Twitter Multilayer Networks of Co–Occurrence: Comparing Communities on Hashtags, Tweeple and Topic Modeling Terms

Moses A. Boudourides[1]

Robert K. Merton Visiting Research Fellow 2019, IAS, LiU
Northwestern University School of Professional Studies

[1] Moses.Boudourides@gmail.com

With Sergios Lenis

LäsIT Seminar

Department of Information Technology

Uppsala University, Sweden

March 25, 2019

The Problem at a Glance

- The aim here is **not** to analyze Twitter data and interpret the findings.
- But the plan is:
  - Starting from **any** JSON data harvested from the Twitter API:
    - to extract for each tweet **only** four objects: *authors*, *hashtags*, *user mentions* and the *text* of messages;
    - to form a Twitter dataframe with tweets as rows and with two columns (categorical variables): **hashtags** and **tweeple** (who are the authors of tweets and the user mentions in tweets);
    - to do a LDA Topic Model of the corpus of all text messages (after removing RTs) in order to discover an *arbitrarily selected* number of (*uninterpreted*) Topics of terms;
    - to append the **Topic Modeling terms** of each tweet (row) as a third column (categorical variable) to the Twitter dataframe.

The Analyses at a Glance

- The aim here is **not** on multilayer networks and communities.
- But the plan is:
  - To extract a multilayer network of **co–occurrences** among hashtags, tweeple and terms in order to:
    - to describe the overall network of co–occurrences and the three layers of subgraphs of co–occurrences separately (together with the inter-layer multipartite subgraph of co–occurrences);
    - to compute the (modularity–maximizing) **communities** of the overall network of co–occurrences and of each layer (as a subgraph of the former);
    - to compute **projected communities** on each layer induced by the restriction on layers of the communities of the overall network of co–occurrences;
    - to compute **reflected communities** from the communities of each layer on any other layer induced by the restriction on every pair of dual layers of the edges (co–occurrences) of the biparite graph among these two layers;
    - to detect similarities among communities, projected communities and reflected communities by computing the *Jaccard indices* among members of all these communities.

The Purpose at a Glance, I

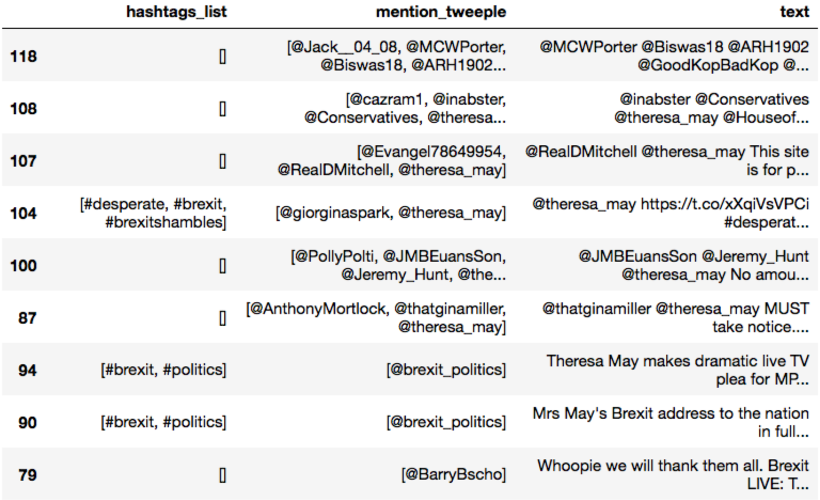
- Nowadays, Twitter data can be easily mined or even be downloaded from publicly available databases.
- At the same time, various types of networks extracted from Twitter data are discussed, qualitatively interpreted, formally analyzed (from the angles of network science or statistics) and displayed in elaborate complex visualizations (due to the typically large size of these networks).
- Of course, the purpose of all these data analyses (including text analyses) and network analyses is to contribute to the understanding of the underlying social (micro–)interactions taking place in social media and often resulting a variety of emerging (macro–)patterns of social behavior (like processes of diffusion, contagion, influence, group polarization, herd behavior, viral phenomena, information cascades, tipping points etc.).

The Purpose at a Glance, II

- Since a typical Twitter data is an example of Big Data, one way to tame the inherent excessive complexity of Twitter networks is by compressing them into smaller parts (groups, clusters or communities or etc.), which would be expected to carry over the characteristic features of the overall uncontrollable complex assemblages.
- However, as Twitter networks pinpoint a richness of multi-criteria and multi-scale effects, studying these networks as multilayer networks and understanding the variety of options for a multilayer decomposition into smaller constituents (communities here) can be a methodologically useful analytical/computational tool.
- The latter is the direction towards which we are moving here, as it is exemplified on a small size case study.

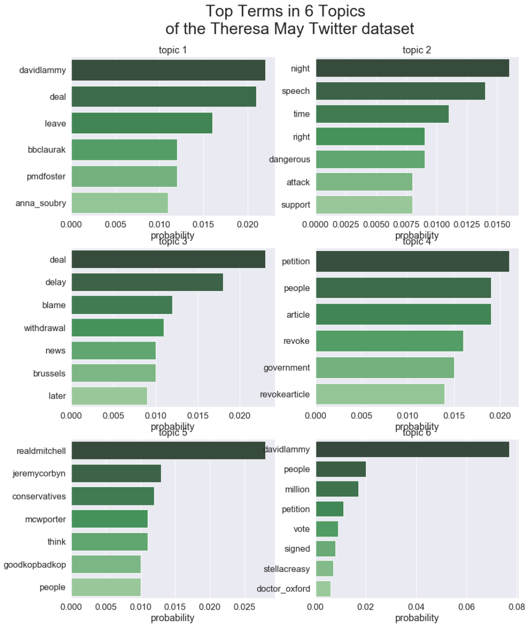
A Small Size Case Study:
The Theresa May Twitter Dataset

