communities and to see how the most influencers behave, who they are, and how they connect to other people. Further communities were constructed to see how networks behave internally.**

***Index Terms*—Twitter, Social Network Analysis, Terrorism, Networks, ISIS**

I. Group information

This project has been done alone. The original project from the given project list was 7 with the title of "Analysis of ISIS Twitter dataset". Project source code and also source code of this article can found from https://github.com/Nicceboy/SNA-project-2021

II. Introduction

Social networks have become part of the most people's everyday life. These networks, such as Facebook, Twitter or Reddit are base for many kinds of groups and people for communicating with each other. They are being used widely for expressing opinions or advertising services and for many other kinds of things. This has naturally risen special interest in radical and terrorism organisations because of the provided possibility for influencing different kind of people with a large scale. While the brutality of terrorism has become even more severe over recent time based on the data gathered by Global Terrorism Database and seen in Appendixes in the figure 14, this raises specific interest in terms of identifying and preventing possible future incidents.

Terrorism is identified to be especially brutal from Islamic State of Iraq and Levant (ISIL), which is also known ISIS or Daesh. [1] Their usage of the social media is also known to be "probably more sophisticated than [that of] most US companies", [2] and has been one of their main campaigning tactics in Syria and Iraq. Twitter has been their primal social network [3] so far. It is known that ISIS has been previously organising for example specific hashtag campaigns to get their topics trending and gain more visibility [3].

Because the social networks have begun to be in key role to motivate support for their actions, raise funds and even for recruiting foreign fighters with huge success [4] allowing organisation to reach worldwide recognition and impact, it is important to study how these networks are formed and what we could learn from them and how we could act based on this information.

Social Network Analysis (SNA) is a process for studying social life by social structures which is constructed by relations and patterns formed by these relations. This is usually done with networks and applying the graph theory. [5], [6]

In this article we will focus particularly on the Twitter platform and for the Social Network Analysis of how ISIS has been using this specific platform to spread their propaganda and organizing recruitment. Twitter data of over 17 000 tweets of pro-ISIS fanboys has been used as dataset for this study.

We will try to identify major characters from the provided data, for example which characters have the most influence and what kind of networks they are constructing. This is evaluated based on different values specific for Twitter platform, such as usage of mentions, retweets or hashtags; how are different Twitter users using them.

We further try to estimate the sentiment of the tweets in terms of negative, neutral and positive and estimate the most frequent hashtags and their possible context and characteristics. Sentiments are finally plotted with ternary plots by different characteristics. Different kind of other graphs will be constructed to develop our understanding of underlying network. Social network is constructed by using hashtags, and their relations to other tweets based on their appearances.

This document is structured as following: in the section III the main problem has been described. In the section IV the exact details of the used dataset has been described. In the section V general methodology has been described and further continued in the section VI in more precise matter. The results are presented in the section VII and paper has been finally concluded in the section VIII.

III. Problem Description

The main problem is to study and tell how we can find specific communities from underlying dataset and identify interesting numerical values from the constructed network. Can we detect specific patterns or behaviours related to specific Twitter users, is there connection between them and how powerful their influence actually is? We further try to find a way to tell about what kind of messages they are distributing in overall. Dataset was not collected by itself, instead it was given in the project assignment.

IV. Dataset description

Twitter dataset of tweets collected from pro-ISIS fanboys of all over the world has been used as a base for this study. This dataset was provided with the project assigment. Based on the same project assigment, the origin of the dataset is unknown, as it is stated to be published in dark web website. However, after doing some research, it seems that this dataset is probably collected by Fifth Tribe digital agency, and published originally on the *Kaggle.* [7] Data is under Creative Commons 0 (CC0) license and can be used without restrictions to the fullest extent allowed by law. Data was created originally with the intention of ”to develop effective counter-messaging measures against violent extremists at home and abroad.” [7]

Tweets are located during the period of 1st of June 2015 and 13rd of May 2016, which contains the November 2015 Paris attack as interesting point of event regarding the context of this study. Tweets are had been written with multiple languages, but in general they are in English. Their content is varying a lot; they could be text with varying context, external links to other places, images and videos or rewteets.

Dataset contains total of 17410 different tweets by 112 different users, and was originally given in the newer Excel format (**.xlsx**). Following data columns can be found from the raw data:

- name
- username
- location
- number of followers
- number of statuses
- time (month/day/year 24-hour clock)
- tweet (multilingual)

Location is user supplied data and can be therefore anything.

A. Pre-processing of the data

As the initial dataset was given in Microsoft Excel Open XML (.xlsx) format, it required some conversion to be more suitable for processing with programmatically with selected programming language (Python in this case) and in general for easier handling and compatibility. Dataset was converted to basic **.csv** file format by using Python **pandas** [8] library with **openpyxl** [9] engine. Successful conversion was verified later by checking that there were no null data shells for columns which are considered as ”important” and the amount of rows matches with original and converted data. ”Important” means in this context that every tweet should have at least username and tweet content to be meaningful.

The data has been on some cases further pre-processed as following to extract some specific information and features. This information is stored programmatically on the run-time-memory by creating specific Python class object to represent single line from the dataset data, which also contains extracted additional data.

1) Mentions: Mentions of different users have been extracted from the every tweet based on the ’@’ symbol in tweet data. Twitter usernames are case-insensitive and therefore as and additional step, they are stored in lowercase to improve accuracy of the data and also to reflect real world behaviour when linking to other tweets. This is implemented by using specific regex patterns. Tweet can contain multiple mentions. Retweet contains mention, and these tweets have been ignored to detect explicitly the use of mentions.

2) Retweets: Retweets are identified from the data based on ’RT’ as first word in the tweet.

3) Hashtags: Hashtags have been extracted from the every tweet based on the ’#’ symbol in tweet data. Twitter hashtags are case-insensitive and therefore as and additional step, they are stored in lowercase to improve accuracy of the data and also to reflect real world behaviour when linking to other tweets. This is implemented by using specific regex patterns. Tweet can contain multiple hashtags.

4) Sentiment analysis: Sentiment analysis is applied for every tweet in the dataset to describe the potential category in terms of *negative, neutral and positive.* Python package named as VADER Sentiment, which was originally presented in the article ”VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text” [10] has been used as tool for scoring the data for these categories. This is discussed in more details on the section VI.

