Uppsala 2019-08-30

The Polarised State of the Swedish Political Twitterverse

Summer Math Camp 2019

Department of Mathematics, Uppsala University, Box 480, 751 06 UPPSALA, Sweden

Participants: Agnes Davíðsdóttir, Magdalena Fischerström, Claes Fälth, Johannes Graner, Andreas Lindgren, Amela Mehic and Albert Nilsson
Supervisors: Raazesh Sainudiin and Tilo Wiklund.

Abstract: An analysis of over 91 million tweets collected during an 8-month period around the Swedish general election in 2018 showed strong evidence of highly polarised communities. These communities were found to differ politically and in their nearly mutually exclusive use of hashtags, links to URLs and preference for news sources. This preliminary study of the Swedish Twitterverse in conjunction with statistical tests is indicative of highly polarised echo-chambers across the left and right political spectrums.

Summary: During Summer Math Camp 2019, the Swedish twitterverse was analysed to try to discern ideological groupings and investigate their statistical behaviours. The data used consisted of 91 million tweets collected during an 8-month period around the Swedish general election in 2018. The data was collected by Raazesh Sainudiin under *Project MEP (Meme Evolution Programme): Sverige*[1] by focusing on a list of political hashtags, a list of politically influential and ideologically diverse twitter accounts, including all the Swedish MPs as well as tweets containing any one of the 400 most common Swedish words. The experiment was jointly designed by Mattias Gardell, Simon Lindgren, Emin Poljarevic and two anonymous Citizen Scientists.

Initially the data had to be cleaned to remove non-Swedish and spam-like tweets that were collected due to the inclusive experimental design. To remove the non-Swedish tweets, a language filter was implemented. This filter calculated the letter frequency in all the tweets of each twitter account and compared that to the Swedish language letter frequency.[2] If the letter frequency of an account's collection of tweets were too far from the Swedish letter frequency, all tweets from that account were removed. The threshold for the distance to Swedish was chosen so that 95% of Swedish text would be accepted.[3] The filter for spam-like behaviour was constructed so that if a user retweeted another user more than 1000 times during the 8-month period, equal to over 4 times a day, that user was considered a spam account and was removed.[4] The effects of the filter was manually checked by examining the twitter user accounts at the thresholds of the filter and for random samples that passed the filter.

When the data was cleaned it was made into a graph or network with twitter accounts as vertices and their retweets as directed edges from the account that made the original tweet to the account that retweeted. An unsupervised machine learning algorithm called label propagation was used on the network to divide the users into different clusters.[5]

---

[1] Project MEP was supported by cloud computing support for academic research from AWS and databricks and on-premise computing support by Combient AB.
https://lamastex.github.io/scalable-data-science/sds/research/mep/
[2] Notebook: languageDistanceParquet
[3] Notebook: languageSampling
[4] Notebook: CleanData
[5] Notebook: data

A well-established fact that laid the foundation for a lot of our analysis is that a retweet is a clear signal of concurrence of the retweeter with the tweeter of the original tweet. Another common type of interaction between twitter accounts is via reply-tweets, whereby one account replies to another. Unlike retweets, reply-tweets may not necessarily be in agreement and in politically engaged debates they can be due to disagreements. To test the hypothesis that there is no difference between clusters of accounts formed due to purely retweet interactions versus those formed due to purely reply-tweet interactions, a network was made with 90% of randomly sampled retweets as its edges. The label propagation algorithm was then used to partition the users into clusters. When looking at the remaining 10% of the retweets it was found that 67% of them were in between users within the same cluster, while 45% of all reply-tweets from the data were between users in the same cluster.
Doing the same test on a network created with the 90% of the reply-tweets from the data, we found that 1.5% of all retweets from the data were between users in the same cluster, and 10% of the remaining 10% of reply-tweets were between users in the same cluster.
This shows that clustering based on a retweet network which yields a community structure reflecting agreement or concurrence is significantly different from that than based on reply-tweets.[6] This is not unnatural as reply-tweet are known to include interactions of disagreement across members who belong to different retweet-based clusters of agreement.

The clustering done by label propagation on the retweet network yielded two larger clusters and a few smaller ones. An analysis of the clustering was done to find possible characteristics of each cluster. This was of great interest since the label propagation algorithm is unsupervised and thus the results of the clustering had to be analysed afterwards.

Limiting the analysis to the ten largest clusters given by label propagation, we found that a further subset of three clusters were primarily composed of Swedish accounts. These three clusters, being the first, second and fourth largest clusters, were therefore the only ones of interest for further analysis. Together, the three clusters consisted of approximately 200,000 twitter accounts. The other, non-Swedish, clusters seemed to be grouped based on their love for K-pop or other personalistic self-indulgent interests of a non-political nature.

When looking at how different members of the Swedish parliament were distributed among the clusters it turned out that the algorithm placed most parliament members from the same party in the same cluster as one would expect more agreement between members of the same party or members of different but ideologically aligned parties. The members of the parties SD, KD, M, and L were found in the same cluster, the largest one, whereas C, MP, S and V were found in another cluster, the second largest. This corresponds well to how the parties are placed on the left-right political spectrum. The cluster with SD, KD, M and L is referred to as the right cluster, and the cluster with C, MP, S and V is referred to as the left cluster, while the third cluster was determined to be non-political with a focus in sports. It is worth noting that both L and C had a fairly large proportion of members placed in the other political cluster than its majority.[7]

Further, the extent to which different news sources were consumed within each of the clusters was investigated. By looking at what newspapers were shared in the retweet interactions, we found a significant difference between the clusters. The right cluster shared news from two right wing alternative newspapers called Samhällsnytt and Fria Tider, while the left cluster shared Dagens

[6] Notebook: Stochastic Block Model
[7] Notebook: CleanExplorationOfLabProp

Nyheter a lot more than the right cluster.[8] This indicates that the left and right clusters, being informed by different sources of news, are in different *news echo-chambers*.