- 17821 tweets were harvested from the Twitter API on the afternoon of March 21, 2018, under the query "Theresa May" (during 4 minutes of mining).
- The retrieved tweets had circulated from Wed Mar 20 17:50:04 +0000 2019 to Thu Mar 21 12:15:27 +0000 2019.
- After removing all the included RTs, the remaining dataset was composed of 4608 tweets and the original dataframe looked like the next display:

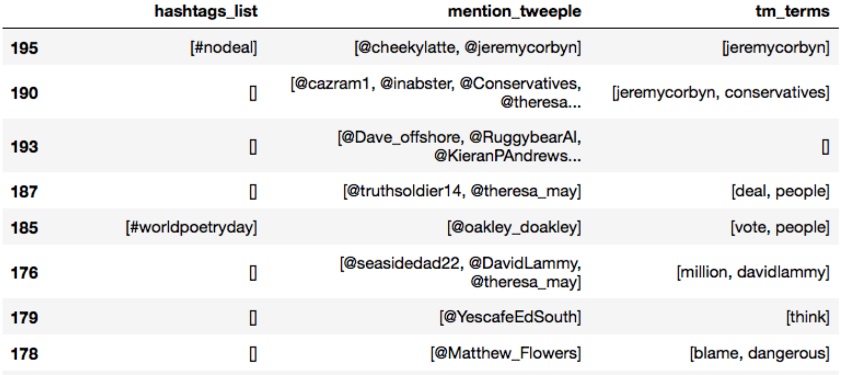


	hashtags_list	mention_tweeple	text
118	[]	[@Jack__04_08, @MCWPorter, @Biswas18, @ARH1902...	@MCWPorter @Biswas18 @ARH1902 @GoodKopBadKop @...
108	[]	[@cazram1, @inabster, @Conservatives, @theresa...	@inabster @Conservatives @theresa_may @Houseof...
107	[]	[@Evangel78649954, @RealDMitchell, @theresa_may]	@RealDMitchell @theresa_may This site is for p...
104	[#desperate, #brexit, #brexitshambles]	[@giorginaspark, @theresa_may]	@theresa_may https://t.co/xXqiVsVPCi #desperat...
100	[]	[@PollyPolti, @JMBEuansSon, @Jeremy_Hunt, @the...	@JMBEuansSon @Jeremy_Hunt @theresa_may No amou...
87	[]	[@AnthonyMortlock, @thatginamiller, @theresa_may]	@thatginamiller @theresa_may MUST take notice....
94	[#brexit, #politics]	[@brexit_politics]	Theresa May makes dramatic live TV plea for MP...
90	[#brexit, #politics]	[@brexit_politics]	Mrs May's Brexit address to the nation in full...
79	[]	[@BarryBscho]	Whoopie we will thank them all. Brexit LIVE: T...

LDA Topic Modeling of the Text of Tweets



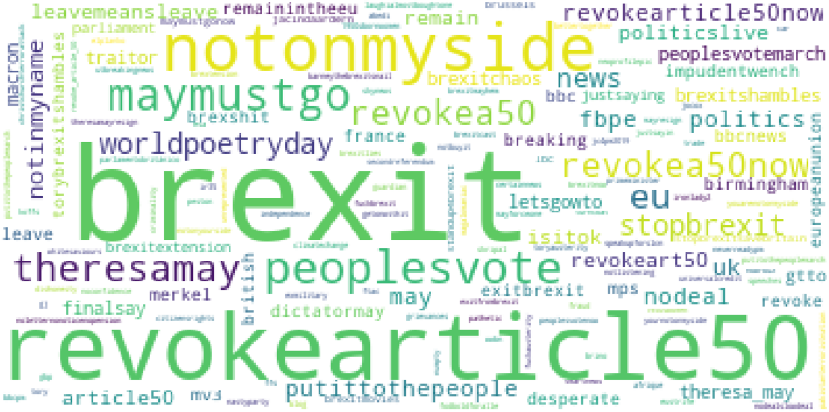
The Dataframe of the Theresa May Twitter Dataset



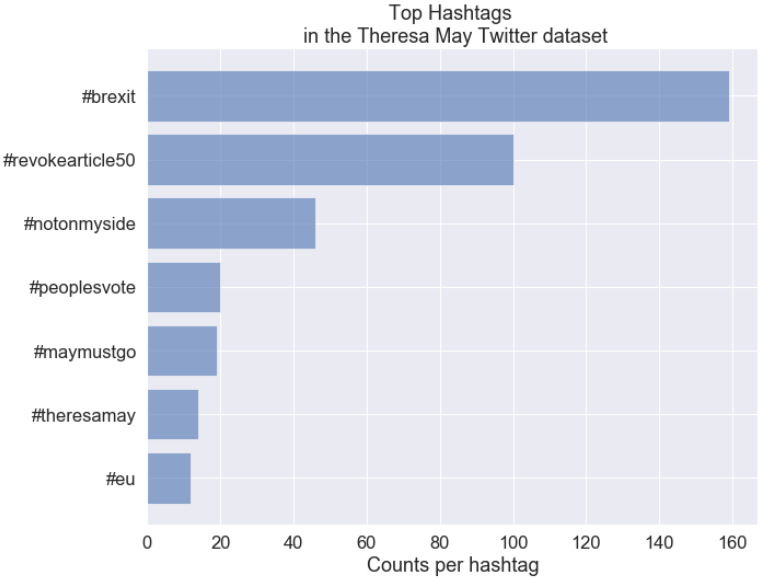
	hashtags_list	mention_tweeple	tm_terms
195	[#nodeal]	[@cheekylatte, @jeremycorbyn]	[jeremycorbyn]
190	[]	[@cazram1, @inabster, @Conservatives, @theresa...	[jeremycorbyn, conservatives]
193	[]	[@Dave_offshore, @RuggybearAl, @KieranPAndrews...	[]
187	[]	[@truthsoldier14, @theresa_may]	[deal, people]
185	[#worldpoetryday]	[@oakley_doakley]	[vote, people]
176	[]	[@seasidedad22, @DavidLammy, @theresa_may]	[million, davidlammy]
179	[]	[@YescafeEdSouth]	[think]
178	[]	[@Matthew_Flowers]	[blame, dangerous]

