An in-depth analysis of the authorship of political affiliated and localities subreddits

Max Gannett
University of Colorado, Boulder
Boulder, CO
max.gannett@colorado.edu

Jay Ghosh
University of Colorado, Boulder
Boulder, CO
jay.ghosh@colorado.edu

ABSTRACT:

It is known that many ICT's foster environments for political discourse and engagement. The social community of subreddits related to political affiliations and localities gives new insight into the politics of Reddit and fostering of such environments. By generating data from PRAW, which pulls Reddit data, we examined 354 subreddit communities with 22664 unique authors. We show through a quantitative methods approach of network science, machine learning and data science to explore three central questions. (1) What insights are generated by network analysis of authors between political and locality-based subreddits? (2) Does generating random forest and eXtreme Gradient Boosting models produce insight into identifying political leaning localities? Last, (3) we demonstrate and hope to add to the growing understanding of politics in online communities with accessible data and easy-to-follow visualizations.

KEYWORDS:

Social networks, Reddit, Network Science, Random forest, eXtreme Gradient Boosting, Political, computational informational politics

MAIN:

Research around social groups and the interaction of information is common in early human settlements. Questions start to arise with the new digital age of Web 2.0 and humans being both producers and consumers of information in online communities proves a breeding ground for data exploration to understand the fundamentals of human interactions with information. There is a need to understand what is evoked by the homophily of humans and groups and how that translates into online platforms. Since individuals with the same sets of homogenous identities tend to "flock together", this can likely be said about the same of online communities like Reddit. 1 Understanding these phenomena in online interaction of social platforms requires the skills of network analysis to study large-scale interpersonal interactions. We hope to inspire the further development of network and data science principles to help understand social phenomena in online communities when it pertains to political conception.

Recently published in *Nature*, Isaac Waller and Ashton Anderson presented a "Quantifying social organization and political polarization in online platforms" paper that examines much of the same questions around how we see the partisan differences in online platforms like Reddit. 2 Here we hope to develop the same purpose of drawing on different mathematical principles and tools of understanding social phenomena online. Waller and Anderson laid the groundwork for more work to be produced around this type of research. We hoped to expand on what they missed in a different light. Rather than focusing on traditional notions of identities and demographics- age, gender, and political orientation - we wanted to drill down on the political subreddits identified by them, paired with a bigger identity: localities. Not that Waller and Ashton did not think of those subreddits as being important to the overall macroscale structure of political Reddit, but we can simply understand the same phenomena found in their results in the same way by looking more closely at two identifiable subreddit categories.

Here we expanded and corroborated a methodology of using network analysis and science paired with machine learning to answer insights about the political ideologies and localities - states and urban cities - to leverage new information around computational informational politics. Computational informational politics is a new discipline that we hope to grow and attribute to the rapid interdisciplinary skills of computer science, information science, and political science. Essentially, computational politics exists with computer scientists helping political scientist scientists answer their questions using computational tools. We hope to do the same in this paradigm of information science, being information practitioners interested in political problems that can be understood through a computational lens.

One of these tools is network analysis which is a powerful way of studying phenomena as diverse as interpersonal interaction, connections among nodes, and the structure of the Internet. 3 Generating these new characteristics of an emerging mode of inquiry in the political and social sciences is something that hasn't been achieved on a large-scale behavioral scale before. We can utilize these tools to create an encompassing field of work that can make high fidelity inferences about the political identities of individuals and groups in the landscape of American online political activity. 4 Our approach, coupled with machine learning insights, answers the same questions posed by prior research but in a new light with new methods. Centrally, do we see the same results and perceived phenomena of political polarization with new insights through these methods and objectives?

NETWORK STRUCTURES OF POLITICAL IDEOLOGY AND LOCALITIES

We can study social aspects of individuals by examining how they interact on both politically identified subreddits and state and cities subreddits with authors' connections to posting on those subreddits presents a bipartite network analysis. This presents a social phenomenon that pairs well for something in this dynamic of social science and network analysis with 2 mode matrices. 5 With 354 subreddits and 22664 unique authors, the bipartite network seems similar to patterns of other authorship like Wikipedia. 6 Below you can see the overview of the whole network in both forms, **figure 1**.