B. Data verification after pre-processing methods

As there were many methods implemented for extracting information from the dataset, some testcases were applied and random data was selected to be sure, that extraction is working as expected. This was also applied for data conversion. Testing was implemented by using Python package **pytest** [11], and based on the limited test cases data is extracted as intended.

V. General methodology

In general, the methodology on study is based on processing the Twitter dataset with Python programming language and selected existing libraries. Dataset was pre-processed at first to convert it to suitable format and further some methods are applied to extract features from the tweets, such as hashtags or mentions, as mentioned in the section IV.

NetworkX [12] is the main package used in the work. It is mainly used to ”study structure” or other properties of potentially complex networks. In this scenario it is used to

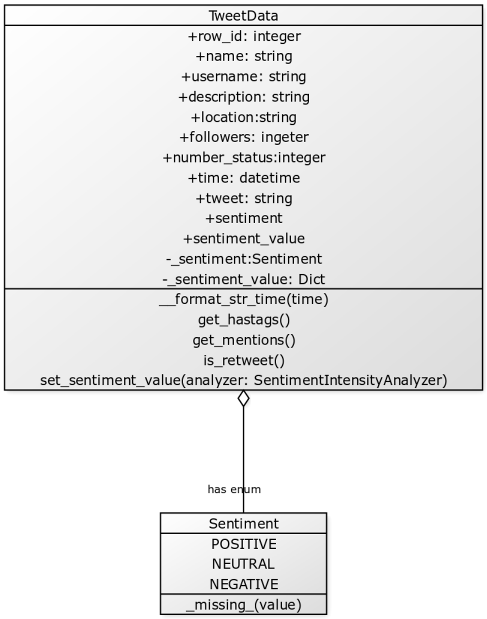

Fig. 1: Designed class for storing and handling the Tweet data

construct a network model based on appearing hashtags in Twitter tweets. Core properties have been measured from the constructed network to make conclusions based on that.

Matplotlib [13] is the main library used for drawing the visible graphs in this document. Also plotly [14] has been used for more advanced visual graphics. Vader Sentiment [10] tool has been used to identify sentiment of the every tweet on some occurrences.

Command-line application has been constructed to make analysis with different parameters easier for the dataset.

VI. Detailed methodology

This section is going through exact details of the process how data was handled and how results were finally obtained. Actual results are described later on the section VII.

A. Initial steps

To get started with the analysis our Twitter data, we are required to shape data suitable for our use case. This was briefly described in the section IV on ”preprocessing” subsection. Data was converted into **.csv** format and loaded into program memory.

Initially, specific Python class was designed to allow ease access for tweet data and feature extraction. UML diagram can be seen in the figure 1. This class has build-in methods for extracting specific features from the tweet data and it converts values to suitable data types e.g. time into build-in Python time object. Methods include extracting hashtags, mentions and identification is the tweet retweet, and evaluation of possible sentiment value of it.

Sentiment analysis is used here to make mass analysis for all tweets to categorize every tweet for one of the NEGATIVE, NEUTRAL and POSITIVE semantic orientations. This is useful stat to generalize what kind message(s) individuals are tweeting.

Individual Twitter users will be sorted and handled separately by using different metrics: by use of the retweets, mentions and use of hashtags. Also, sentiments among all the tweets will be evaluated. VADER (Valence Aware Dictionary for sEntiment Reasoning) tool [10] has been used for making the sentiment calculation. It was released in 2014 and has acquired fame from its efficiency and accuracy. By default, tool is producing four different values for provided input as results. Three of the values (pos, neu and neg) are pointing the portion of the three orientations, from range to 0 into 1, totalling to sum of 1. One is describing overall result (compound), which ”is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive).”

Therefore, sentiment orientation is finally defined by compound value as following:

score $\geq$ 0.05	Positive sentiment
score $>$ -0.05	Neutral sentiment
score $\leq$ -0.05	Negative sentiment

These threshold values are often used among different researchers and are commonly agreed. This classification is implemented on the Sentiment enum in the figure 1. It is giving one of the three sentiment description for every tweet based on calculated compound value, automatically.

Ternary plots are used to describe the results, as they often can give clear visual evidence which variable from the three is the controlling one. During the assessment of class creation, it was observed that instancing SentimentIntensityAnalyzer from Vader tool is resource consuming, hence analyzer is instanced only once outside TweetData class and passed into object afterwards to significantly increase performance.

After we have loaded the **.csv** file, it is parsed in a such a way that from each row in the data, new instance of TweetData object has been created and list form them is formed. Further, this list is processed in such a way, that every Twitter user is separated to be part of the new <username>:USER_META dictionary object, which has following attributes; amount_tweets, amount_retweets, amount_mentions, amount_hashtags and all the tweets this user has tweeted. This gives us access by username for metadata and user-specific tweets.

Some specific functions have been additionally created that previously created dictionary can be easily sorted, based on it’s key-value attributes and finally plotted with suitable graphs.

B. Visualising typical properties

We can plot and sort users based on their amount of tweets, amount of retweets, amount of mentions and by the amount of hashtag usages. Matplotlib Python package has been used for this.

We are also interested to know, whether power law is applying here, in the context of tweets by single users.

Power law is known to be functional relationship between two quantities, where probability distributions are "heavy-tailed". Pareto principle (80-20) is commonly used as rule to see if data fits under power law [15].

Probability distribution of power law can be seen as:

$$p(x) \propto x^{-\alpha} \tag{1}$$

where α is a constant known usually as scaling parameter and *x* as quantity. Quantity *x* obeys the power law if it is drawn from this distribution. [16]

Python package called as **powerlaw** [17] was used for fitting data into distributions and estimating statistical significance. With powerlaw, data can be fitted into Complementary Cumulative Distribution (CCDF) and compared to theoretical power law, truncated power law, exponential and lognormal distribution candidates for example. It has specific function called as *distribution_compare* which automates a lot of underlying math for calculating two values *R* and *p*. Kolmogorov-Smirnov distance algorithm has been used to estimate whether data fits to theoretical estimate. R is the loglikelihood ratio between the two candidate distributions, which will be positive if the data is more likely in the first distribution. P value is describing statistical significance of this value.

C. Network construction and main properties

NetworkX Python package was used for constructing network from hashtag usage in tweets. Every hashtag is describing unique node in the network, and edges are describing tweets where two or more different hashtags have been used: there is link between diffrent hashtags by tweets.

