Examining Political Homophily of Trump and Clinton Supporters During The 2016 US Election

u1623195

Mark Artyukh

13/05/2020

Word Count: 4444

Introduction

In this paper I analyse the political homophily and heterophily levels of Donald Trump and Hillary Clinton supporters during the 2016 US presidential election using a large corpus of data from the social media platform Reddit; As well as, I introduce, as far as I am aware, a new method of tracking the levels of homophily for a given community on Reddit through time, which can add to the tools available at hand, when analysing idealogical segregation on Reddit and wider internet (Gentzkow & Shapiro, 2011). Homophily describes the preference of individuals in a social network to form ties with individuals which have similar political views (McPherson, Lovin, & Cook, 2001). Heterophily, is the opposite; it describes a preference for alternative opinions. The most popular explanation in academia for such behaviour is that interactions with someone of a different opinion on a given topic may lead to cognitive dissonance, which induces stress and pressures us to conform or defend our opinion. Whereas, exposure to opinions similar to ours reaffirms our belief in our worldview being correct, and generally sends us a positive signal. Through selective exposure, individuals end up forming ties with like-minded individuals, which leads to the formation of homogeneous groups. Such behaviour has been noted when meeting people and forming friendships (McPherson et al., 2001). In the context of online political discourse, selective exposure leads to the formation of political 'echo-chambers', which are communication spaces where the individuals' views are similar and their opinions are reinforced with no one to give an opposing view (Flaxman, Goel, & Rao, 2016).

Democracy works most effectively when citizens have beliefs accurately reflecting their surroundings (Becker, 1958). But in order to have such beliefs, one must expose themselves to diverse, and sometimes opposing arguments so as to gain more information. Thus 'echo-chambers' lead to increased polarization of the individuals that are part of these communication spaces, and polarization of a nation in turn leads to the deterioration of social cohesion and leaves citizens uninformed and unable to make optimal decisions in the democratic process (Becker, 1958).

'Social Networking Sites' (Danah & Ellison, 2007; Sponcil & Gitimu, 2012), also referred to as social media platforms,have become arguably the most relevant topic of modern times, having completely transformed the way we communicate with one another. The nature of such platforms allows for rapid transmission and diffusion of information across the network of participants. This has made possible the viralization of certain news within minutes of its entering into the stream of the social platform (Honeycutt & Herring, 2019). It follows, in the context of political discourse, social media platforms are very important as they allow for real-time exchange of political opinion with others with geographical distance not being a factor. Literature on political discourse on social media platforms mainly focuses on two theories. The first being that social media platforms foster a space of open communication

providing exposure to alternative opinions and opposing opinions (Danah & Ellison, 2007; Kwak, Lee, Park, & Moon, 2019). The second theory, being the opposite, is that social media platforms are highly segregated due to selective exposure and the formation of 'echo-chambers', leading to consumption of one sided information, and radicalization of opinions (Conover et al., 2016). Colleoni, Rozza, and Arvidsson (2014) looks at political discussions of Democrats and Republicans on Twitter from a different angle, by building on the work of Kwak, Lee, et al. (2019). Colleoni et al. viewed twitter as both 'newsy' medium, where user interactions are very one-sided within the social graph, as well as, a 'reciprocating' network, where users reciprocate efforts to communicate with one another. The authors found mixed results, supporting both of the above theories. In general, Democrats were found to be experiencing high levels of homophily whereas Republicans showed low levels of homophily.