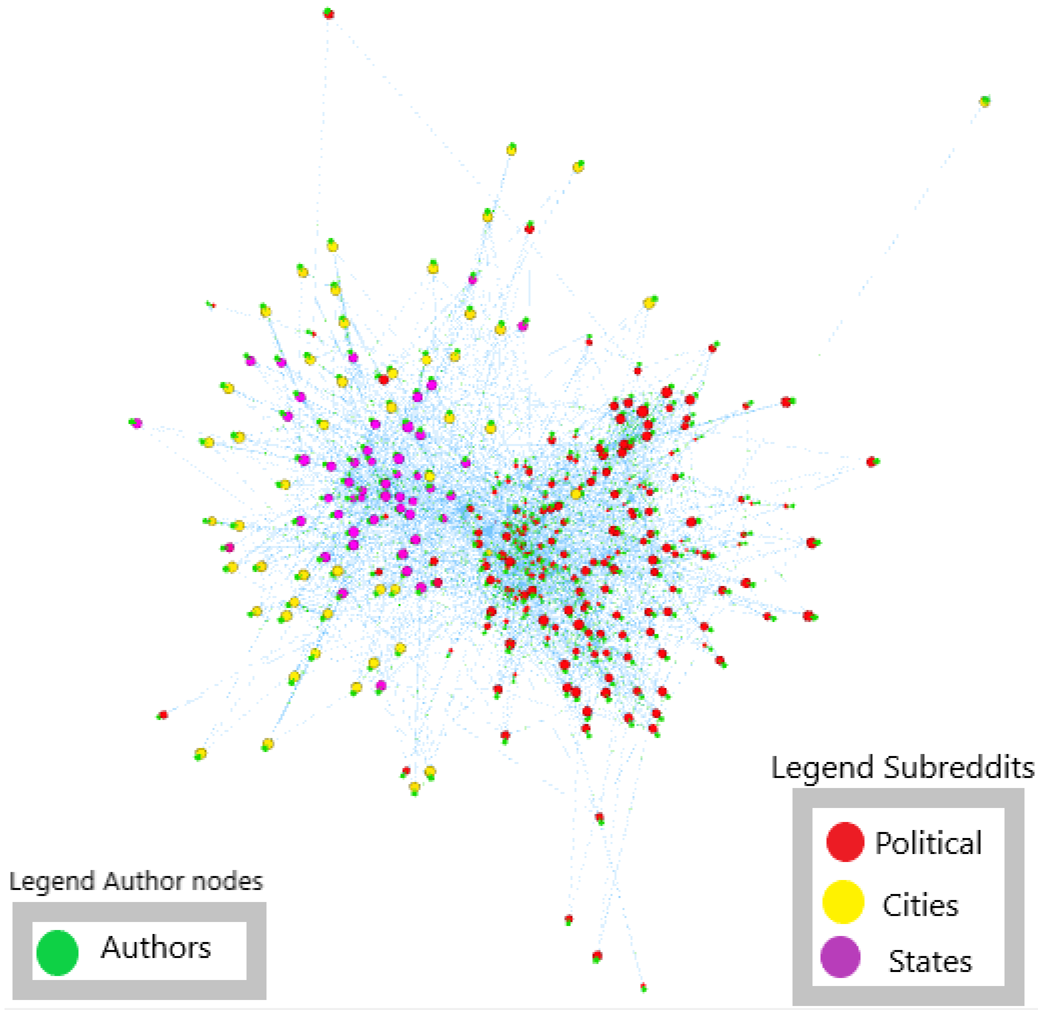


Figure 1: Subreddits of political and localities with authorship

From the above overview of the featured dataset of political identifiable subreddits - classed by Waller and Ashton- and state and city subreddits a divide of two clusters appears. **(Figure 2)** On the left cluster, you can see the connection between states and cities clustered around each other. The right is our politically identifiable subreddits with the heart of clustering being mostly ideological leaning nodes of Democrats and Republicans. However, there are some interesting cities that have authors and centrality towards that political cluster, notably Los Angeles and Austin. **(Figure 3)**

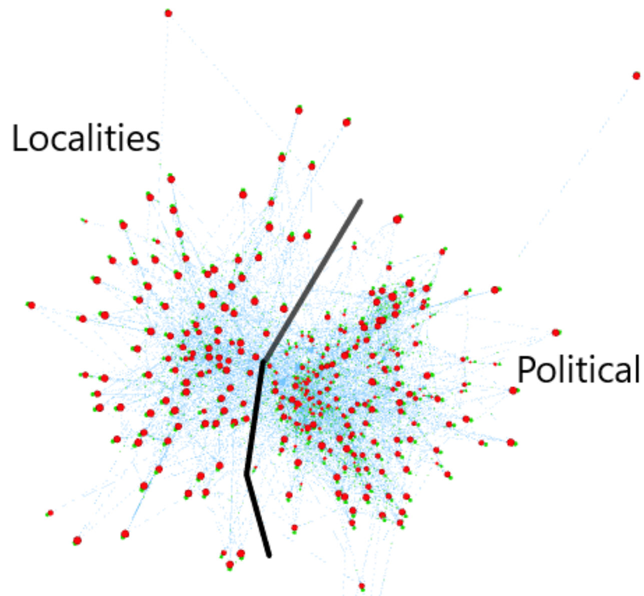


Figure 2: Clustering around Localities and Political subreddits

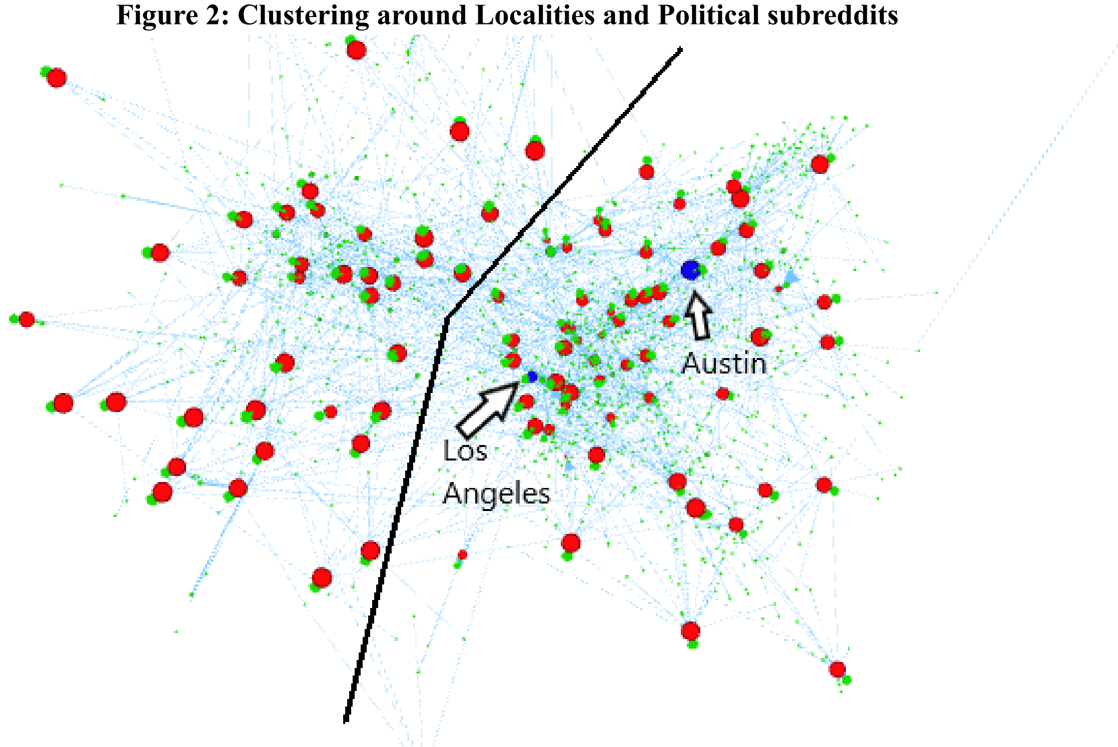


Figure 3: Los Angeles and Austin nodes on the political cluster

With the nodes of Los Angeles and Austin being in close proximity to the politically affiliated nodes like democrats and conservatives, means that those subreddits of the cities have authors contributing to both politically affiliated subreddits and the cities. Both cities leaned heavily Democrat in the 2020 U.S. presidential election with 71% of votes going democratic. However, there is more Democratic vote share for other states and cities with higher percentages than LA and Austin. Trends around these clusters and authorship are a small sample of Reddit data - 22,982 nodes and 25,906 edges - paints more of a picture \of the divide between the subreddits categories than one would say polarization of the authors. What's more, we see some general insight into politically based subreddits like thedonald, hillaryclinton, politics, democrats by having a 2 depth ego network of the subreddit nodes we begin to see a picture of authors' contributions to both pages. (**Figures 3,4,6,7**) Ego networks are representative samples of the social environments surrounding particular elements and are compatible with conventional statistical methods of generalization to large populations. [7]