Graph will be studied by the amount k of occurrences of hashtag pairs: e.g. at least 1, 2, 3, 4, 5, 10, 15 total pairs must be found from the network. For example with k = 5, there must be at least 5 tweets from all the tweets where this specific hashtag pair is appearing. With help of this, we are able to measure graph to see which are the most used hashtags based on Degree Centrality, with different thresholds.

This is applied in the graph generation phase, as following:

1) Extract hashtags from the tweet
2) Generate unique pairs from detected hashtags (With Python itertools.combinations() function)
3) Use dictionary to measure count. Order does not affect.
4) Finally add edges into the graph, which fulfill k threshold. Graph is undirected.

Graph must be recreated for every k value. To visualize how the selection of k affects the total network, heatmap will be drawn.

Some core measurements will be applied for the constructed network: calculation of number of nodes, edges, diameter, average clustering coefficient, average degree centrality, average closeness centrality, average betweenness centrality and the size of the largest component. Diameter can be calculated usually only then, if network is fully connected, otherwise it is infinite. However, we can try to estimate diameter by finding the biggest component, and creating subgraph from that, finally calculating diameter for that.

1) Clustering coefficient: Clustering coefficient value describes how close nodes neighbours are to being full graph (a clique). Value can be calculated for undirected graph as:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}. \tag{2}$$

where N is neighbourhood, E describes all nodes, v is single node, e is single edge and k as number of nodes.

2) Degree Centrality: Degree centrality is the sum of in-degree and out-degree. It is representing the amount of edges entering and leaving the nodes respectively. The most important nodes have the most direct connections with others under degree centrality. The value can be computed as,

$$C_D(v_i) = \sum\nolimits_j A_{ij} \tag{3}$$

3) Betweeness Centrality: Betweeness centrality is another way to measure importance of nodes. It is describing the amount of the shortest path passing the node. Important nodes have high betweeness centrality, information is flowing through them, and they are connecting multiple nodes into the network.

$$C_B(v_i) = \sum\nolimits_{v_s \neq v_i \neq v_t \in V, s<\mathrm{t}} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \tag{4}$$

4) Closeness Centrality: Closeness Centrality means how center the node is compared to all other nodes. It is calculated by measuring the shortest paths; more closer the node is towards center, closer it is to all other nodes. Calculating Closeness Centrality for undirected graphs (as in this case) is special case, and can be calculated by using Dangalchev's method [18].

$$D(x) = \sum_{y \neq x} \frac{1}{2^{d(y,x)}} \tag{5}$$

where d means distance function.

5) Average Values: Average values are calculated in general as:

$$M = \frac{1}{n} \sum_{v \in G} c_v \tag{6}$$

where n is the number of nodes in G and c is this specific value of single node.

D. PageRank

Further, PageRank has been calculated for the network. It can be thought as variant of eigenvector centrality. Originally PageRank was used to rank web pages, but is suitable for ranking other networks as well. [19] With PageRank, we can rank network components by their incoming links. Explanation how PageRank works is left out the scope, as it was handled in the course lectures as well. Corresponding NetworkX function for PageRank will be used, NumPy variant selected for speed.

PageRank distribution will be plotted by rank and value. Additionally, we will plot local clustering coefficient (LCC) for comparison. Plot type for LCC will be histogram as duplicate values are expected.

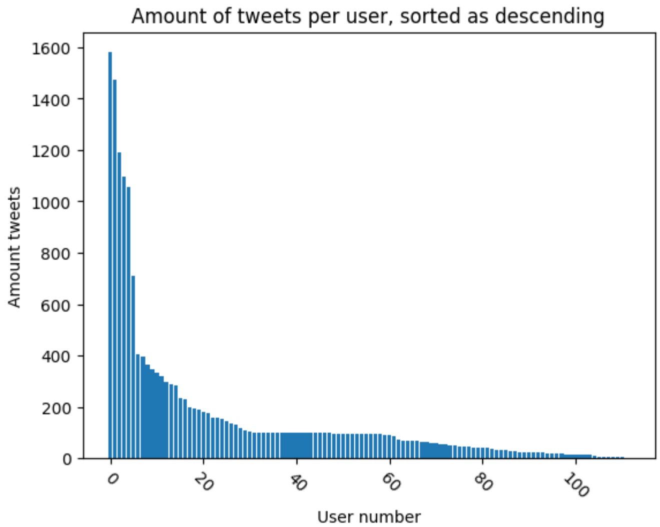

Fig. 2: Amount of tweets per user

VII. Results and Discussion

Initial dataset contained more than 17 000 tweets and 114 different users. Multiple different methods we applied for analysing data, and results will be discussed in this section.

A. Basic measurements

Distribution of the tweets per single user can be seen on the figure 2. Some users have tweeted significantly more than most.

This data distribution seems to be "heavy-tailed" which is typical in power law theory, but for further clarification, it was fitted into Complementary Cumulative Distribution Function (CCDF) and compared into other candidate distributions.

As seen in the figure 3, it is not so clear what is the closest suitable distribution. This strongly suggests, that Twitter user behaviour follows the traditional clause "rich get richer", in terms of Twitter tweets.

Loglikelihood ratios where compared to estimate the most likely fit. We can see from the table I the most likely fit might be Truncated Power Law. Comparison between different distributions are pointing that lognormal or truncated power law are the most likely, where truncated power law is significantly stronger fit ($p > 0.05$). Usually this means we might have too many data entries and the rest of the data "fall-off" from the power law or simply not enough data. We might be able to analyse tails bit better, but it is left out of the scope in this assignment.

Data was further sorted for top 10 users by mentions, retweets and hashtag usages and plotted by Matplotlib. Results are shown in the figure 6, 4, and 5. Usernames are partially shortened to not make them totally public.