This paper is written specifically on political discourse on Reddit, and Reddit is relatively new to academia and is new in the research of political discourse. Given Reddit is a social media platform and is now the third most popular website in the US with the number of monthly unique visitors exceeding that of twitter (Peterson, 2019), it plays an important role in the domain of politics, and it's presence in academia will only grow in the coming years. There is one paper written on political homophily of the supporters of Donald Trump and Hillary Clinton on Reddit (Kwak, An, Posegga, & Jungherr, 2019). Kwak, An, et al. looked at the user interactions in a number of communities where political discourse takes place and calculated and compared the expected and observed political homophily levels for both groups of users. They found that both Donald Trump and Hillary Clinton supporters experienced a high degree of homophily when communicating in their respective 'homogeneous' communication spaces, which are spaces with little flexibility in proposing alternative arguments, and a high degree of heterophily when communicating in 'cross-cutting' communication spaces, spaces where any arguments would permitted. In their paper, communication spaces are subreddits, which are communities on Reddit, which users can choose to be part of. Additionally, they looked into the nature of these discussions and found that the interactions between Trump and Clinton supporters in cross-cutting spaces took on a persuasive style of communication. The authors, however, limited their analysis to a subset of political discourse on Reddit, by looking at only 4 different communities, or communication spaces, if you will. Additionally, they didn't look at homophily levels across the board, rather, they calculated separate scores for each community, thus they didn't conclude whether in general there was a high or low level of homophily. Finally, the potential impact of the presence of bot accounts, or automated accounts and deleted accounts was not addressed, even though it made up a significant portion of their data. The authors made an important and first of a kind contribution to research on internet communication by looking at Reddit, however these limitations need to be addressed.

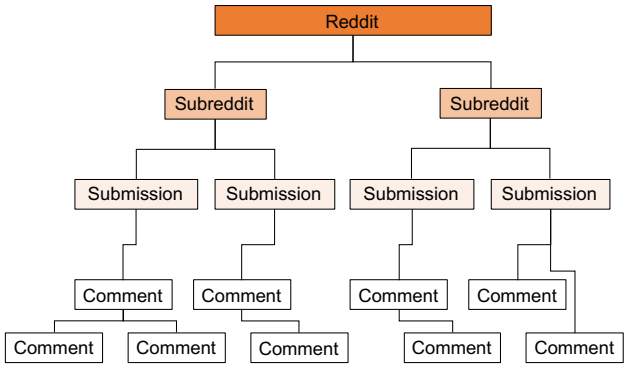


Figure 1: Structure of Reddit threads

Building on the work of Kwak, An, et al. this paper aims to expand the domain of communities looked at, in an attempt to look at all political discourse on Reddit, as opposed to a fraction, as well as address the problems of deleted comments and bot accounts. The first question I look at is what the levels of political homophily and heterophily of Trump and Clinton supporters are and test whether they are above the levels expected given no selection bias. Given Reddit's structure is similar to that of a forum, where users can browse through different messaging boards, join existing threads or start their own, it's reasonable to expect that users would exhibit similar behaviour to users on a forum. There is evidence from a large online forum 'Tianya' that users experienced high levels of homophily (Medaglia & Yang, 2016). Thus, it is expected that Reddit would similarly have high levels of homophily. Political heterophily, on the other hand, is less obvious and one would expect it to be lower than in the case of no selective exposure, as it would require additional effort in order to seek out individuals with opposing opinions. The second question is how the level of political homophily and heterophily varied during the months before and after the US presidential election. In answering the second question I show that it is possible to track the level of political homophily on Reddit across time.

Methodology

Reddit Structure:

Reddit is a social media platform where individuals have the ability to join topic-based communities, called 'subreddits'. Like in the popular social media platforms Facebook and Twitter, users can make a post in the form of a comment, picture, video, or link, as well as other mediums, and other users can then reply to the post by commenting on it, show approval by giving it a point ('up-vote') or disapproval by removing a point ('down-vote'), as

well as, users can comment on other users' comments, thereby creating a tree-like structure like (see Figure 1). Users can also communicate directly with one another through private messages. One unique aspect of Reddit is that every post (and therefore every comment), has to be made in a specific subreddit; There is no open space for communication like there is on Twitter.