When looking at how the different clusters used hashtags and URLs, it turned out that the right cluster used hashtags a lot more than the left cluster. Some common hashtags in the left cluster were *svpol* (Swedish politics)*, svtnyheter* (Swedish public service news)*, klimat* (climate) and *metoo,* whereas the right cluster used hashtags *svpol, migpol* (migration politics)*,* and *SD2018.*[9] The third cluster seemed to be more focused around sport with the largest use of hashtags like *worldcup, leotrean* (a hashtag connected to a betting site), and *val2018* (election 2018).[10] The most tweeted URLs were mainly newspapers, and the right clusters shared more news than the left cluster.[11] The distribution of hashtags and URLs were significantly different between the clusters, thus pointing to distinction of issues and concerns.

To find additional divisions within each cluster it was of interest to look at the shortest paths between users. This was done by computing the shortest weighted path[12] from all users to a set of key users. The key users, referred to as landmarks, were selected based on their political affiliation, influence and twitter activity. Around 70 landmarks were chosen, representing a wide political spread, along with a set of newspapers and media outlets. From the shortest path-distances it was possible to compute the percentage[13] of the population with similar distances to a specific set of landmarks. The result seemed to indicate that a large part of the Swedish twitter users are not very politically engaged.

Further, the following landmarks were chosen; *Hanif Bali* (M)*, Katerina Janouch* (right wing influential)*, Paula Bieler* (SD)*, Jonas Sjöstedt* (V)*, Isabella Lövin* (MP)*, Jeff Ahl* (AfS) and the account *Nazispotting* (far left) as starting points for the algorithm bisecting k-means[14]. The algorithm divided the three largest clusters into 10 sub clusters. Further research is required to find meaningful structure behind the formation of the smaller clusters.

An attempt to analyse the content of each tweet was done by looking at the tweet text and identifying specific positive and negative words from a Swedish lexicon. A sentiment score was given to each tweet but the method was inconclusive perhaps due the brevity of tweets.

In conclusion, it can be said that unsupervised machine learning algorithms for forming clusters based on a retweet network yields interpretable community structures and can be useful in understanding the extent to which various politically ideological communities are structured. For the Swedish twitterverse the clustering was subject to further analysis and the clusters were found to differ politically as well as in their use of hashtags, URLs and their choice of news sources. This preliminary study of the data in conjunction with statistical tests is indicative of fairly polarised echo-chambers across the left and right political spectrums.[15]

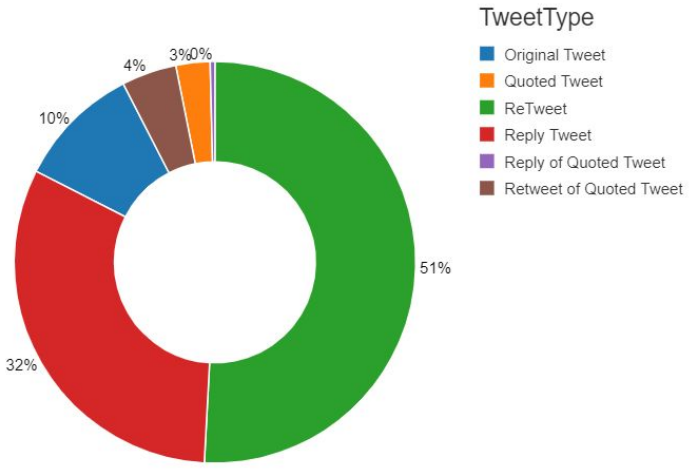
[8] Notebook: CleanExplorationOfLabProp
[9] Notebook: CommunityAndHashtags
[10] Notebook: Hashtag-Bootstrapping
[11] Notebook: CommunityAndURLs
[12] Notebook: SWP1_labelprop
[13] Notebook: SWP2_analysis
[14] Notebook: CleanBisectingKMeans
[15] Slides of the presentation made to the Department of Mathematics is attached in the following pages.

Summer Math Camp 2019

Analysing the Swedish Twitter network
around the 2018 general election

Agnes Davíðsdóttir
Albert Nilsson
Amela Mehic
Andreas Lindgren
Claes Fälth
Johannes Graner
Magdalena Fischerström

Twitter interactions



Aron Flam Retweetade

Peter Sellei
@PeterSellei

Antar att ni då fryser bistånd till Palestina tills öppna och demokratiska val sker samt att mänskliga rättigheter respekteras?
Ordföranden i ert systerparti är inne på sitt 15:e år som president efter att ha blivit vald för en 4-årig mandatperiod.