These three figures show that similar usernames are appearing on all three categories, but mostly in different order. This might be partially explained with the power law, as minority of the user amount are mostly responsible from the tweets, and

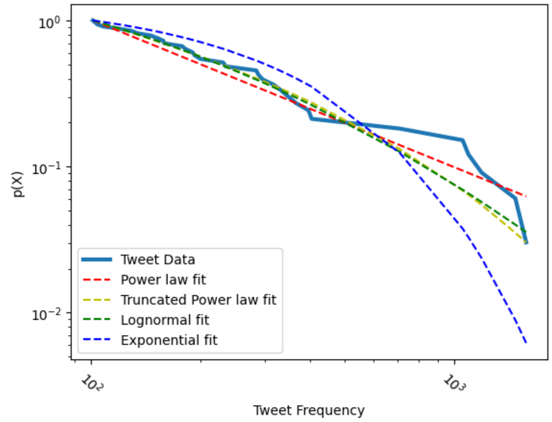

Fig. 3: Distribution candidates for amount of tweets per user

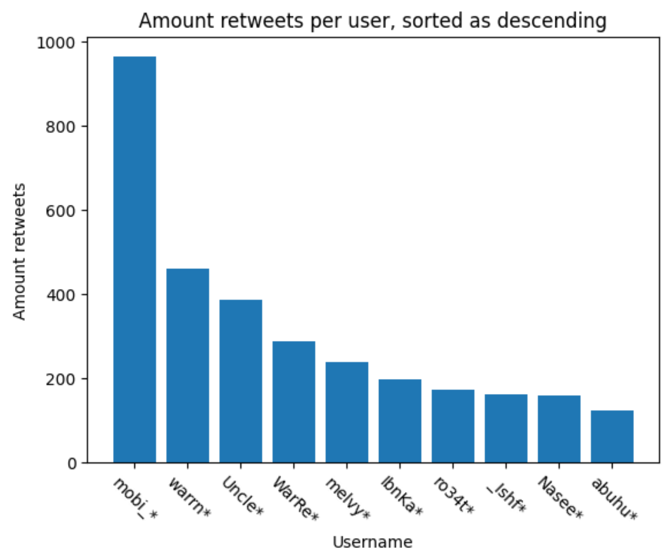

Fig. 4: Top10 users by retweeting.

with enough content diversity of the tweets, they can reach top position on multiple categories. Based on this, these users are effectively using properties of Twitter platform to connect to other users, potentially gaining more visibility. These users are mainly responsible on distributing messages and sharing other messages, giving visibility with hashtags.

B. Sentiment analysis

Sentiment analysis was applied for whole tweet data and separately for tweets of previously calculated top10 users by different categories.

Distribution 1	Distribution 2	R	p
Power Law	Lognormal	-0.96	0.34
Power Law	Exponential	0.83	0.40
Exponential	Lognormal	-1.65	0.10
Power Law	Truncated Power Law	-1.34	0.10
Lognormal	Truncated Power Law	-0.36	0.27

TABLE I: Loglikelihood rations of the different distributions

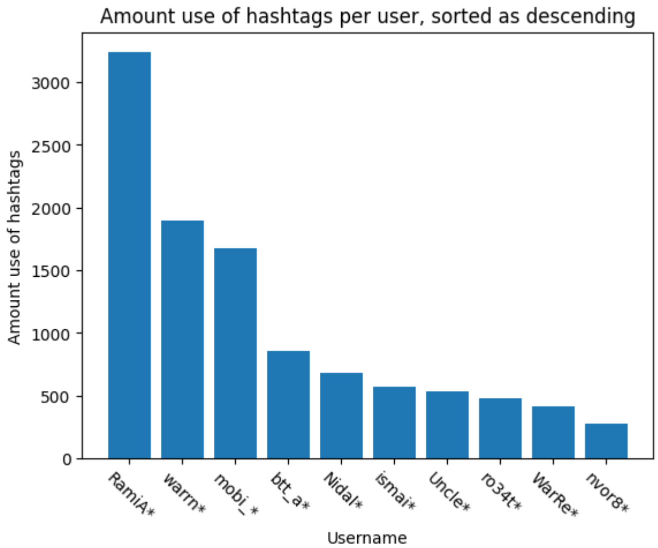

Fig. 5: Top10 users by usage of hashtags.

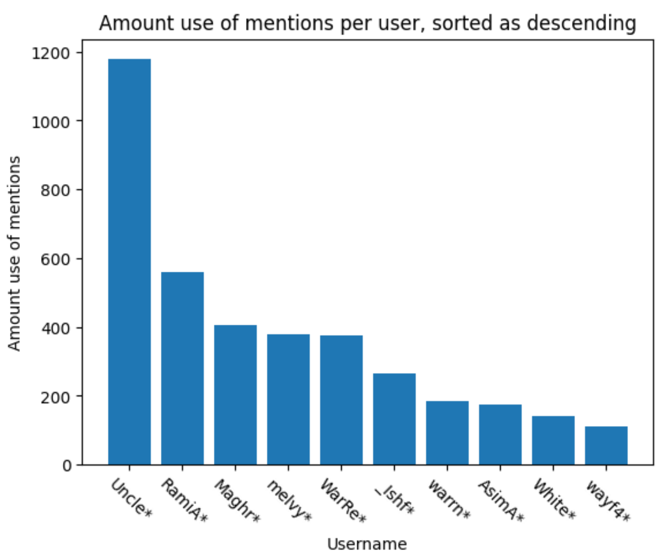

Fig. 6: Top10 users by usage of mentions.

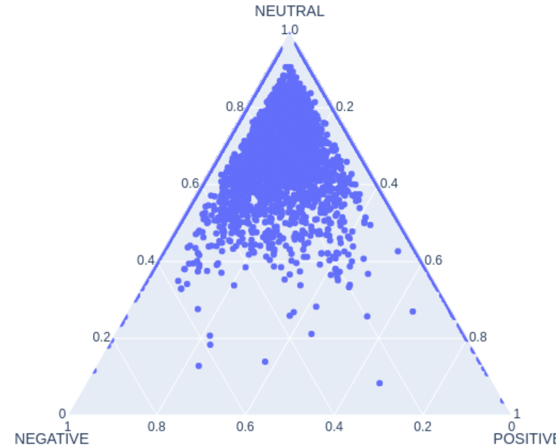

Fig. 7: Ternary plot of sentiment analysis of all tweets.

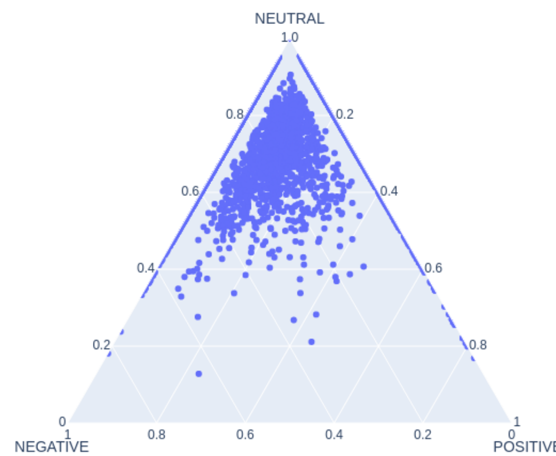

Fig. 8: Ternary plot of tweets combined of top10 users by usage of hashtags

In the figure 7 we can see ternary plot of sentiment analysis among all tweets. Distribution isn't totally clear, as there are many values with zero negative or positive weight, which leads points to be on the border of the ternary.