The ego networks of specific nodes like the ones described above tell a story of authorship around many subreddits. For example when looking at thedonald 2 deep ego network - meaning the nodes connected to the ego node (thedonald) would grab the authors who contributed to thedonald and then those subreddit nodes the authors also contributed to - you see all the localities and other political subreddits that are within one author. The authors that are contributing to thedonald subreddit also contributed to state and cities, but also political ones like conservatives, politics, and panamapapers. Compared to r/hillaryclinton, thedonald has more authors to more subreddit localities. Though more political subreddits authors than hillaryclinton authors. An interesting take that needs more investigation is are more democratically leaning users contributing to more political subreddits - vice versa for states and cities for republicans - or does Reddit have more democratically leaning political subreddits? More data encapsulating prior authors and trends around temporal patterns of elections might prove more insightful for future data analysis as well. Seen in both the Waller and Austin paper and Alexa Pavliuc work is a framework for advancing the analysis of these questions surrounding political Reddit. We implore you to take a look at the visual yourself found in the data resources below to play around with other ego networks. [8]

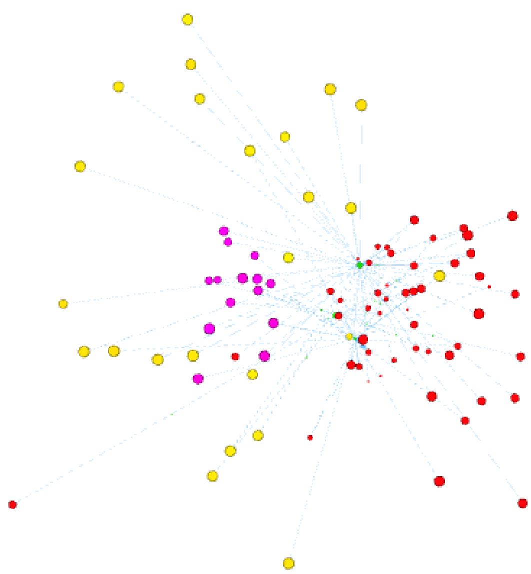


Figure 3: r/thedonald ego network 2 depth

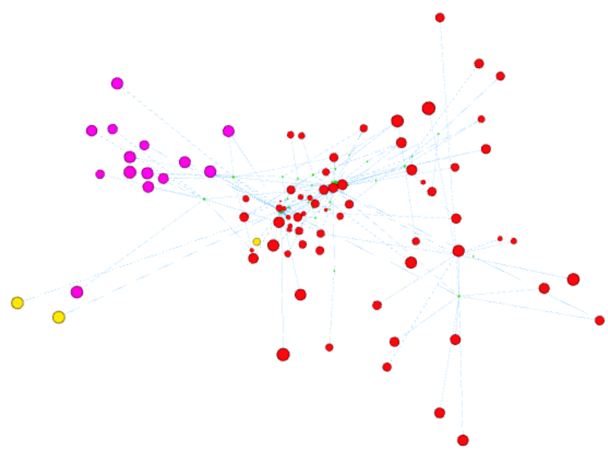
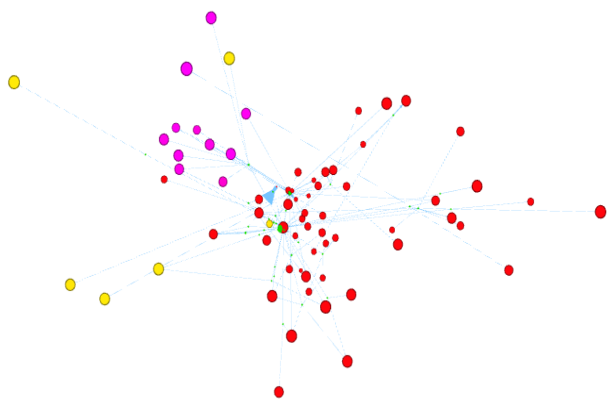


Figure 4: r/hillaryclinton ego network 2 depth



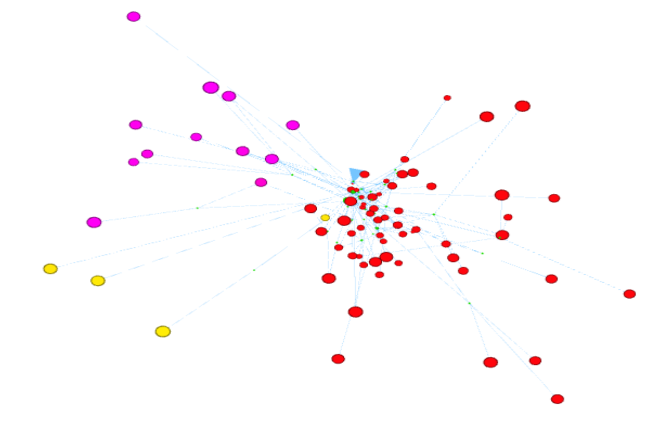


Figure 6: r/politics ego network 2 depth

Figure 7: r/democrats ego network 2 depth

DATA CLEANING AND PREPARATION

We started by using the data from Waller and Austins political subreddits. This study had very valuable information on multiple different subreddits. An important field we used was the "partisan_raw" metric that was calculated for their selected subreddits. This "partisan_raw" metric had a score that ranged from -1.0 to 1.0, with -1.0 representing an extreme liberal partisan bias and 1.0 representing an extreme conservative partisan bias on the associated subreddit. We can see the distribution of these "partisan_raw" scores below.