Margot Wallström @margotwallstrom · 4h
Vår demokratioffensiv innebär också mer av vår feministiska utrikespolitik. I den rådande globala polariseringen är det viktigare än någonsin att Sverige fortsätter vara en stark röst för jämställdhet och åtnjutande av mänskliga rättigheter för alla. #Priosutrikespol

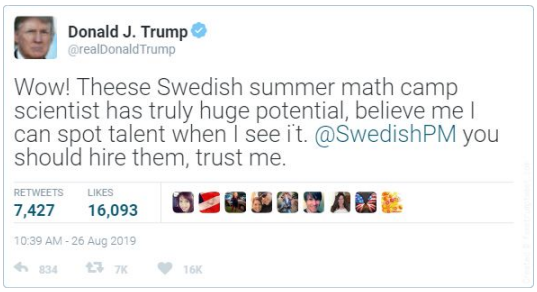
10:22 fm · 28 aug. 2019 · Twitter for iPhone

45 Retweets **197** gilla-markeringar

Eric Danell @eric_danell · 3h
Svarar @PeterSellei och @konsensuseliten
Finns det mer hyckleri än hos regeringspartierna med kringfjäskande C och L samt V. Varför ställer inte C och L krav på regeringen bl.a. om biståndet till Palestina. Låt inte regeringen härja fritt inom utrikespolitiken! C och L, sluta fjäska för Löfven o co! Visa lite kurage!

1

Data



- Meme evolution programme Sverige, Raazesh Sainudiin
- Mattias Gardell, field ethnography, department of theology Uppsala
- Simon Lindgren, digital sociology, department of sociology Umeå

- 91 million tweets

What data do we want?

VAL 2018

- Structure of the Swedish Twitter network around the 2018 general election

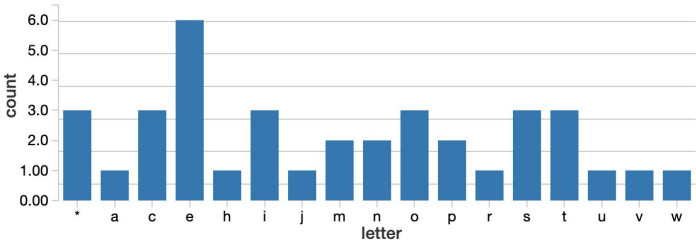
- Swedish Twitter users
- Not bots or other spam

Cleaning Data

- Language filter
  - exclude non-Swedish Twitter users
- High frequency filter on retweets

Language Filter

- Every language has a unique letter frequency
- Swedish letter frequency compared to user letter frequency



- Remove punctuations, emojis, white space, etc ✨🎉🧶조선글
- Removepunctuationsemojiswhite spaceetc조선글
- non-Swedish letter is mapped to *

Tolerance of Distance to Swedish

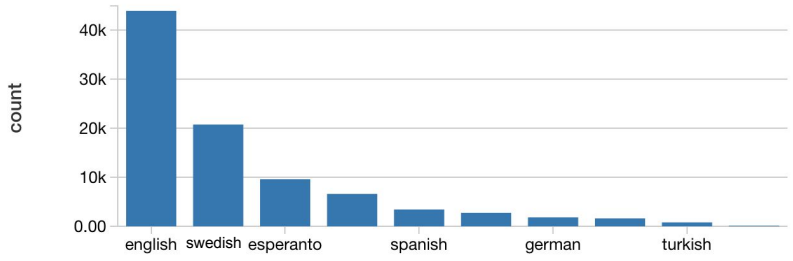
- A “Swedish word” in our sense: ajhbksdnmäaösfmsakfpåslfafebgh mtk
- Result: Tolerance 5% gave max distance 0.485

Distributions

- Swedish letter frequency
- Users tweet length

Language filter

- We made sure that language was closer to Swedish than English
- Result: Reduced data from 91 to 29 million data points



Sample of the language distribution before cleaning

UPPSALA UNIVERSITET

High Frequency Filter

- Frequent retweets
  - Anyone retweeting the same person more often than 1000 times over the 8 months was taken out of the data (more than 4 times a day on average)

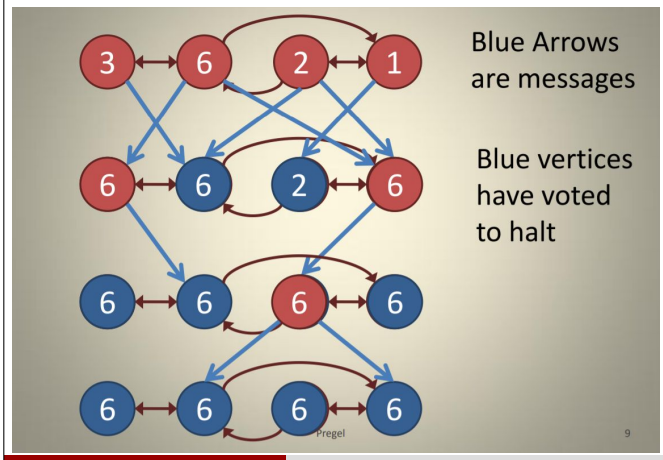
UPPSALA UNIVERSITET

Distributed Vertex Programming

- With the clean data, we want to group similar users for analysis
  - This is easily done if we have a graph
- Problem: Computing things on large graph
  - Solution: Only pass messages between neighbours!