Hashtags

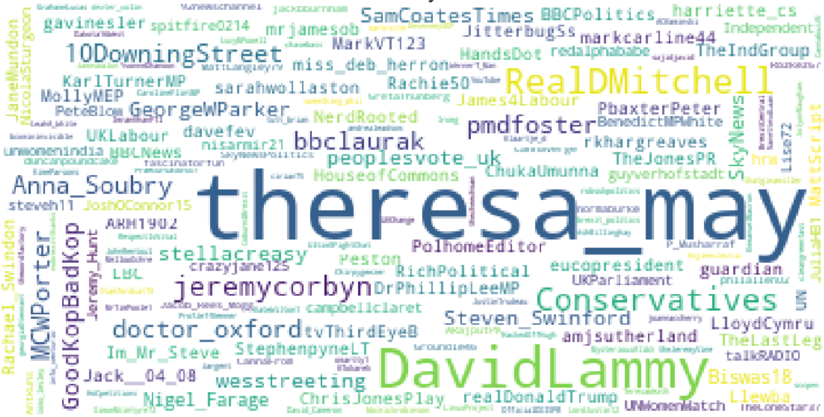
The wordcloud of Hashtags
in the Theresa May Twitter dataset



Top Hashtags

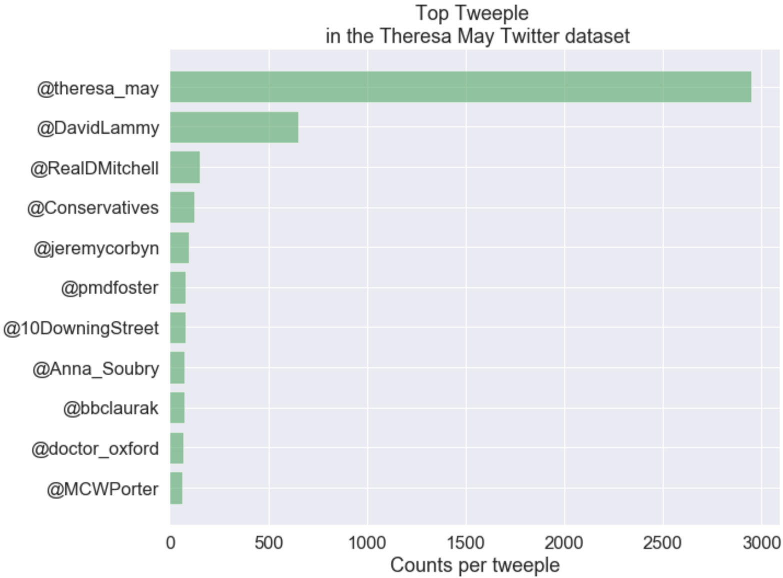


Tweeple



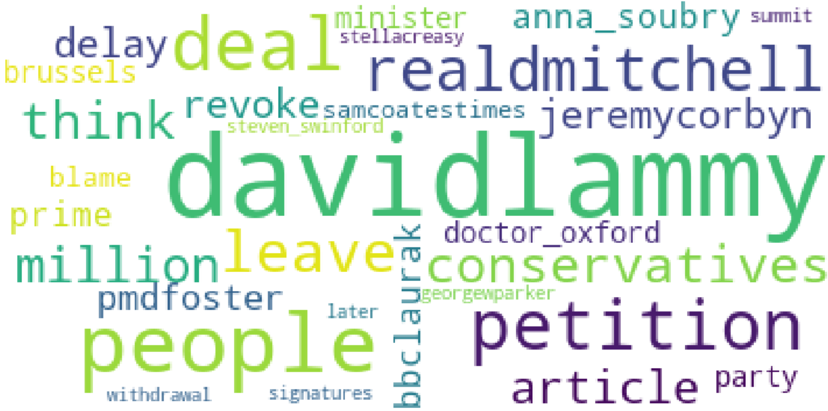
The wordcloud of Tweeple
in the Theresa May Twitter dataset