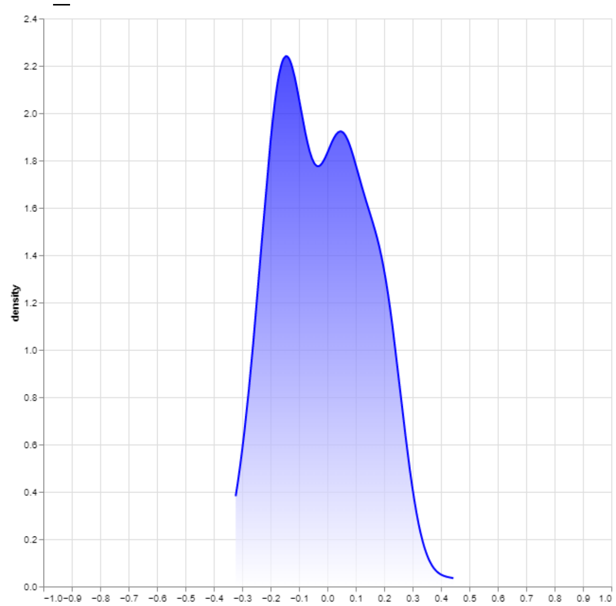


Figure 8: In this we can see a fairly normal distribution that is also slightly bimodal, with the two peaks likely representing mean liberal and conservative leanings.

Once we had the data from Waller and Austins' political subreddits, we decided to scrape more Reddit data ourselves. We compiled a list of cities and state-based subreddits which we will broadly label as localities from now on, in addition to a series of political subreddits. We used

PRAW which is a Python Reddit API Wrapper module/library to scrap our social media data from Reddit. We pulled the most recent 150 posts from each of the following 354 subreddits:

'Portland, Austin, chicago, Seattle, houston, LosAngeles, nyc, Atlanta, boston, philadelphia, washingtondc, Denver, sanfrancisco, rva, SeattleWA, Dallas, sandiego, pittsburgh, baltimore, Columbus, kansascity, newjersey, NewOrleans, texas, cincinnati, Charlotte, Sacramento, California, minnesota, sanantonio, tampa, Connecticut, Detroit, Louisville, Hawaii, wisconsin, vegas, Michigan, Cleveland, madisonwi, Rochester, Miami, indianapolis, milwaukee, raleigh, Buffalo, Maine, NorthCarolina, oklahoma, Tucson, Albuquerque, boulder, jacksonville, memphis, florida, tulsa, Chattanooga, oregon, alaska, Albany, Charleston, FortWorth, newhampshire, Delaware, maryland, Knoxville, Ohio, arizona, Indiana, Iowa, vermont, Pennsylvania, Arkansas, desmoines, Louisiana, Alabama, Virginia, southcarolina, Washington, RhodeIsland, Kentucky, Montana, LasVegas, massachusetts, Utah, kansas, WestVirginia, illinois, Georgia, Tennessee, bullcity, mississippi, northdakota, missouri, wyoming, Idaho, newyork, NewMexico, Nebraska, SouthDakota, Colorado, Libertarian, Conservative, Anarcho_Capitalism, socialism, Anarchism, Republican, progressive, democrats, Liberal, conservatives, feminisms, Objectivism, Egalitarianism, GreenParty, alltheleft, labor, republicans, americanpirateparty, Capitalism, LibertarianLeft, Marxism, politics, worldnews, news, The_Donald, TumblrInAction, SubredditDrama, SandersForPresident, MensRights, PurplePillDebate, EnoughTrumpSpam, hillaryclinton, Drama, PoliticalHumor, forwardsfromgrandma, ShitRedditSays, AskTrumpSupporters, GamerGhazi, LateStageCapitalism, TheBluePill, CapitalismVSocialism, uncensorednews, WayOfTheBern, Gamingcirclejerk, circlebroke, conspiratard, FULLCOMMUNISM, AskThe_Donald, pussypassdenied, badhistory, MarchAgainstTrump, FeMRADebates, AgainstGamerGate, european, esist, Buttcoin, Political_Revolution, TopMindsOfReddit, circlebroke2, SRSsucks, ShitPoliticsSays, EnoughLibertarianSpam, GenderCritical, Enough_Sanders_Spam, CoonTown, samharris, Kossacks_for_Sanders, AskFeminists, badphilosophy, undelete, GGFreeForAll, HillaryForPrison, Feminism, enoughsandersspam, ShitWehraboosSay, Fuckthealtright, DebateFascism, WikiLeaks, moviescirclejerk, Shitstatistssay, Negareddit, SRSDiscussion, shitpost, COMPLETEANARCHY, sjwhate, OneY, againstmensrights, TiADiscussion, ShitLiberalsSay, GaryJohnson, BestOfOutrageCulture, BlueMidterm2018, beholdthemasterrace, altright, EverythingScience, subredditcancer, Trumpgret, PoliticalVideo, AgainstHateSubreddits, BannedFromThe_Donald, DebateaCommunist, DebateAltRight, ShitRConservativeSays, WhiteRights, DebateAnarchism, SRSGaming, MensLib, communism, tucker_carlson, EnoughPaulSpam, GoldandBlack, antisrs, Anarchy101, POTUSWatch, TrumpCriticizesTrump, DebateCommunism, worstof, no_sob_story, AskALiberal, metanarchism, SocialJusticeInAction, askaconservative, Mr_Trump, jillstein, PussyPass, Impeach_Trump, libertarianmeme, communism101, DNCleaks, the_meltdown, GreatApes, WhereIsAssange, TumblrPls, thedavidpakmanshow, RightwingLGBT, women, OurPresident, Blackout2015, TumblrAtRest, PanicHistory, The_Mueller, SRSWomen, Equality, killthosewhodisagree, SubredditDramaDrama, randpaul, LateShow, JustUnsubbed, bidenbro, badpolitics, Gender_Critical, EnoughCommieSpam, antifa, lastweektonight, DailyShow, AntiPOZi, shittankiessay, askhillarysupporters, DarkEnlightenment, HillaryMeltdown, FemmeThoughts, media_criticism, austrian_economics, socialjustice101, Socialism_101, EnoughHillHate, AskBernieSupporters, MarchForScience, Physical_Removal, TheNewRight, The_Congress, BadSocialScience, SRSBusiness, Le_Pen, PanamaPapers, OffensiveSpeech, badscience, Stuff, AsABlackMan, SRSMeta, Maher, HillaryForAmerica, WikiInAction, racism, forwardsfromhitler, new_right, thedonald, ThePopcornStand, LibertarianDebates, bad_religion, The_Farage, RussiaLago, dancarlin, NeverTrump, openbroke, europeannationalism, justicedemocrats, CNNmemes, TinyTrumps, hottiesfortrump, Voat, redacted, GrassrootsSelect, Oppression, StormfrontorSJW, AskLibertarians, SRSFunny, u_washingtonpost, The_Europe, ShitThe_DonaldSays, Drumpf, LeftWithoutEdge, TrumpForPrison, StillSandersForPres, INeedFeminismBecause, AnarchistNews, trump, TumblrCirclejerk, The_Dotard, RedditArmie, ABoringDystopia, TIL_Uncensored, AntiTrumpAlliance, FULLDISCOURSE, NewPatriotism, Full_news, GasTheSnoo, whitebeauty, sjsucks, SRSFeminism, SummerReddit, SRSMen, goldredditsays, SRSQuestions, RedditCensors, forwardsfromreddit, WatchRedditDie, ForwardsFromKlandma, NewYorkForSanders, tulsi, Political_Tweets, MensRants, BernTheConvention, menkampf, pol, FreeSpeech, PaoYongYang, uncensorship, anarchy, Sorosforprison, Infowars, censorship, Ellenpaohate'