Distributed Vertex Programming

Example. Finding largest state in strongly connected graph:



- Every vertex begins with an *initial state*
- Vertices *send a message* to their neighbours
- Each vertex *updates* its state based on incoming messages
- A vertex can choose to halt, not participating in the next iteration
- This is called a Pregel program

Clustering

- We want to create clusters of retweeting users
- First we need a directed graph of retweet network
  - Vertices: Users and their unique ID
  - Edges: If A retweets B, include edge from B to A
  - We get a *directed, multi-edged, looped graph*

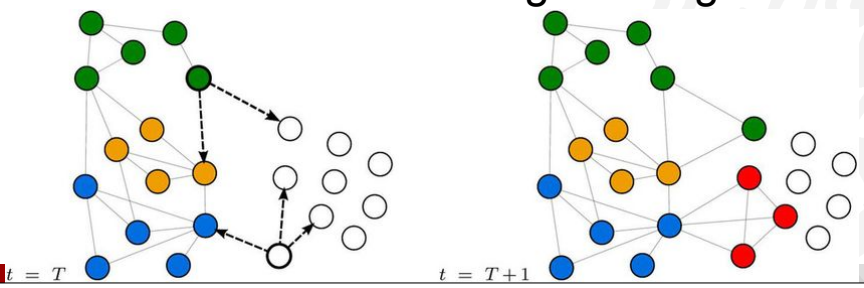
Clustering

component	size
528	1157438
159429982	1873
15261401	493
15005510	170
913991205159096320	157
53614956	130
73332929	123
69549657	86
319579913	85
40227371	84
25949039	82
148152741	77
94012801	77
460893708	71
20087934	71
46806220	70
15021968	62
161418081	59
35569691	57
954735119805292544	57

- We only consider users who appear in the largest connected component of the graph
  - Computing the connected components is done by a Pregel program
- 89.9% of users are in the largest component

Clustering

- We cluster users based on who they retweet
  - Pregel program implementing *label propagation*
    - Initial state: ID of the vertex
    - Sent Message: Current state
    - Update: Take mode of incoming messages

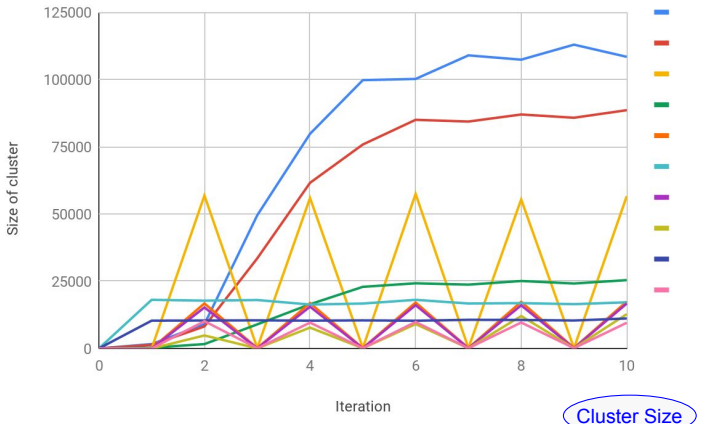


Clustering

- Pros
  - Relatively cheap
  - No information about graph necessary
- Cons
  - Convergence not guaranteed
  - Can put all vertices in the same cluster

Clustering

- After 10 iterations, the three Swedish clusters have settled



Swedish!

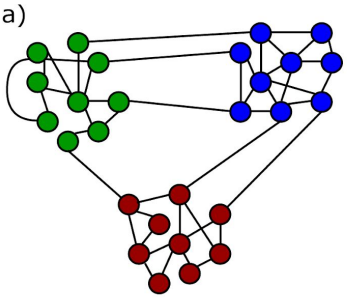
Cluster Size

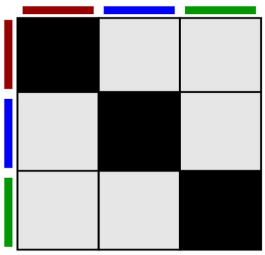
label	size1	size2	size3	size4	size5	size6	size7	size8	size9	size10
343663197	1641	9033	49519	79794	99747	100188	108949	107359	112934	108408
3048723709	1330	8197	33493	61579	75819	84996	84346	86978	85763	88576
3285105132	1	56869	167	55939	202	57465	144	55412	177	56632
434315852	281	1630	8957	16436	22913	24232	23726	25084	24152	25405
700434386160852992	1	16737	138	16874	149	17082	157	17339	194	17475
1283934055	18081	17769	18005	16330	16700	18076	16718	16855	16474	17152
832135844706148353	1	15242	55	15530	74	16092	83	16037	86	16708
22558580	1	4837	43	7756	78	9068	79	12067	95	12764
2982376457	10345	10398	10460	10340	10448	10319	10661	10597	10402	11149
2297996020	2	10143	75	9618	126	9857	100	9654	118	9562

Twitter Interactions Between Clusters

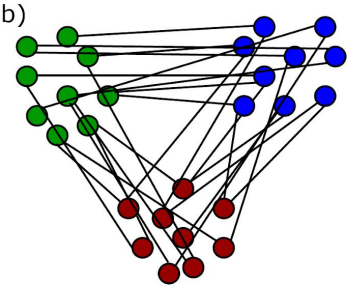
- Build network based on retweets or reply tweets
- Form clusters on the network with label propagation
- Test the clustering with unseen retweets and replies
- Stochastic block model
- Probability for edge, *p* within clusters, *q* between clusters

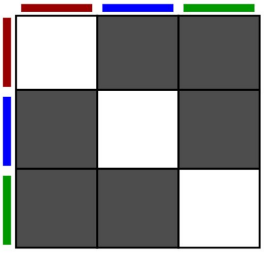
a)





b)