Top Tweeple

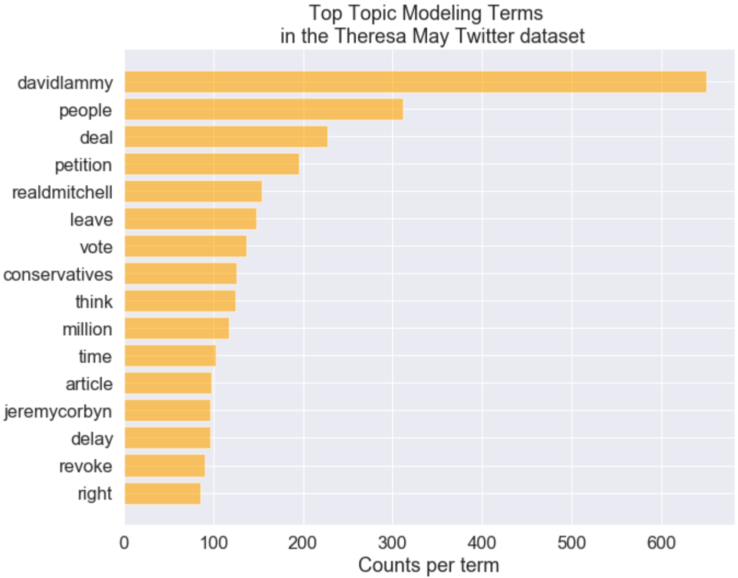

Top Tweeple
in the Theresa May Twitter dataset
@theresa_may
@DavidLammy
@RealDMitchell
@Conservatives
@jeremycorbyn
@pmdfoster
@10DowningStreet
@Anna_Soubry
@bbclaurak
@doctor_oxford
@MCWPorter
0
500
1000
1500
2000
2500
3000
Counts per tweeple

Terms

The wordcloud of Topic Modeling Terms
in the Theresa May Twitter dataset



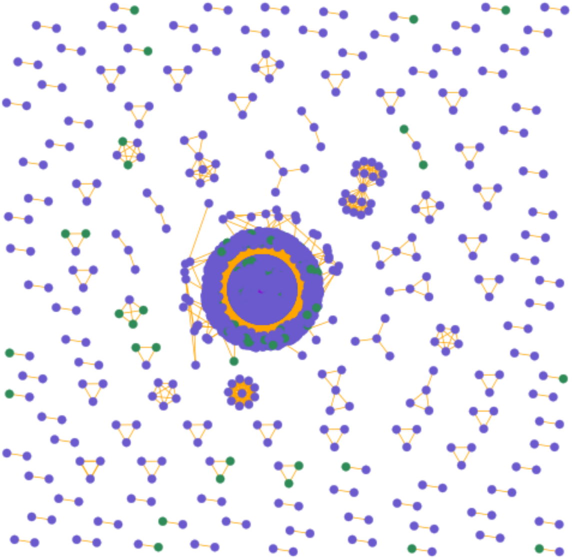
Top Terms


Top Topic Modeling Terms
in the Theresa May Twitter dataset
davidlammy
people
deal
petition
realdmitchell
leave
vote
conservatives
think
million
time
article
jeremycorbyn
delay
revoke
right
0
100
200
300
400
500
600
Counts per term

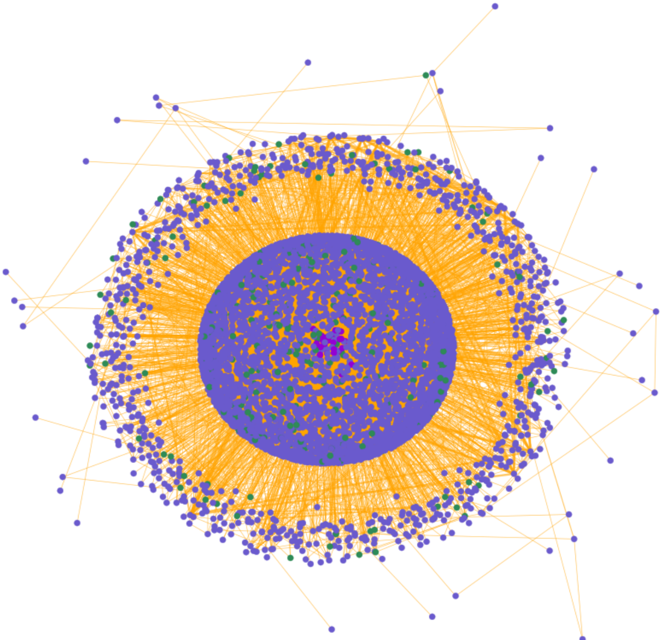
The Graph of All Co-Occurrences

- The co-occurrence graph has 4973 nodes and 36407 edges
- The nodes of the co-occurrence graph are:
  - 285 hashtags
  - 4653 tweeple
  - 35 terms
- This graph is not connected and has 119 connected components
- The largest connected component has:
  - 4656 nodes and 35951 edges
- The nodes of the largest connected component of the co-occurrence graph are:
  - 260 hashtags
  - 4361 tweeple
  - 35 terms

The co-occurrence graph
in the Theresa May Twitter dataset

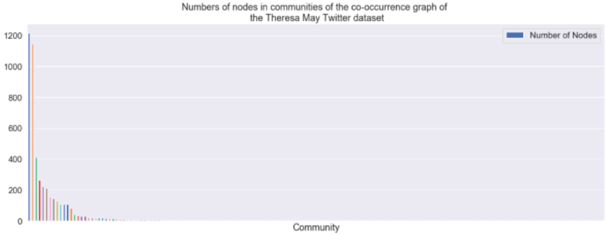


The largest connected component of the co-occurrence graph
in the Theresa May Twitter dataset



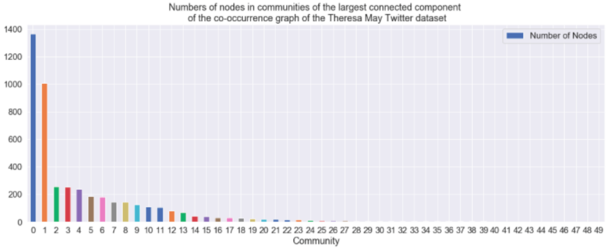
Communities of the Graph of All Co-Occurrences

```
The co-occurrence graph has 165 communities and modularity coefficient
equal to 0.5211
```