On a quick glance it seems that tweets are more weighted on negative side in general. After taking exact numbers, negative seems to have slight majority, and positive is in clear minority.

- Amount of NEGATIVE values: 6948
- Amount of NEUTRAL values: 6657
- Amount of POSITIVE values: 3805

It implies that content of tweets is siding more on negative/neutral side than neutral/positive side in general.

In figure 8 we can see ternary plot, where tweets are limited to top 10 users by usage of hashtags. In figure 9 ternary plot is constructed by usage of mentions of top 10 users. Finally in the figure 10 ternary plot is created by the use of retweets.

It is difficult to see drastic differences by the eye on ternary plots. The most significantly it seems that users with the most use of hashtags, sentiment value seems to be mostly negative.

TODO add table about values and percentage portions

C. *Hashtag Network Analysis*

Hashtag network was build successfully with NetworkX. To get initial visual view, we can take a glance on figure 15 and figure 16 on Appendixes. There is one giant community. This graph was created with $k = 0$; e.g. all nodes are included. We wanted also to examine graph by the amount of hashtag pair occurrences k, which limits only those hashtags pairs into the network which are used at least k times. The core properties

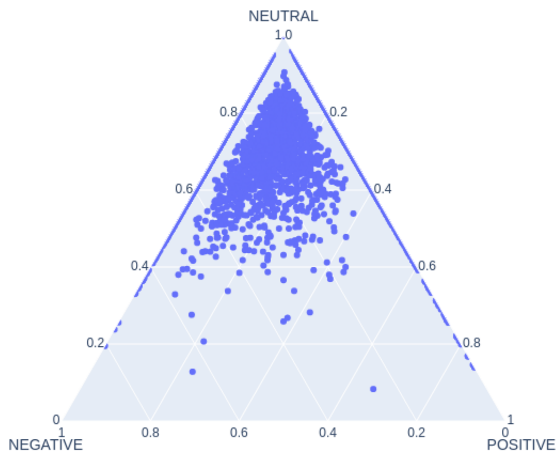

Fig. 9: Ternary plot of tweets combined of top10 users by usage of mentions

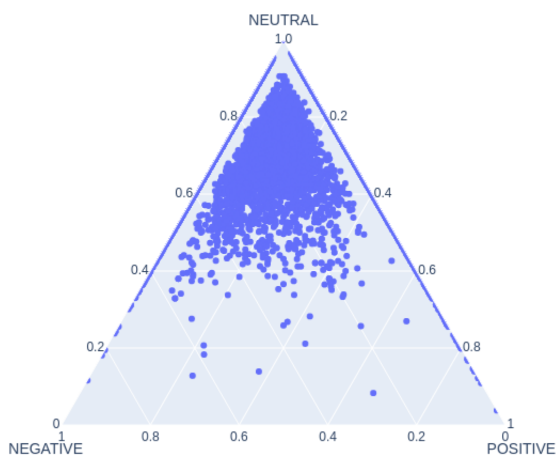

Fig. 10: Ternary plot of tweets combined of top10 users by retweeting

can be seen in the tables III and IV. Only difference with these two tables are, that table IV has nodes removed which are not fulfilling the k requirement. Visual graph with reduced graph (k = 2 was selected), we can see the core structure bit better in the figure 17.

There are around 4x amount of edges compared to nodes, hence the network has many connections, when looking at without k limitations. All nodes are not connected and diameter was estimated based on the biggest component. It was expected that network will not be totally connected - some specific hashtag might appear only alone in the tweet.

The effect of k requirement can be seen clearly. Some hashtag pairs are used very often with other hashtags as we can find plenty of nodes even with k = 15, and they are creating

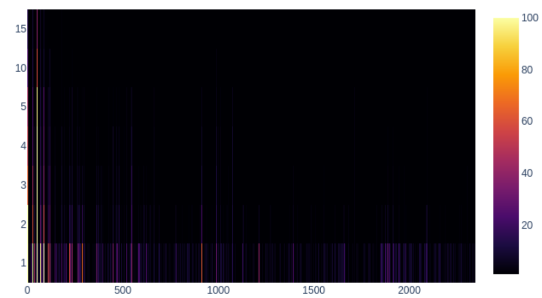

Fig. 11: Heatmap of the hashtag graph, with k values 1, 2, 3, 4, 5, 10, 15

Rank	Hashtag
1	isis
2	islamicstate
3	syria
4	is
5	breaking
6	amaqagency
7	iraq
8	caliphate_news
9	aleppo
10	assad

TABLE II: Top10 hashtags by PageRank

close communities as closeness and inbetweenness values are increasing. Degree centrality and clustering coefficent are not changing significantly, which implies that on average hashtags are used quite "equally" with other hashtags. Plotted heatmap is pointing also the importance of some hashtags over others in the figure 11. Maximum value in the heatmap is limited into 100, because few hashtag occurrences were significantly more frequent (400+) than others, making heatmap hard to observe.

For further ranking these connections, PageRank distribution has also been calculated and can be seen on the figure 12. Based on the figure, we can deduce that small portion of hashtags has significantly higher rank value that most of them. Out of the curiosity, top 10 hashtag names based on PageRank are listed in the table 7. Note, that names are normalized to be lowercase, since they are case-insensitive when construction networks on Twitter platform behind the scenes.

Histogram of LCC can be seen in the figure 13. As the total amount of nodes (unique hashtags) were 2347 in initial graph, we can see that clear majority of them has either LCC value close to zero (600+) or 1 (more than 1000), meaning they have barely connections or they are almost fully connected, being "complete graph".