Twitter interactions between clusters

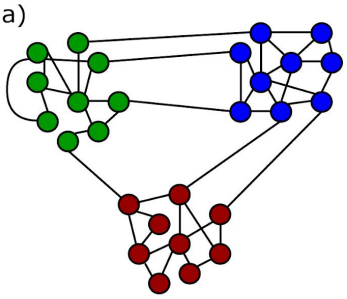
Retweet Network

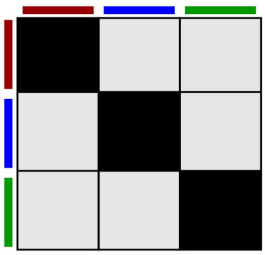
Tweet Type	Within	Between
Retweet	67.6%	32.4%
Reply Tweet	45.4%	54.6%
Random connection	0.38%	99.62%

Reply Network

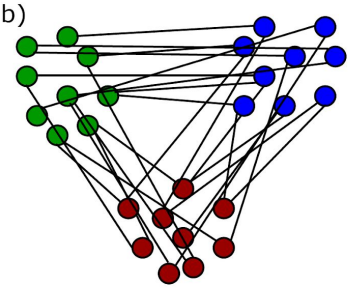
Tweet Type	Within	Between
Retweet	1.49%	98.51%
Reply Tweet	9.79%	90.21%
Random connection	0.00057%	99.99943%

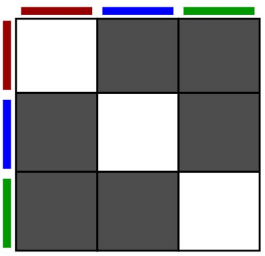
a)



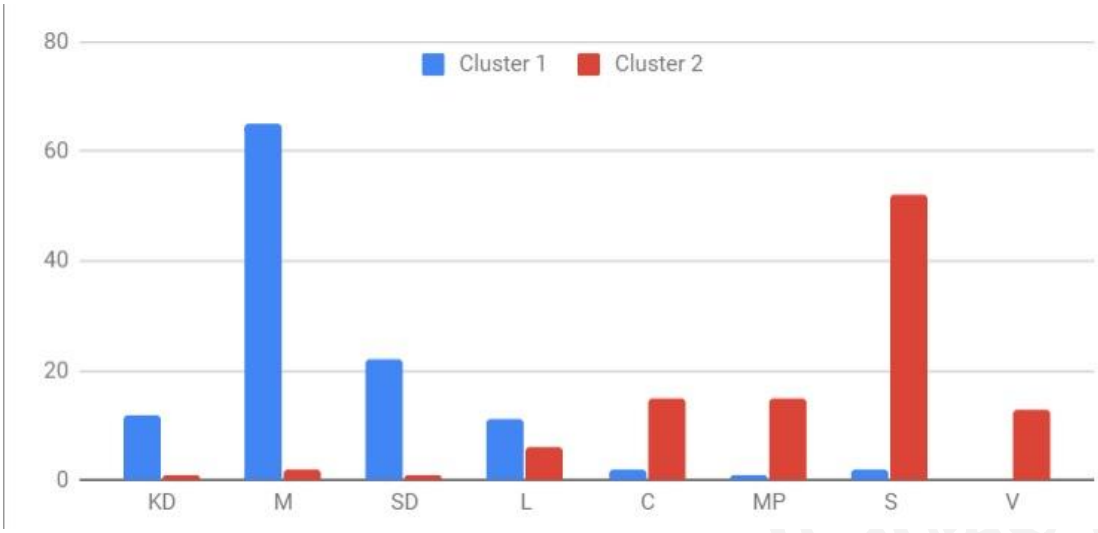


b)