We pulled the timestamp, author, title, subreddit, and ID of each of those 25,906 posts. 25,906 posts may seem small given the amount we intended to scrape. However, we had to clean the data and thus eliminated observations that lacked values and additionally we lined up the political subreddits we personally identified with the political subreddits identified in Waller and Austin's study. In that study, they labeled subreddits by type or genre under the "cluster_name" tag. They actually mislabeled several subreddits, for example, they mislabeled the democrats subreddit as "General interest 2" instead of "politics". So our final data included an inner join of subreddits we personally labeled as being political that were present in the entire Waller and Austin's database, combined with the subreddits they labeled as political; in addition to of course the locality-based subreddits we personally identified.

After collecting this network data, we had it essentially formatted as one column being the author and another column being the subreddit, with a final column representing how many times that author posted to that subreddit as our edge weight; with each observation/row representing a post. This data format was very useful in being able to visualize the network structure of the data as this was essentially an edge list. However, it was not yet ready for machine learning and predictive modeling.

	author	subreddit	post
0	---Sanguine---	news	6
1	---Sanguine---	worldnews	3
2	--MadMatt--	Feminism	1
3	--MadMatt--	Liberal	1
4	--Spartan45--	Conservative	1
...	...	...	...
26447	zwrite	electionreform	1
26448	zzarkino	lgbt	1
26449	zzdanny	Conservative	1
26450	zztop610	politics	2
26451	zztx2zzm8stop	Anarchism	1

Figure 9: Edge Table

To transform this network data into a feature vector needed for machine learning, we had to perform several complex data transformations. In order to do so, we had to iterate through the nodes list data frame of the locality-based subreddits and then iterate through our edges list data frame nested within our first iterative process. As we iterated through the edge list within our locality-based nodes list iteration, we checked to see if the author in the edge list posted to the locality-based subreddit. If it passed this check, we then iterated through the list of political subreddits and checked to see if the author also posted to any of the political subreddits in addition to the locality. If they passed this check, we added the edge weight, (the number of times the author posted to the political subreddit), to our new feature vector. In this feature vector, every observation/row is a locality-based subreddit, with the features/columns, representing the political subreddits. For the columns, we have the aforementioned sum of counts of how many posts, users who posted in the locality-based subreddit, posted to the political subreddits. In addition, we had a column that was the “raw_partisan” score for the political subreddit that was calculated by Waller and Austin’s study. We then added 0.25, to account for zero values, to each count and performed a log10 transformation on the ‘count’ feature to normalize the data. Then we multiplied this log10 count feature by the partisan score and calculated a new influence metric. Finally, we dropped the partisan score as it has no variance. We gathered electoral results for each of these localities in the 2020 elections to serve as our labels, with 0 representing a Democrat victory and 1 representing a Republican victory calculated on a vote share total of all votes. Now was ready for machine learning.

	locality	Libertarian	Libertarian_influence	label
0	Portland	-0.602060	-0.105277	0.0
1	Austin	-0.602060	-0.105277	0.0
2	chicago	-0.602060	-0.105277	0.0
3	Seattle	-0.602060	-0.105277	0.0
4	houston	-0.602060	-0.105277	0.0
...	...	...	...	...
95	Idaho	-0.602060	-0.105277	1.0
96	newyork	0.352183	0.061583	0.0
97	NewMexico	-0.602060	-0.105277	0.0
98	Nebraska	1.183270	0.206907	1.0
99	SouthDakota	1.051153	0.183805	1.0

101 rows × 507 columns

Figure 10: Feature Vector table of post counts and partisan influence scores

DUMMY MODEL AND LOGREG MODEL

Now that we had our feature vector prepared, it was time to conduct our machine learning for the purposes of predicting whether that locality leans Democrat or Republican. This was an imbalanced dataset. We had a ratio of 74 Democratic localities to 26 Republican localities. In order to have a more balanced training set, we randomly sampled 80% of the Republican observations and matched it with the same count of Democratic observations. This resulted in a ratio of 21 Democratic localities to 21 Republican localities in our training dataset. The other observations were lumped into our testing set with a ratio of 51 Democratic localities to 5 Republican localities. Given how 74% of the data was Democratic, we felt that sampling by class for the training data was the best way for our model to predict Republican observations.

To evaluate our machine learning models, we will be measuring the accuracy, precision, and recall of our models. Accuracy simply measures how many of the observations in the dataset we correctly predicted. Our predictions are either 0 for Democrat or 1 for Republican. Precision measures how out of all of the observations we predicted to be positive, the amount of those that were actually positive. So having high precision often results in low false-positive rates. Recall measures the models' ability to capture all of the relevant/positive items. So recall essentially measures how out of all of the positive observations, how many did we actually predict to be positive. Having high recall results in having low false negatives, but often high false positives. This is what is known as the precision/recall tradeoff and for our analysis, we will be focusing on increasing our precision.