To get more information about underlying communities inside network, Girwan-Newmann algorithm was used. Based on the initial visual graphs, k value of 3 was initially selected as depth to get communities. It looked a good candidate to limit communities to the biggest ones inside the first giant

k	Nodes #	Edges #	Largest component	Diameter	Avg. Clustering Coef.	Avg. Degree	Avg. Closeness	Avg. Betweenness
1	2347	8215	1877	9	0.59	7.0	0.2	0.0
2	2347	2149	591	9	0.16	1.83	0.02	0.0
3	2347	1094	339	11	0.07	0.93	0.01	0.0
4	2347	753	218	6	0.05	0.64	0.0	0.0
5	2347	561	177	6	0.05	0.48	0.0	0.0
10	2347	249	89	5	0.02	0.21	0.0	0.0
15	2347	156	59	4	0.01	0.13	0.0	0.0

TABLE III: Core measurements of hashtag graph by minimal hashtag pair occurrences (k)

k	Nodes #	Edges #	Largest component	Diameter	Avg. Clustering Coef.	Avg. Degree	Avg. Closeness	Avg. Betweenness
1	2078	8215	1877	9	0.66	7.91	0.26	0.0
2	653	2149	591	9	0.57	6.58	0.28	0.0
3	374	1094	339	11	0.46	5.85	0.28	0.0
4	244	753	218	6	0.51	6.17	0.31	0.01
5	192	561	177	6	0.55	5.84	0.33	0.01
10	96	249	89	5	0.51	5.19	0.37	0.01
15	64	156	59	4	0.55	4.88	0.39	0.02

TABLE IV: Core measurements of hashtag graph by minimal hashtag pair occurrences (k), removing nodes without connections

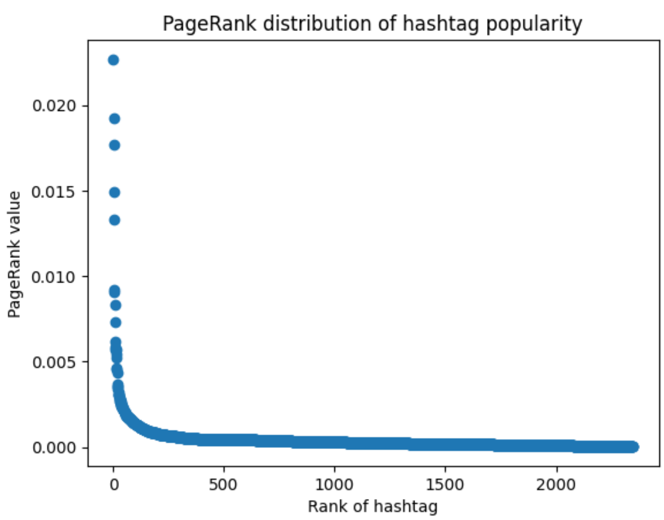

Fig. 12: PageRank of different hashtags.

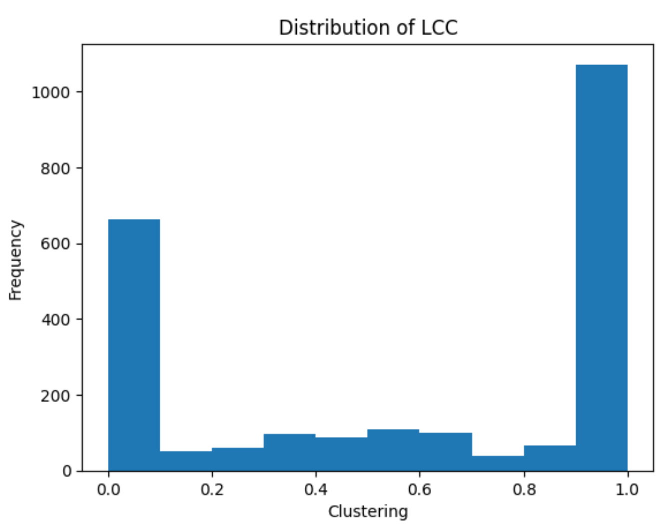

Fig. 13: Local Clustering Coefficient distribution

community.

VIII. Conclusion and Perspectives

Graph study seems to be efficient way to study Twitter network. I was able to identify

A. Authors and Affiliations

B. Identify the Headings

C. Figures and Tables

Acknowledgment

References

[1] Wikipedia contributors, "Islamic state of iraq and the levant — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Islamic_State_of_Iraq_and_the_Levant&oldid=1024074085, 2021, [Online; accessed 19-May-2021].

[2] R. Khalaf and S. Jones. (2014) Selling terror: how isis details its brutality. [Online]. Available: https://www.ft.com/content/69e70954-f639-11e3-a038-00144feabdc0

[3] J. Berger. (2014) How isis games twitter. [Online]. Available: https://www.theatlantic.com/international/archive/2014/06/isis-iraq-twitter-social-media-strategy/372856/

[4] J. Stern and J. Berger. (2015) Isis and the foreign-fighter phenomenon. [Online]. Available: http://www.theatlantic.com/international/archive/2015/03/isis-and-the-foreign-fighter-problem/387166/

[5] A. Marin and B. Wellman, "Social network analysis: An introduction," *The SAGE handbook of social network analysis*, vol. 11, p. 25, 2011.

[6] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002. [Online]. Available: https://doi.org/10.1177/016555150202800601

[7] F. Tribe. (2019) How isis uses twitter. [Online]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

[8] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 2010, pp. 51–56.

[9] E. Gazoni and C. Clark. (2021) openpyxl - a python library to read/write excel 2010 xlsx/xlsm files. [accessed 19-May-2021]. [Online]. Available: https://openpyxl.readthedocs.io/en/stable/

[10] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, May 2014. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[11] H. Krekel, B. Oliveira, R. Pfannschmidt, F. Bruynooghe, B. Laugher, and F. Bruhin, "pytest 6.2.3," 2004. [Online]. Available: https://github.com/pytest-dev/pytest

[12] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, and J. Millman, Eds., Pasadena, CA USA, 2008, pp. 11 – 15.

[13] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[14] P. T. Inc. (2015) Collaborative data science. Montreal, QC. [Online]. Available: https://plot.ly

[15] Wikipedia contributors, "Pareto principle — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Pareto_principle&oldid=1023956740, 2021, [Online; accessed 21-May-2021].