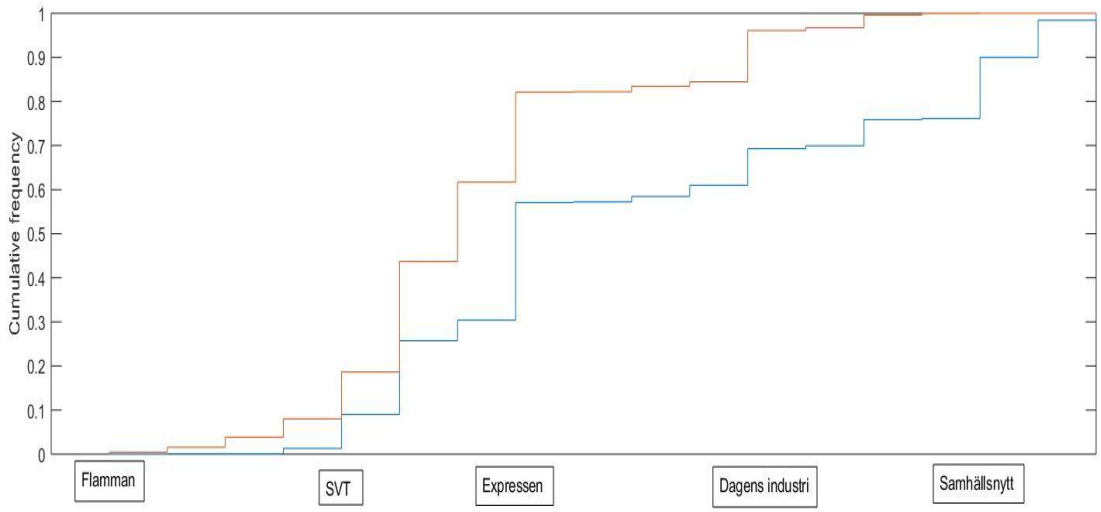


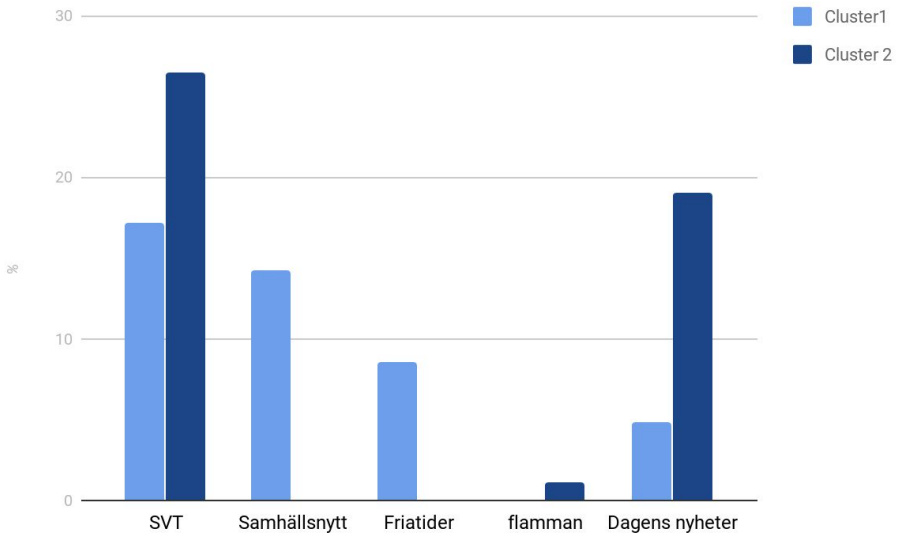


Exploring the Clusters

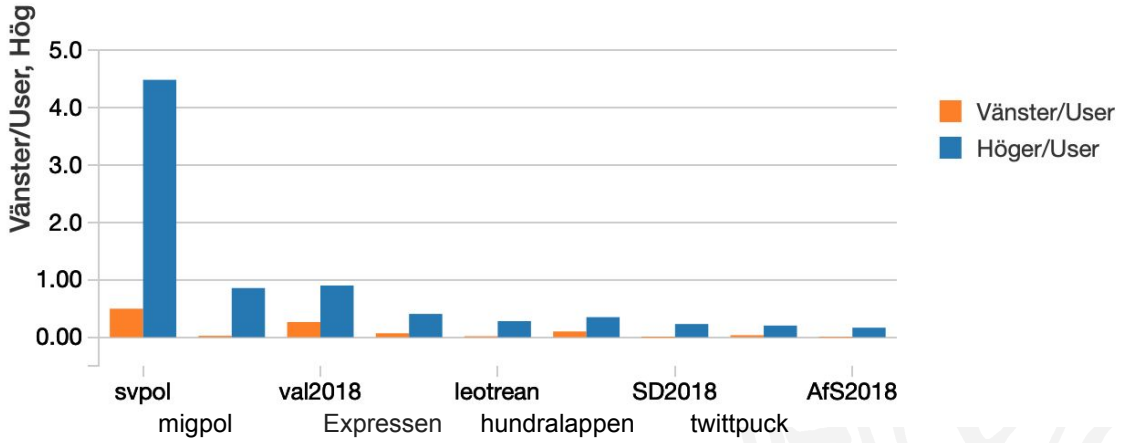


Kolmogorov–Smirnov Test





Hashtag-Distribution



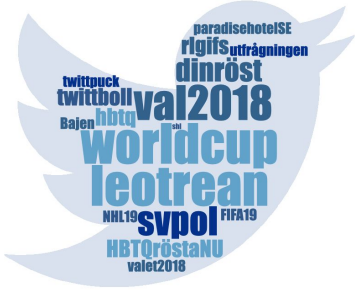
Hashtag Distribution in Retweets

- h0: Is the hashtag distribution of the clusters different from the global distribution?
- Sampled random subgraphs 1000 times with same size as the cluster for each cluster
- Total variation distance from sampled graphs to global hashtag distribution
- Null hypothesis was rejected for all 3 clusters with 0.1 % significance

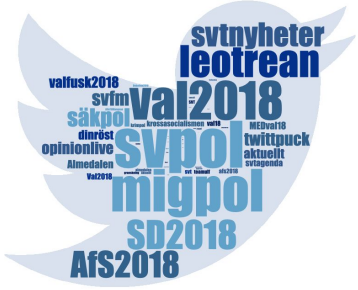
Clusters	Interval	Obs
Right-wing	[0.042,0.044]	0.149
Left-wing	[0.078,0.082]	0.541
Sport	[0.261,0.284]	0.744