Numbers of nodes in communities of the co-occurrence graph of the Theresa May Twitter dataset

```
The largest connected component of the co-occurrence graph has 50 comm
unities and modularity coefficient equal to 0.5117
```

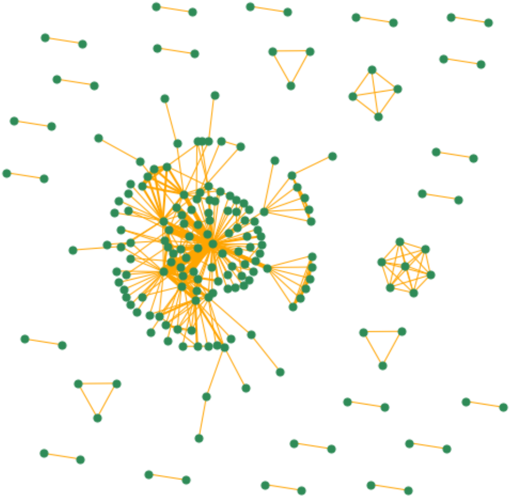


Numbers of nodes in communities of the largest connected component of the co-occurrence graph of the Theresa May Twitter dataset

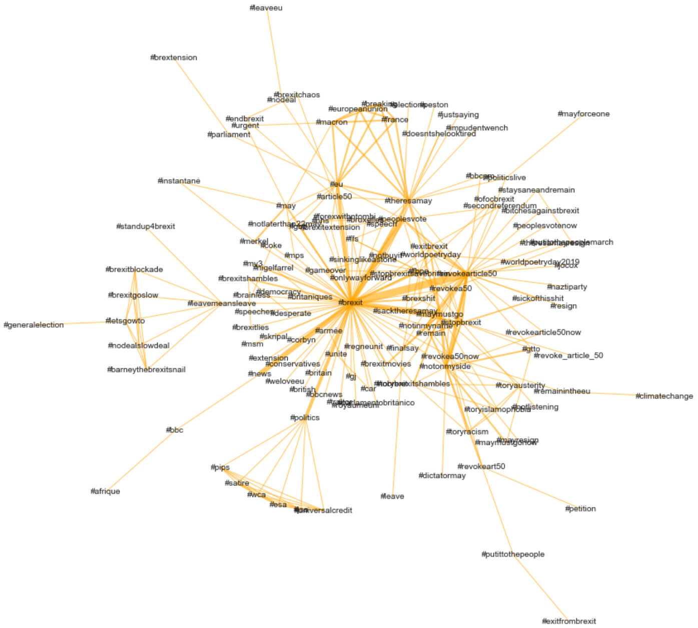
The Layer of Hashtags

- The hashtags layer has 285 nodes and 408 edges
- This graph is not connected and has 118 connected components
- The largest connected component has:
    - 132 nodes and 351 edges

The layer of co-occurring hashtags in the Theresa May Twitter dataset

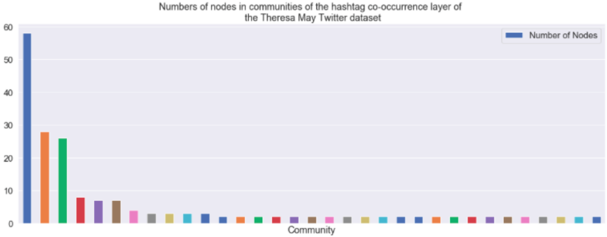


The largest connected component of the layer of co-occurring hastags in the Theresa May Twitter dataset

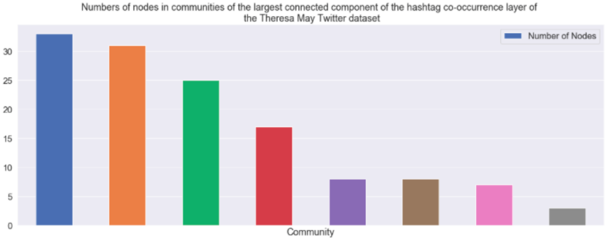


Communities of the Layer of Hashtags

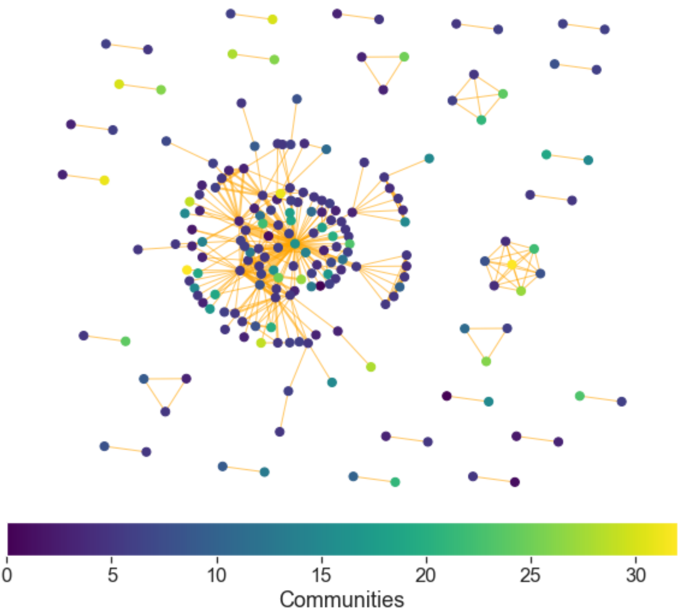
```
The hashtags co-occurrence layer has 33 communities and modularity coe
fficient equal to 0.6000
```