Data:

Subreddit	# Comments	# Submissions	# Total
The_Donald	7977562	990928	8968490
news	3688655	427713	4116368
politics	3723589	212957	3936546
ukpolitics	802791	20049	822840
PoliticalDiscussion	687326	13797	701123

Table 1: Top 5 subreddits by number of comments and submissions

Data used in this paper is part of a larger data repository which was collected by Jason Baugaertner and is available at 'Pushshift.io' (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, n.d.). Jason has collected all publicly available comments and submissions available on Reddit from 2006 until present, and regularly updates the scores of older comments and submissions. For the purpose of this research, a total of 28,521,660 comments and submission were collected from political subreddits from 1 August, 2016 to 28 February, 2017 (see Appendix 1). The list of political subreddits was taken from a list compiled by the user u/IranianGenius who is a moderator of the subreddit r/ListOfSubreddits – a subreddit responsible for classifying other subreddits on Reddit by topic of relevance. There are multiple academics that have already used the classifications of subreddits offered by this user (Nobles et al., 2019; Barros, Buitelaar, Duggan, & Rebholz-Schuhmann, 2019; Kang, 2019); however, this user is not responsible for mapping out all of Reddit, and there may be inaccuracies.

A total of 2,956,555 comments and submissions were made by accounts which have since been deleted and had their names changed to '[deleted]'. They account for 10.4% of the whole dataset. The deleted accounts do not have any kind of ID, which means they cannot be seperated into different users and classified by political orientation accordingly, thus they have been removed. Subreddits also tend to have so-called moderator bots as well as user bot accounts that perform different automated tasks. Moderator bots are typically bots which remove comments or submissions which aren't following the rules of that subreddit or make

certain automated remarks in the comment section under a submission (see appendix 2). 63 of users were identified as bots by searching for the comment 'I am a bot, and this action was performed automatically.' in all comments and submissions, and were also removed in addition to the deleted accounts. All 'moderator' bots include this phrase in their message as required per Reddit rules however not all user bot accounts have a single identifiable characteristic, thus those are left unidentified.

Identification Strategy:

The strategy used to identify Donald Trump and Hillary Clinton supporters is a modified version of the strategy used by Kwak, An, et al. (2019). In order to identify the supporters of Trump and Clinton, the subreddits r/The_Donald and r/hillaryclinton were looked at. r/The_Donald explicitly states that it is a subreddit for individuals that support Donald Trump and that critical comments directed at him or his policies will not be tolerated. r/hillaryclinton similarly states that it is a subreddit meant for Clinton supporters and any critical comments directed at Hillary Clinton will not be tolerated. If the interests of the subreddit community are aligned with the rules prescribed by that subreddit, then the comments seen as being in violation of the rules will be less popular and rejected by the community in the form of receiving down-votes from other users, decreasing the score of comment or submission. This would imply that comments critical of Trump or his policies would be rejected by r/The_Donald and would either be voted negatively and/or be removed by the moderator. The same applies to critical comments of Hillary Clinton made on r/hillaryclinton. By looking at the average scores across users in the 2 subreddits, this can be exploited to find Donald Trump and Hillary Clinton supporters by taking users with an average score above 1 (since all comments start with 1 up-vote) in one of the subreddits as being a supporter of that subreddit, and hence also being a supporter of the candidate that that subreddit is supporting. Users that have average comment scores above 1 across both subreddits have their average scores compared to the median comment scores of each subreddit, and are labelled as a supporter of the subreddit where their average comment score to subreddit median score ratio is higher. We are interested in users that are active supporters, thus, only users with more than 5 comments in r/The_Donald or r/hillaryclinton are considered. Following this strategy, I have identified 83,645 active Trump supporters and 6,011 active Clinton supporters. The comments and submissions of the active Trump supporters on r/hillaryclinton were found to have an average score of 3.0 which is significantly lower than the average of 6.0 on that subreddit. The comments and submissions of the active Hillary supporters on r/The_Donald were found to have an average score of 5.0 which is significantly lower than the average of 9.2 on that subreddit. This result is consistent with the assumption that supporters of the opposing presidential candidate will have opposing

opinions to that subreddit, hence are more likely to be rejected by that community in the form of getting lower than average scores. Let S_T, S_C, denote the sets of users which have been labelled as supporters of Donald Trump, Hillary Clinton, S_U, the set of unassigned users, and U, the universal set ($S_T \cup S_C \cup S_U$).