[16] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009. [Online]. Available: https://doi.org/10.1137/070710111

[17] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: A python package for analysis of heavy-tailed distributions," *PLoS ONE*, vol. 9, no. 1, p. e85777, Jan 2014. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0085777

[18] C. Dangalchev, "Residual closeness in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 365, no. 2, pp. 556–564, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437105012768

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

APPENDIX

Rank	Hashtag
1	isis
2	syria
3	iraq
4	is
5	aleppo
6	islamicstate
7	breakingnews
8	assad
9	ypg
10	turkey
11	palmyra
12	mosul
13	usa
14	russia
15	amaqagency
16	breaking
17	saa
18	deirezzor
19	ramadi
20	fallujah
21	homs
22	iraqi
23	aamaq
24	us
25	khanasir
26	damascus
27	fsa
28	anbar
29	wilayatninawa
30	infographic
31	caliphate$_{news}$
32	raqqa
33	sinai
34	egypt
35	pkk
36	twitterkurds
37	saudi
38	kirkuk
39	baghdad
40	sirte
41	libya
42	rebels
43	sdf
44	website
45	cybernews
46	status
47	ankara
48	iran
49	latakia
50	syrian
51	hama
52	photoreport
53	daraa
54	idlib
55	qalamoun
56	israel
57	geneva
58	hamas
59	gaza
60	hezbollah
61	wilayathalab
62	shaddadi
63	jordan
64	islamic$_{state}$

TABLE V: Ranked hashtags by Degree Centrality when k = 15

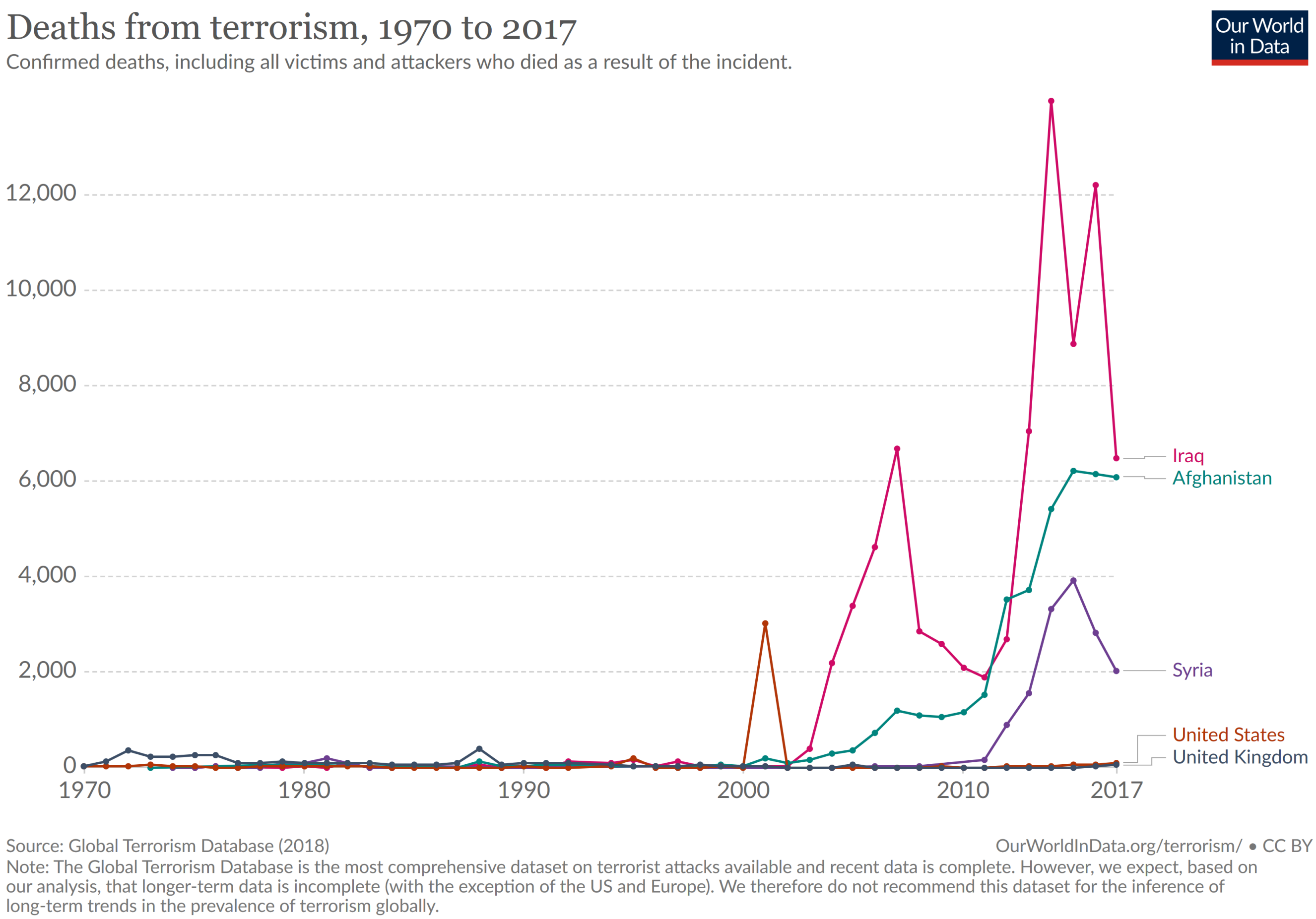

Fig. 14: Deaths from terrorism, 1970–2017.