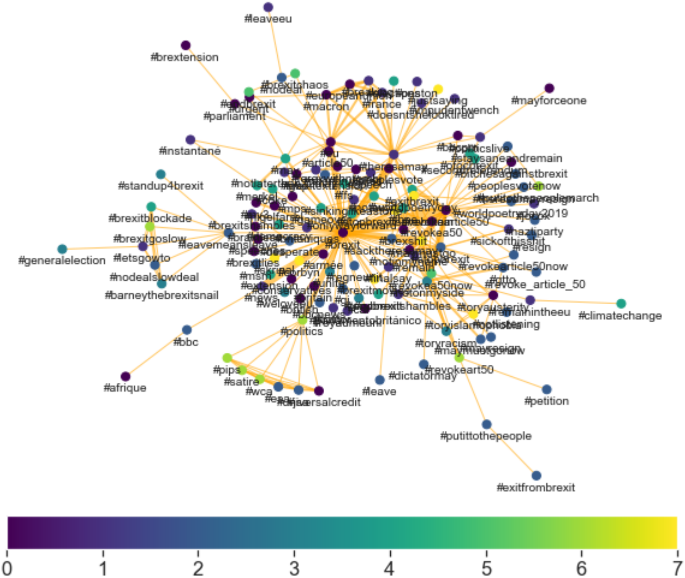
```
The largest connected component of the hashtags co-occurrence layer ha
s 8 communities and modularity coefficient equal to 0.5258
```



The 33 communities (with modularity = 0.6000) of the hashtag co-occurrence layer in the Theresa May Twitter dataset



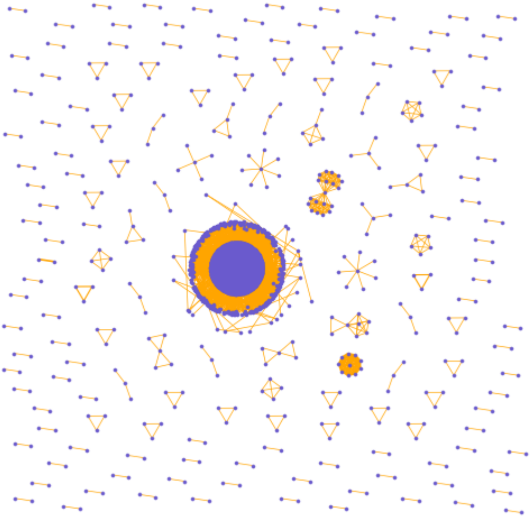
The 8 communities (with modularity = 0.5258) of the largest connected component of the hashtag co-occurrenc
the Theresa May Twitter dataset



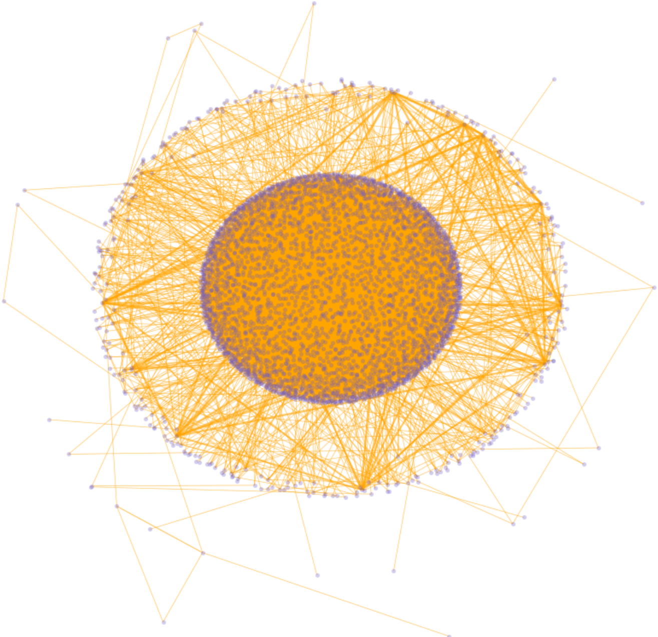
The Layer of Tweeple

- The tweeple layer has 4653 nodes and 20391 edges
- This graph is not connected and has 724 connected components
- The largest connected component has:
    - 3658 nodes and 19845 edges

The layer of co-occurring tweeple
in the Theresa May Twitter dataset

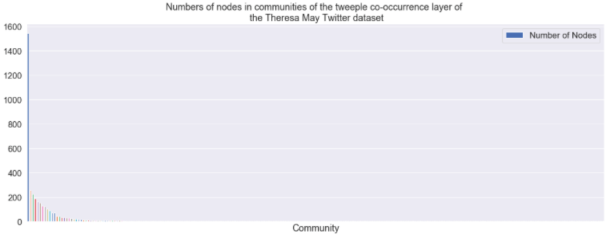


The largest connected component of the layer of co-occurring tweeple in the Theresa May Twitter dataset