With our train/test split finalized, we moved forward to creating a ‘dummy model’ for baseline performance. A dummy model simpy predicts everything as being the class with the highest frequency. For instance, our dummy model predicted everything in the test set as being Democratic, resulting in an accuracy of 0.9137931034482759, but having 0 recall and 0 precision since it merely classified everything as Democrat.

We then moved towards tuning and training a logistic regression model as another baseline to measure against future models’ performance. After tuning and training the logistic regression model, we found an accuracy of 0.08620689655172414 and precision equally at 0.08620689655172414. The recall for the model was 1.0. Overall, this model expectedly performed very poorly, indicating that this is not a simple problem that can be solved via logistic regression.

Random Forest Model

In addition to creating a dummy model and a logistic regression model as a baseline, we created a random forest model to try and actually predict the political affiliations of these localities with strong performance and accuracy. We decided to try a random forest model because random forest models are very fast and handle large numbers of features very well. We tuned and trained the random forest model using a randomized search, searching the optimal number of trees, number of features to consider at every split, the maximum number of levels in the trees, the minimum number of samples required to split a leaf node, the minimum number of samples required at each leaf node, the method of selecting samples for training each tree, and whether the class weights should be considered balanced for our hyperparameter tuning. After performing this randomized search for the optimal hyperparameter values we fit and trained the random forest model on our training dataset.

In examining its performance on the test data, we found a Mean Absolute Error of 0.36 degrees, an accuracy of 0.6379310344827587, precision at 0.1, and recall at 0.4. We then calculated shapley values to conduct feature importance analysis to understand our model’s output. A visualization of the shapley values can be viewed below, Class 0 being Democrat and Class 1 being Republican:

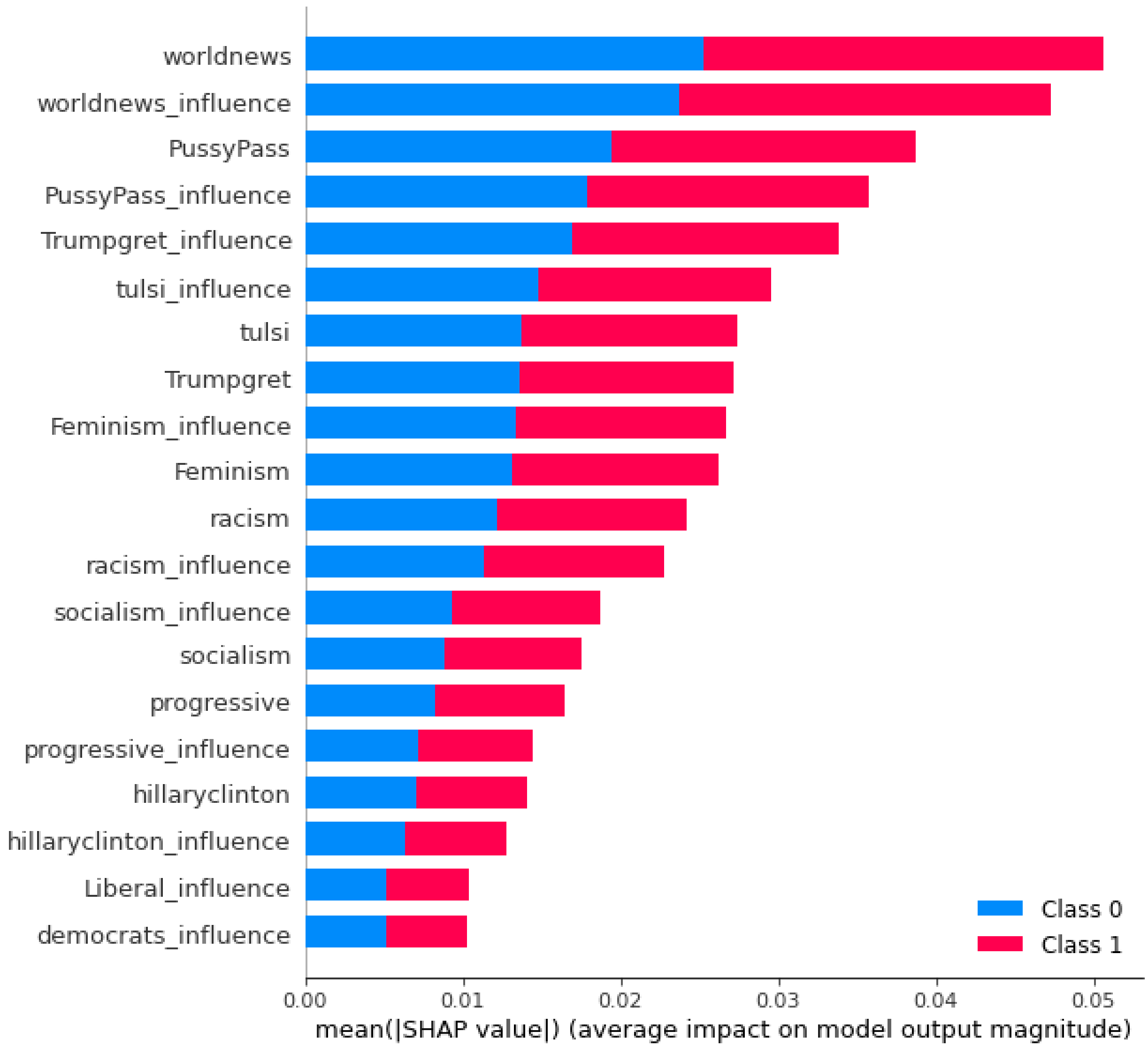


Figure 11: Mean values of computed scores of democrat and republican subreddits

Below, we can see how this random forest model performs at different classification thresholds.