The method of identifying the supporters of each candidate inherently suffers from some level of endogeneity since there is a comment threshold in my definition of an active supporter. Given it takes time and effort to make a comment to any one subreddit, and a user has a finite amount of time and effort he's willing to spend, a user is willing to make only a certain amount of comments on all of Reddit. By increasing this comment threshold, we are potentially leaving in users which have dedicated a large portion of their time spent on Reddit to that one political subreddit, hence communicating mostly to members of that subreddit, thereby inflating the level of political homophily. I believe 5 comments over a period of 7 months is reasonable and introduces an insignificant level of endogeneity.

The removal of comments and submissions from the dataset which were made by users that have since deleted their accounts might have distorted the calculated homophily scores, depending on those users' distribution of political beliefs. If deleted users came from the same distribution of political leaning as users that are still present, then this would not affect the reported result, however there is reason to believe this is not the case. In both subreddits, the comments made by deleted users have a significantly lower average score than that of not deleted users (see appendix 3). This implies that the deleted users on r/hillaryclinton on average contain a different and less popular message to the comments made by users still present, implying that the distribution of deleted users across political leanings is different to that of present users. The same is implied for deleted users on r/The_Donald.

Upon further investigation, it appears that the average comment score made by a deleted user on r/hillaryclinton towards a Clinton supporter is the same as the average score of a comment made by a Trump supporter towards a Clinton supporter. This implies that the content of the messages of deleted users and Trump supporters towards Clinton supporters has the same level of support/rejection by the community of r/hillaryclinton. Similarly, on r/The_Donald, the average score of comments by deleted users towards Trump supporters is the same as the average score of comments made by Clinton supporters towards Trump supporters (see Appendix 4, Appendix 5). Accounts which are created to be deleted shortly after are called 'throw-away' accounts and are typically used in order to preserve anonymity as this technique removes the users past comment and submission history on Reddit , allowing the users to engage in political discourse without the context of their main account (Nagel, 2013). I hypothesise that a large portion of deleted users on r/hillaryclinton are Trump supporters using throwaway accounts to mask the identity of their main account. This would explain the same mean score. Similarly, deleted accounts on r/The_Donald may be largely

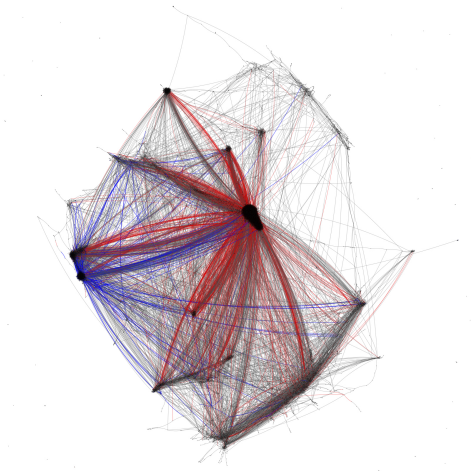


Figure 2: User interactions on election Day

Date: 8 August, 2016. Comments made by Trump supporters are colored red, and comments made by Clinton supporters are colored blue.

composed of Hillary Clinton supporters. In spite of this, the Kolmogorov-Smirnov tests on the comment score distributions reject these hypotheses (see Appendix 6). We proceed under the assumption that comments made by deleted users would not have a significant impact on the results. This is one of the main limitations in this paper and possible solutions to this problem are discussed in the limitations section.