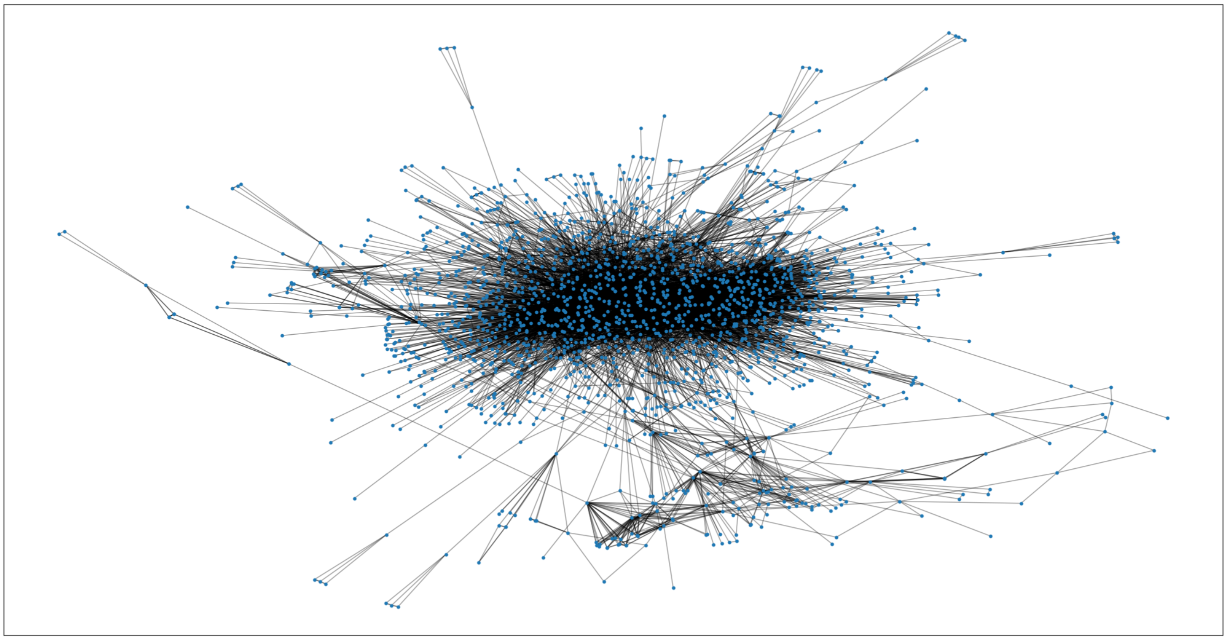

Fig. 15: General view of the hashtag graph. One giant commumity.

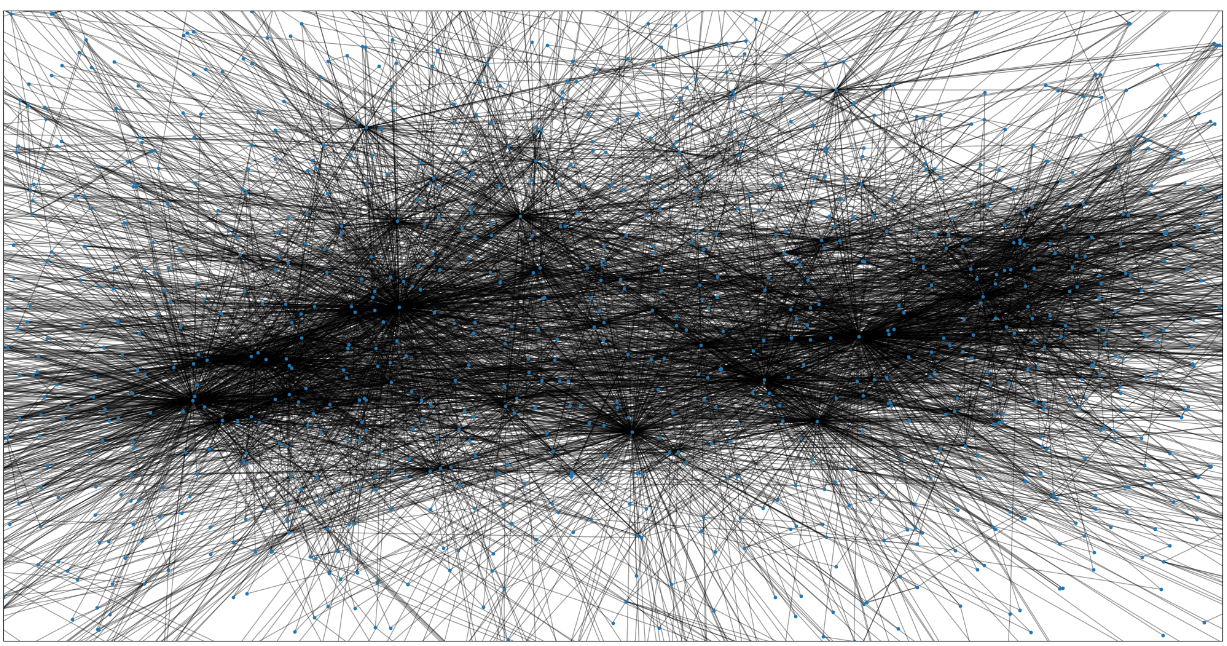

Fig. 16: Zoomed view to the giant community reveals more communities.

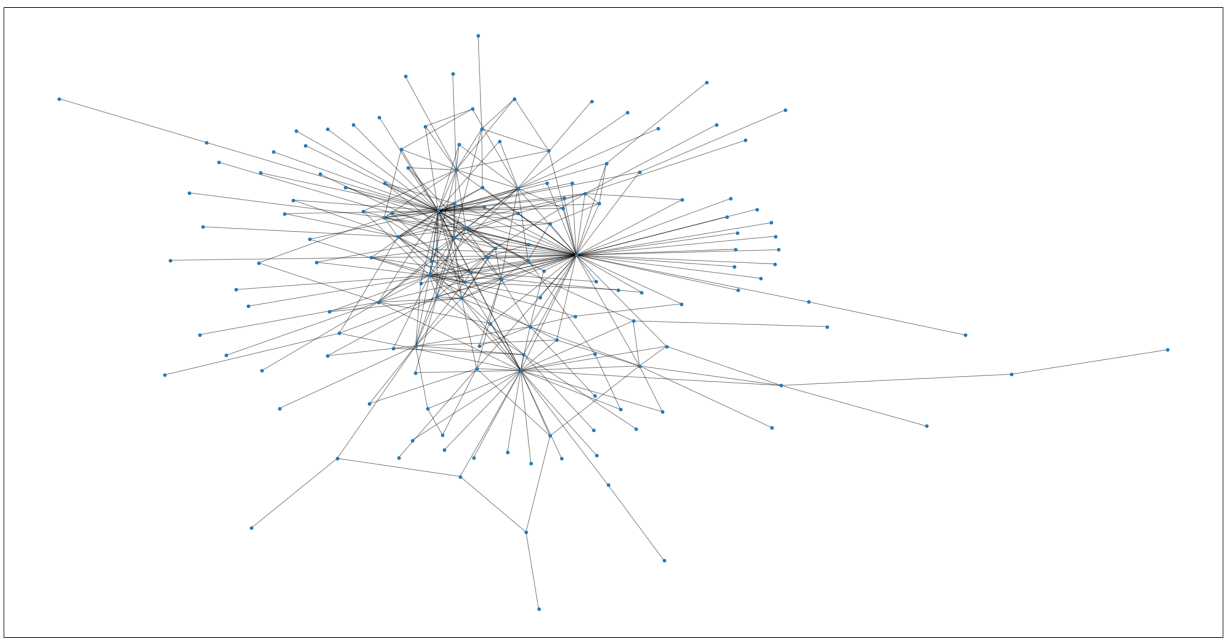

Fig. 17: General view of the hashtag graph. k = 2