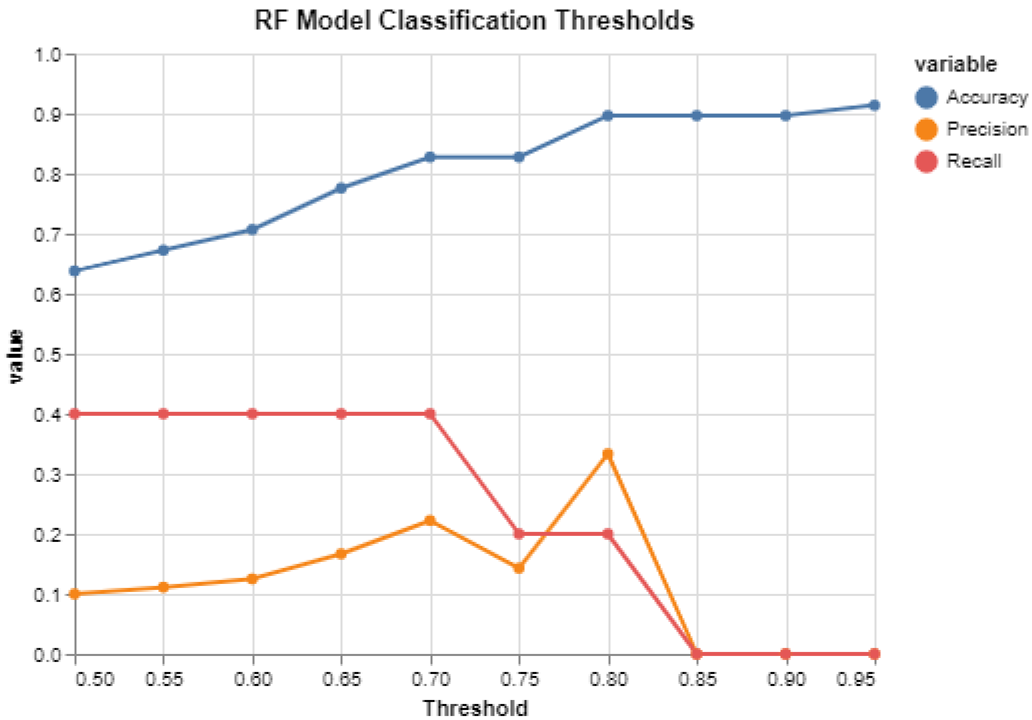


Figure 12: Random Forest Model Performance Metrics at Different Classification Thresholds

After setting a classification threshold of 85%, this model does not predict any positive samples. That is to say that it did not predict any localities as being Republican with a probability of over 80%, so it cannot properly calculate precision and recall. For accuracy's sake, we would recommend that the classification threshold for this model be set at 80%.

XGB Model

While the random forest model performed much better than the logistic regression model, we were still not satisfied with that model's performance. So we decided to tune and train an XGB model. XGB simply means gradient boosted trees. It's essentially a 'super' version of a random forest model. After performing a randomized search for the best hyperparameters, including the max depth, the learning rate, number of estimators, the fraction of samples per tree, alpha levels, lambda levels, and finally which tree method to use, we fit and trained our model on the training data set.

In examining its performance on the test data set, we can observe a mean absolute error of 0.29 degrees, an accuracy of 0.7068965517241379, with recall at 0.6 and precision at 0.16666666666666666. This has better performance than our random forest model. We analyzed the feature importance of this model by calculating and visualizing the shapley values for this model on the test data, seen below:

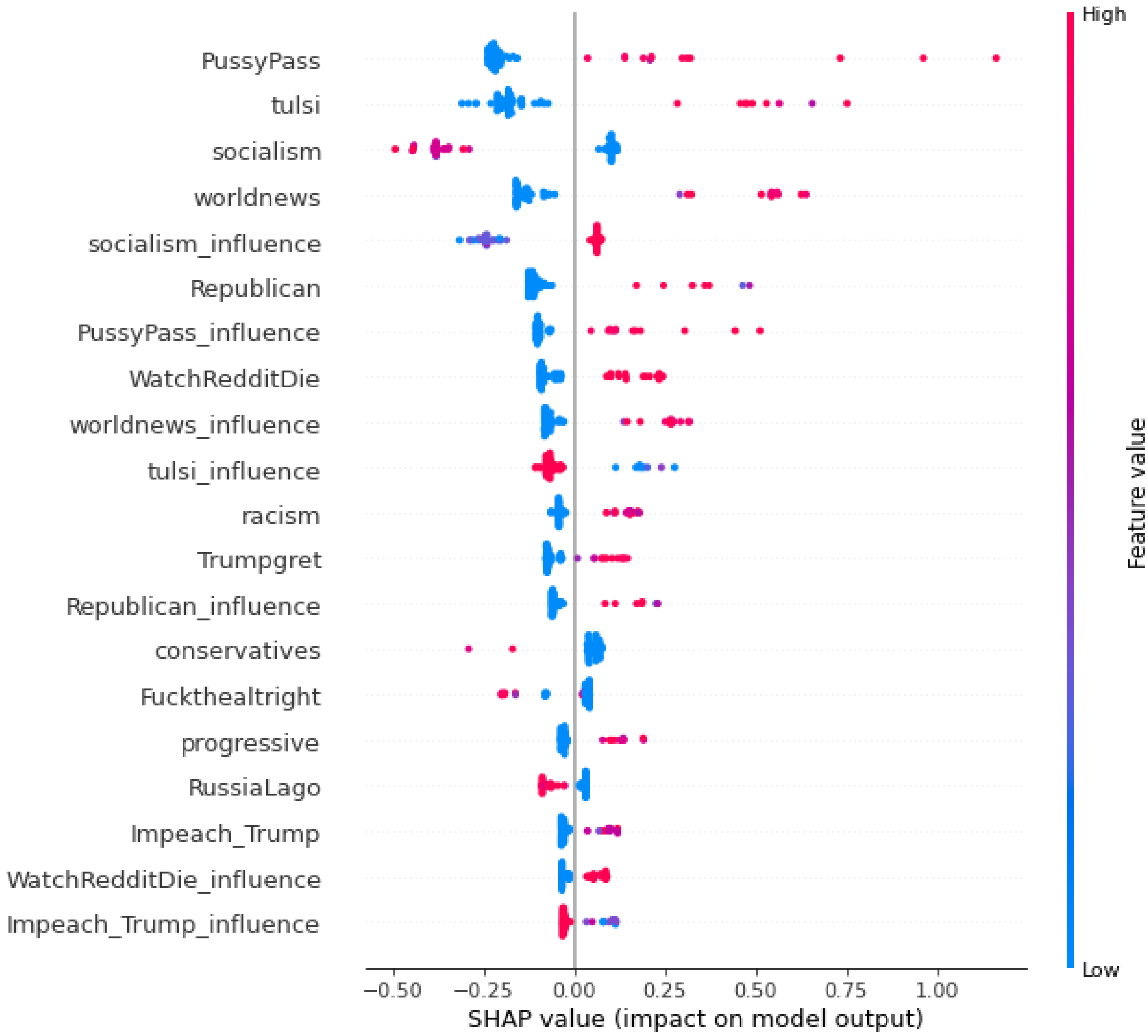


Figure 13: Shapely values of XGB model Features(subreddits)