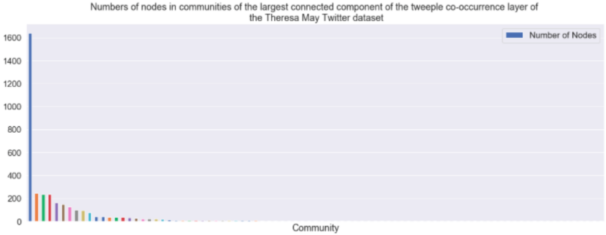


Communities of the Layer of Tweeple

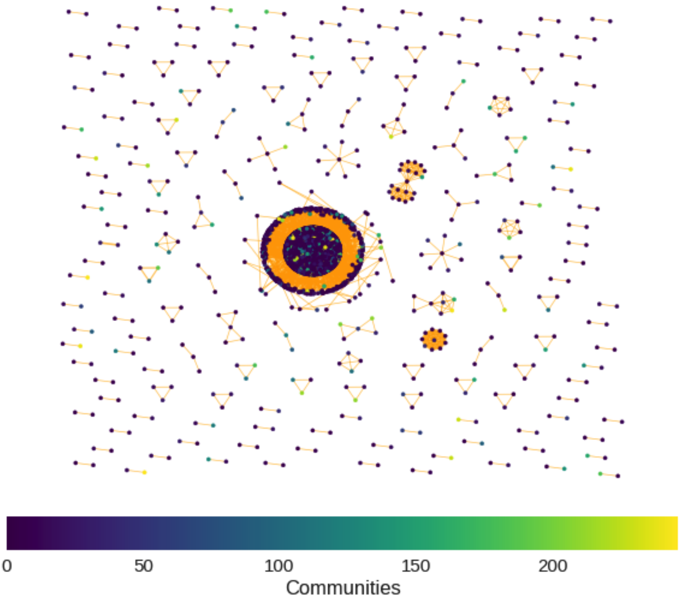
```
The tweeple co-occurrence layer has 240 communities and modularity coe
fficient equal to 0.5639
```



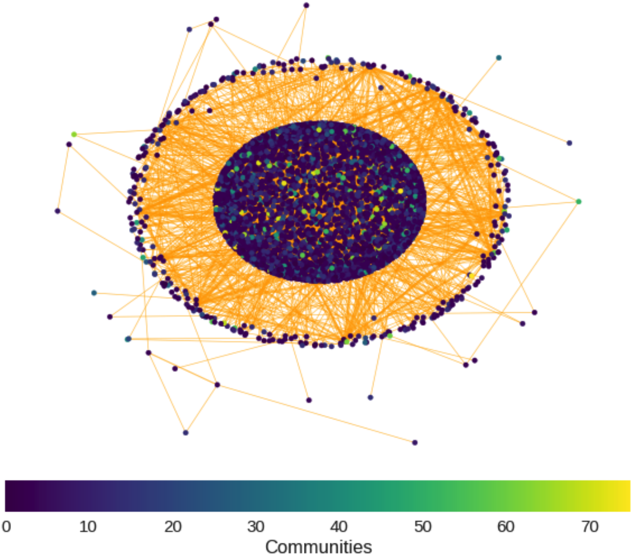
```
The largest connected component of the tweeple co-occurrence layer has
87 communities and modularity coefficient equal to 0.5443
```



The 240 communities (with modularity = 0.5623) of the tweeple co-occurrence layer in the Theresa May Twitter dataset



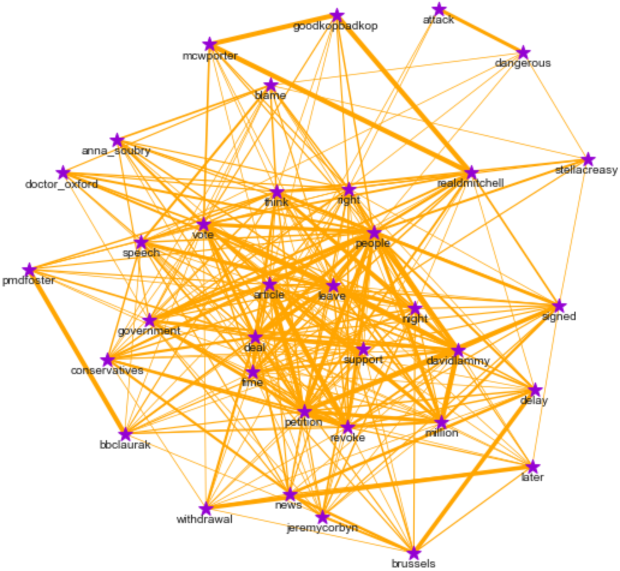
The 87 communities (with modularity = 0.5463) of the largest connected component of the tweeple co-occurrence layer
the Theresa May Twitter dataset



The Layer of Terms

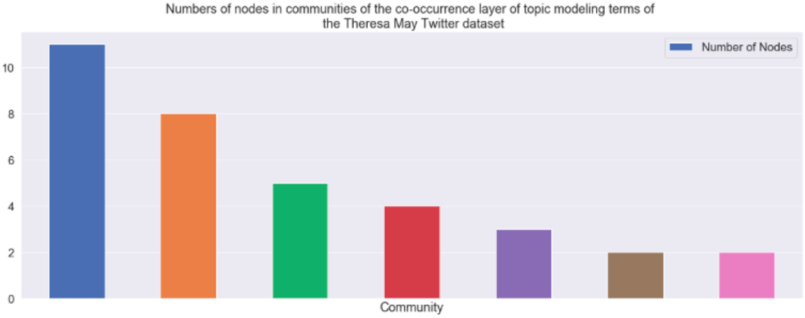
- The topic modeling terms layer has 35 nodes and 1703 edges
- This graph is connected

The layer of co-occurring topic modeling terms
in the Theresa May Twitter dataset