Graph:

In order to perform analysis of political homophily and heterophily a directed multigraph $G(V, E)$ was constructed representing all user interactions from the dataset. For clarity, $V(G)$ represents the list of all vertices in G and $E(G)$ represents the list of all the edges $(u \to v)$ in G. Unique users from the dataset are represented by individual vertices in G, and interactions between users are represented by edges between those vertices. Thus, every vertex in G also belongs to one of the sets S_T, S_C, S_U. As an example, a comment made by some user u on a comment or submission made by another user v is represented in the graph as a directed edge $(u \to v)$.

Measure of Homophily:

In current literature, the concept of of heterophily is seldom used as it is often times defined as the opposite of homophily. This makes sense when there exist only 2 categories for political leaning (typically Liberals and Republicans, or Democrats and Republicans, or Left and Right). Given we have 3 categories for users (Trump supporters, Clinton supporters,

and unassigned users), both homophily and heterophily are useful in defining user behaviour.

The level of homophily for the set of users $S \in \{S_T, S_C\}$ is defined as the sum of all outgoing edges from users in set S to other users in the set S, divided by all outgoing edges from users in set S. Formally it is defined as:

$$\frac{\sum\limits_{\substack{(u \to v) \in E(G) \\ u,v \in S}} (u \to v)}{\sum\limits_{\substack{(u \to v) \in E(G) \\ u \in S}} (u \to v)}$$

The level of heterophily for the set of users $S \in \{S_T, S_C\}$ is defined as the sum of all outgoing edges from users in set S to users in the other set S', where $S \neq S'$, divided by all outgoing edges from users in set S. Formally it is defined as:

$$\frac{\sum\limits_{\substack{(u \to v) \in E(G) \\ u \in S, v \in S'}} (u \to v)}{\sum\limits_{\substack{(u \to v) \in E(G) \\ u \in S}} (u \to v)}$$

Both, heterophily and homophily range from 0 to 1 , and a result of 1 (absolute homophily) for a given group of users implies that every comment made by a member of that group is directed at the member within that very group, whereas 0 implies that every comment made by a member of that group is directed at someone outside that group.

In order to evaluate the computed levels of homophily and heterophily, we need to establish baseline levels to which we can compare these scores (Wasserman & Faust, 2012). In our graph, baseline homophily and heterophily is defined as the expected measure of homophily and heterophily given the absence of selective exposure, and is solely dependent on the distribution of users across the sets S_T, S_C, and S_U. It is calculated by copying the original graph, and relabeling every vertex as a Trump supporter, Clinton supporter, or as 'unassigned', drawing the labels from the distribution of political labels of the original graph. By repeating this process 100 times, we can calculate the probability distribution of random homophily and heterophily rates, the means, and standard deviations. To compare whether the calculated levels of homophily and heterophily differ significantly from the baseline levels, a standard 2-tail z-test is used. The values we get after subtracting baseline levels from homophily and heterophily levels, are referred to as 'relative homophily' and 'relative heterophily'.

For the second question we look at the changes in homophily and heterophily levels across time. For that, the original graph is broken down into 7 smaller ones, each graph representing the communications that took place in every month from August, 2016 to

	value	baseline	Z score
Trump Homophily	0.7039	0.0496	305.8128***
Trump Heterophily	0.0065	0.0026	7.4687***
Clinton Homophily	0.2693	0.0119	91.7034***
Clinton Heterophily	0.0701	0.0407	11.3580***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: Homophily and Heterophily scores, baseline, and significance test results for Trump and Clinton supporters on political subreddits from Aug '16 to Feb '17

February, 2017, and the same procedure is performed to derive the levels of homophily and heterophily above baseline levels of each month for supporters of Trump and Clinton