Since this was our best performing model, we analyzed how this model performed at different classification thresholds as visualized below:

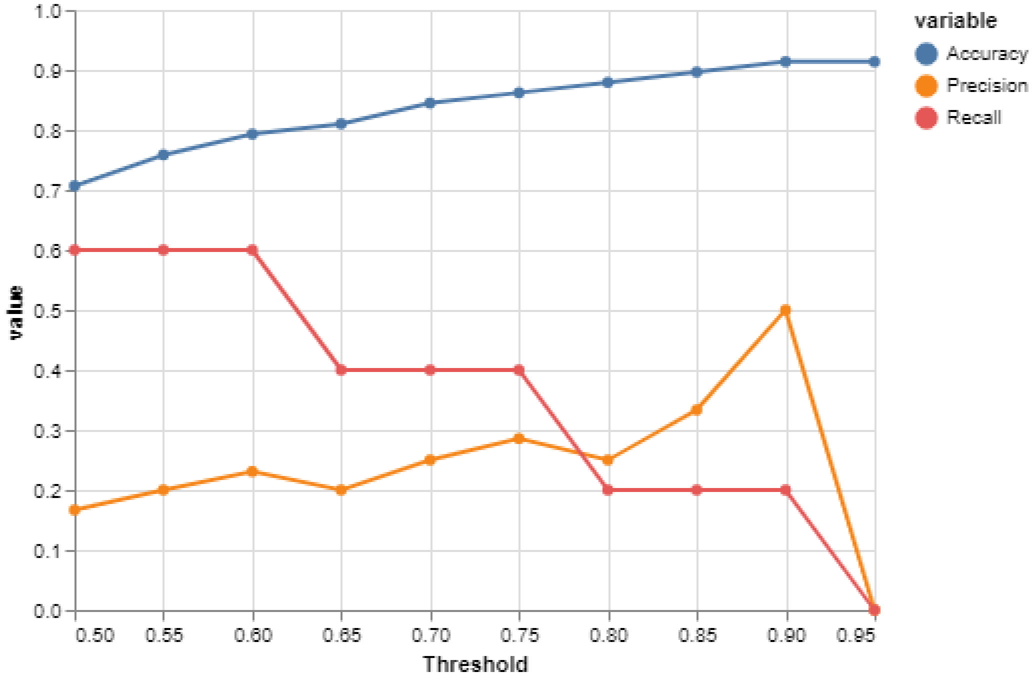


Figure 14: XGB Model Performance Metrics at Different Classification Thresholds

This shows us that to maximize our precision and accuracy we should set a 90% classification threshold as this results in around 50% precision and over 90% accuracy. If instead, we want to maximise our recall, we could set our classification threshold to 60%, resulting in 60% recall and around 80% accuracy; however, this comes at the tradeoff of only around 24% precision.

DISCUSSION OF RESULTS AND TRAJECTORY OF FUTURE WORK

In predicting political leanings and/or electoral outcomes in different localities using a variety of machine learning models, including logistic regression, random forest, and gradient boosted trees models, we've found that the gradient boosted trees model had the strongest performance. These results of 50% precision and over 90% accuracy at the 90% classification threshold are very promising, though it is clear that this model does not completely capture the political sentiments of these localities. There is certainly a liberal bias in the data we collected and that likely plays a role in how effective our models can be. While the social media network data of connections between locality-based subreddits and political subreddits seem to be useful in predicting political affiliations, it does not provide a complete picture. To gain a more complete picture and provide better predictions, perhaps other data is needed or a larger sample.

In the future, we might change the post count we collected to comment count and maybe weight that by the average comment score of that user divided by the average comment score of the subreddit. Since discussion subreddits are largely driven by comments we think comment numbers are a better indicator of political influence. Weighting by average user comment score/average subreddit comment score will better quantify the fact that people who get more upvotes are more influential than those who don't get many or who often have negative comment scores. This type of scoring could lead to stronger performance of our models. However, transforming the network data with our relatively small sample into a feature vector for machine learning took a little bit over 6 hours, so this form of analysis is very computationally expensive.

In conjunction with network science and machine learning techniques, we hope to grow the available knowledge and process of understanding political engagement online. This is just one part of a series of exploring complex problems in our political environments that will add in capstone projects for the final semester of our undergraduate degrees. We hope to contribute and grow this discipline that we have displayed in our main statement of computational informational politics. With an ever-growing need to understand and provide new insight into political problems facing the world, we hope to bring the tools such as network science and data science to aid in answering these tough questions and provide a better place for a healthy political discourse in society. That is our aim and hope in this contribution and the ongoing role we both play in shaping these fields.

Data from figures and this research can be retrieved from Github

https://github.com/jayghosh8973/politicalData

References:

1. McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27,* 415–444. https://doi.org/10.1146/annurev.soc.27.1.415
2. Waller, I., Anderson, A. Quantifying social organization and political polarization in online platforms. *Nature* **600,** 264–268 (2021). https://doi.org/10.1038/s41586-021-04167-x
3. Butts, C.T. (2009). Revisiting the Foundations of Network Analysis. *Science, 325*, 414 - 416.
4. Conover, M.D., Gonçalves, B., Flammini, A. *et al.* Partisan asymmetries in online political activity. *EPJ Data Sci.* **1,** 6 (2012). https://doi.org/10.1140/epjds6
5. Borgatti, S.P. and Everett, M.G. (1997) Network Analysis of 2-Mode Data. Social Networks, 19, 243-269.https://doi.org/10.1016/S0378-8733(96)00301-2
6. Keegan, Brian http://www.brianckeegan.com/2014/12/the-news-on-wikipedia-in-2014/
7. Marsden, Peter. (2003). Network Data Measurement. Annual Review of Sociology. 16. 435-463. 10.1146/annurev.so.16.080190.002251.
8. Pavliuc, Alexa https://towardsdatascience.com/disinfovis-how-to-understand-networks-of-disinformation-through-visualization-b4cb0afa0a71