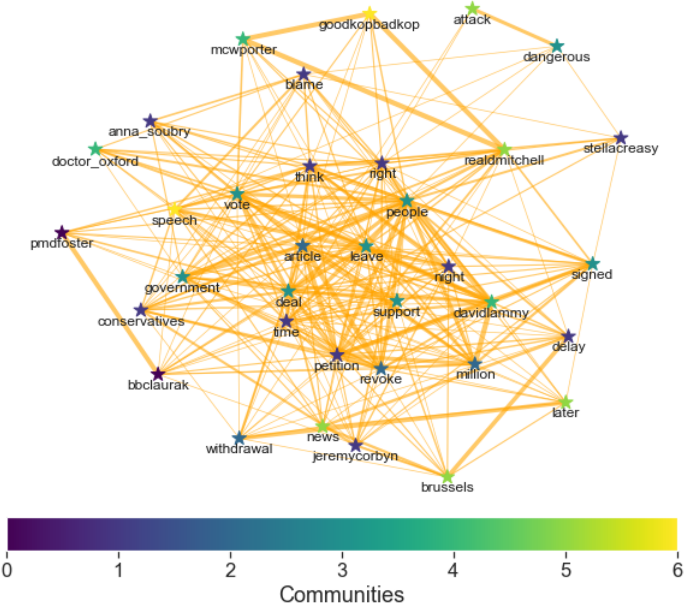


Communities of the Layer of Terms

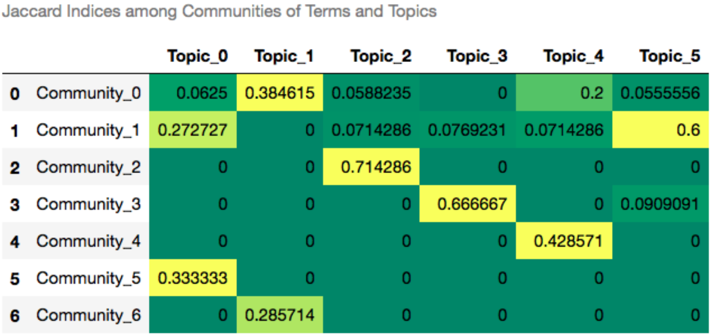
```
The co-occurrence layer of topic modeling terms has 7 communities and
modularity coefficient equal to 0.3543
```



The 7 communities (with modularity = 0.3543) of the co-occurrence layer of topic modeling terms in the Theresa May Twitter dataset



Comparing Topics with Communities on the Layer of Terms



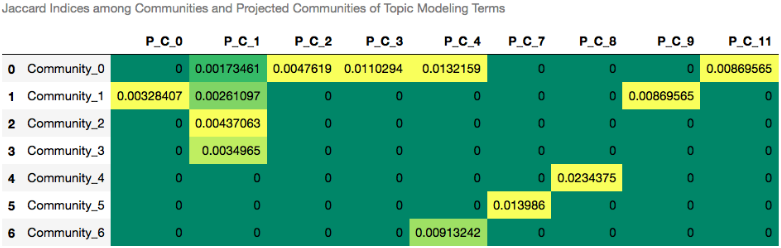
Jaccard Indices among Communities of Terms and Topics

		Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5
0	Community_0	0.0625	0.384615	0.0588235	0	0.2	0.0555556
1	Community_1	0.272727	0	0.0714286	0.0769231	0.0714286	0.6
2	Community_2	0	0	0.714286	0	0	0
3	Community_3	0	0	0	0.666667	0	0.0909091
4	Community_4	0	0	0	0	0.428571	0
5	Community_5	0.333333	0	0	0	0	0
6	Community_6	0	0.285714	0	0	0	0

- The maximum Jaccard index is 0.7143 for Community-2 and Topic-2

Comparing Communities of Layers with Projections on Layers of the Communities of the Co-Occurrence Graph

- On Hashtags the maximum Jaccard index is 0.7500
- On Tweeple the maximum Jaccard index is 1



Jaccard Indices among Communities and Projected Communities of Topic Modeling Terms

		P_C_0	P_C_1	P_C_2	P_C_3	P_C_4	P_C_7	P_C_8	P_C_9	P_C_11
0	Community_0	0	0.00173461	0.0047619	0.0110294	0.0132159	0	0	0	0.00869565
1	Community_1	0.00328407	0.00261097	0	0	0	0	0	0.00869565	0
2	Community_2	0	0.00437063	0	0	0	0	0	0	0
3	Community_3	0	0.0034965	0	0	0	0	0	0	0
4	Community_4	0	0	0	0	0	0	0.0234375	0	0
5	Community_5	0	0	0	0	0	0.013986	0	0	0
6	Community_6	0	0	0	0	0.00913242	0	0	0	0

- On Terms the maximum Jaccard index is 0.0234

Comparing Communities of Layers with Reflections on Layers of the Communities of the Dual Layers

- Reflections of Communities on Hashtags:
  - Reflection of Communities of Tweeple on Hashtags
    - The maximum Jaccard index is 0.1696
  - Reflection of Communities of Terms on Hashtags
    - The maximum Jaccard index is 0.1720

- Reflections of Communities on Tweeple:
    - Reflection of Communities of Hashtags on Tweeple
        - The maximum Jaccard index is 0.3206
    - Reflection of Communities of Terms on Tweeple
        - The maximum Jaccard index is 0.2978

Comparing Communities of Layers with Reflections on Layers of the Communities of the Dual Layers

- Reflections of Communities on Terms:
    - Reflection of Communities of Hashtags on Terms
        - The maximum Jaccard index is 0.4
    - Reflection of Communities of Tweeple on Terms
        - The maximum Jaccard index is 0.3143