Common Hashtags in Clusters

Left-wing cluster

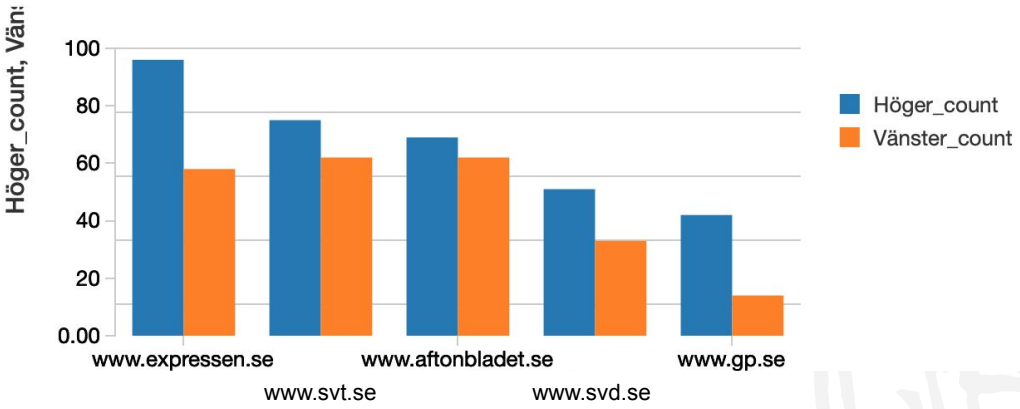


Cluster with a sporty touch



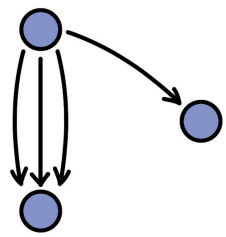
Right-wing cluster

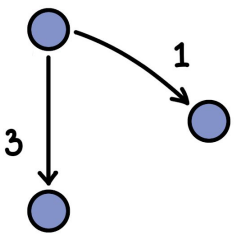
Retweets - URLs

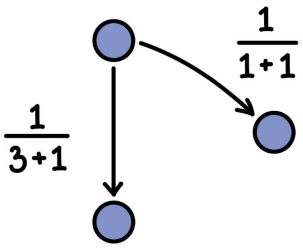


Shortest Weighted Path

- Distributed Dijkstra's algorithm
- Weights - Number of retweets
- Landmarks - Key users
- Degrees of separation





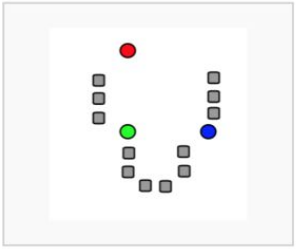


Shortest Weighted Path

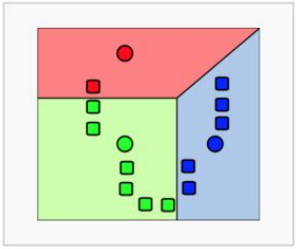
freq	percentage	cs	hanifbali	sdriks	katjanouch	socialdemokrat	strandhall	jsjostedt
84371	37.9	37.9	≥4	≥4	≥4	≥4	≥4	≥4
32732	14.7	52.6	∞	∞	∞	∞	∞	∞
8995	4.04	56.7	3	≥4	≥4	≥4	≥4	≥4
8312	3.73	60.4	≥4	≥4	≥4	≥4	≥4	3
7914	3.56	64.0	≥4	≥4	2	≥4	≥4	≥4
5988	2.69	66.7	≥4	≥4	≥4	≥4	≥4	2
5742	2.58	69.3	2	≥4	≥4	≥4	≥4	≥4
4192	1.88	71.2	≥4	≥4	3	≥4	≥4	≥4
3598	1.62	72.8	3	≥4	≥4	≥4	≥4	3
3005	1.35	74.1	≥4	≥4	≥4	3	≥4	≥4
2892	1.30	75.4	≥4	≥4	≥4	≥4	≥4	1
2885	1.30	76.7	≥4	≥4	≥4	≥4	3	3
2866	1.29	78.0	≥4	≥4	≥4	≥4	3	≥4
2515	1.13	79.1	1	≥4	≥4	≥4	≥4	≥4
2429	1.09	80.2	≥4	≥4	≥4	2	3	≥4
2280	1.03	81.3	≥4	≥4	1	≥4	≥4	≥4
2240	1.01	82.3	≥4	3	2	≥4	≥4	≥4
2107	0.947	83.2	≥4	2	2	≥4	≥4	≥4
1854	0.834	84.0	2	≥4	2	≥4	≥4	≥4
1720	0.773	84.8	3	3	3	≥4	≥4	≥4

K-Means

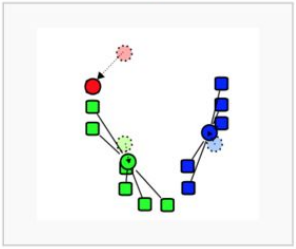
Demonstration of the standard algorithm



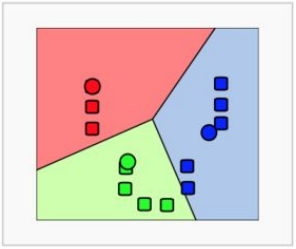
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the *k* clusters becomes the new mean.

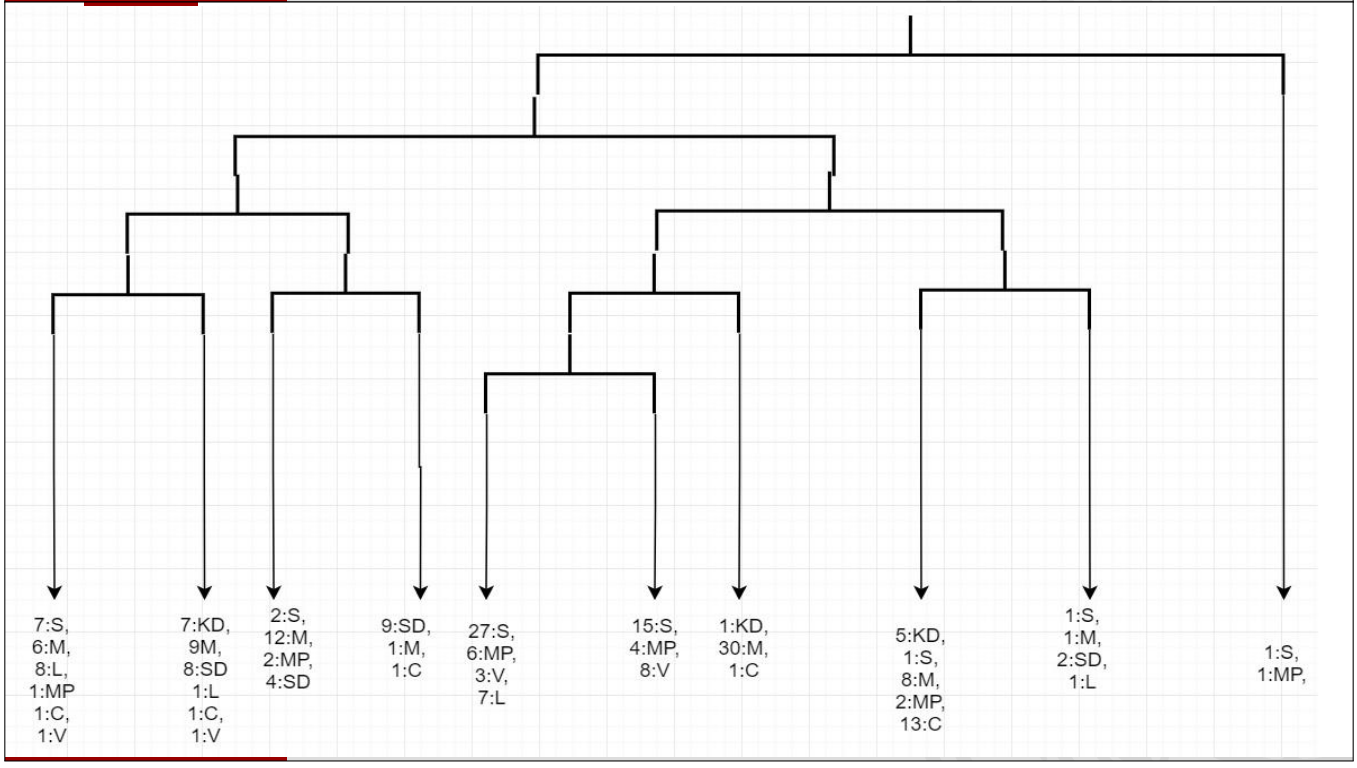


4. Steps 2 and 3 are repeated until convergence has been reached.

Bisecting K-Means

Hanif Bali(M) Isabella Lövin(MP) Jonas Sjöstedt(V)
Paula Bieler(SD) Jeff Ahl(AFS)

Nazispotting katjanouch



Swedish Twitter

Overall Conclusions:

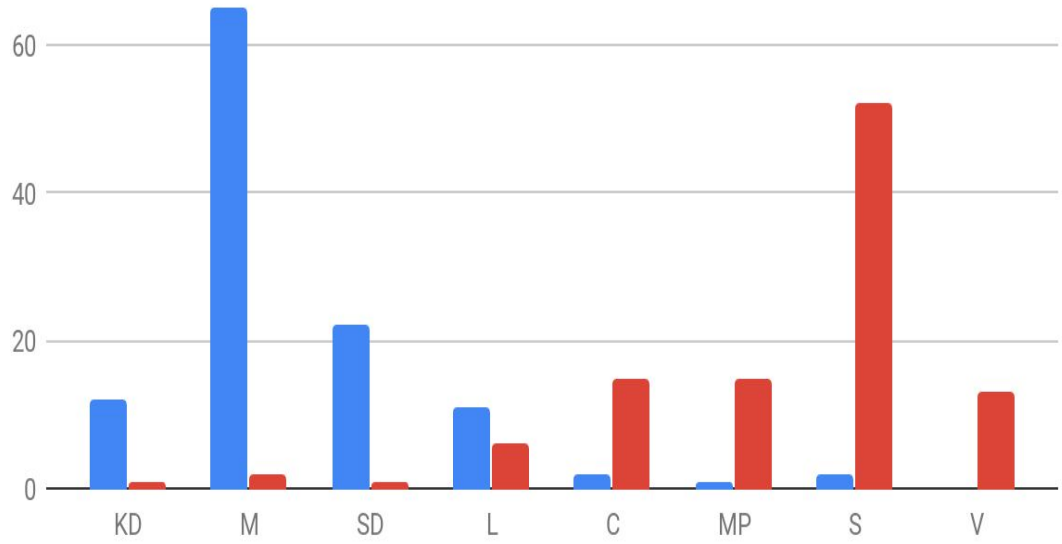
- Clustering captures political affiliation
- The clusters use different news sources
- Different distributions of hashtags and URLs
- We intended to also look closer at the content of the tweet-text itself but the method we had available didn't give meaningful results

Acknowledgements

- Databricks in USA provided free cloud computing
- Combient AB in Sweden donated 4 NUCs
- Tilo has worked hard on on-premise computing support

databricks

Members of Parliament



Kolmogorov–Smirnov Test

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$$