Results

The first research question was to evaluate the levels of political homophily and heterophily of Trump and Clinton supporters. From Table 2 we can see that the level of homophily found in the graph is 0.704 for Donald Trump supporters which is 0.653 greater than the baseline homophily of 0.050 and statistically significant. The level of homophily for Clinton supporters is 0.269 which is 0.257 greater than the baseline homophily of 0.012 and is statistically significant. This confirms the results of previous work on twitter, that politically active users as a whole tend to be more likely to interact with individuals with a similar opinion to them than someone that is not (Conover et al., 2016; Medaglia & Yang, 2016). The difference between relative homophily for Trump and Clinton supporters is 0.396 and statistically significant, which implies that Trump supporters are experiencing a higher level of relative political homophily than Clinton supporters (see Appendix 7). This is unexpected as previous literature highlights that left leaning users tend to experience higher levels of relative homophily (Colleoni et al., 2014; Pennacchiotti & Popescu, 2011). However, Trump supporters represent only a portion of Republicans and Clinton supporters represent a portion of Liberals, thus the two might possess different qualities and experience different behaviours.

Trump supporters had a heterophily score of 0.007 which is 0.004 greater than the baseline score of 0.003 and statistically significant. Clinton supporters had a heterophily score of 0.070 which is 0.029 greater than the baseline level of 0.041 and statistically significant. This implies that both Trump and Clinton supporters have exposed themselves to individuals with opposing beliefs more often than would have been expected in the baseline cases. The

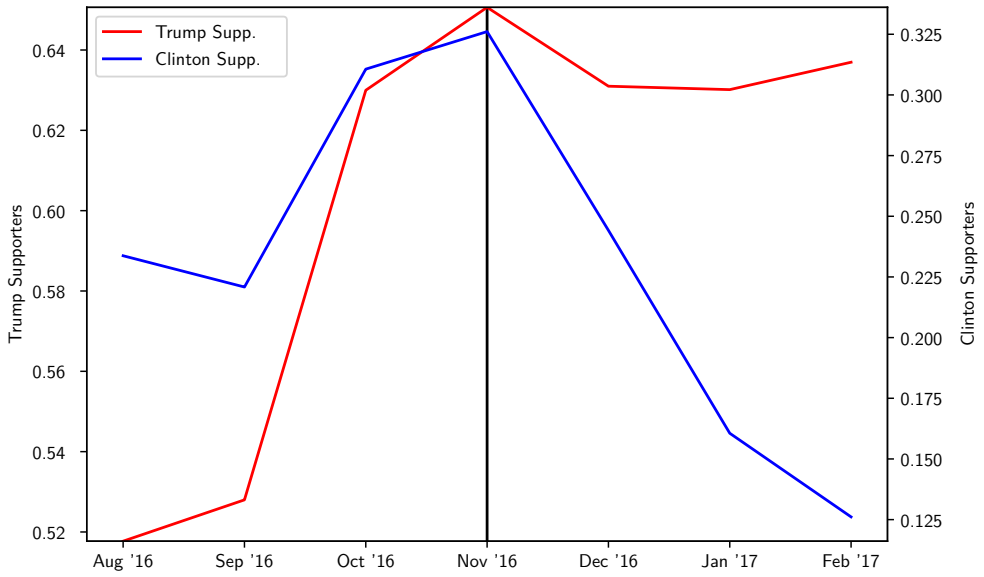


Figure 3: Relative homophily of Trump and Clinton supporters. The black line represents the month the US presidential election took place.

difference between levels of relative heterophily for Trump and Clinton supporters is -0.025, and statistically significant, which implies that Clinton supporters on Reddit are more open to engage in conversation with users with opposing beliefs than Trump supporters. There isn't evidence necessarily supporting or contradicting these results. But, in the paper by Kwak, An, et al. (2019), the authors found that Clinton Supporters are more likely to engage in political discourse using persuasive language in cross-cutting communication spaces, where as Trump supporters used persuasive language only against 'unassigned' users, and used more defensive language when interacting with Clinton supporters. Possibly the Clinton supporters willing to expose themselves to opposing arguments hence are more willing to engage in dialogue with Trump supporters, therefor being described as using using more persuasive language.

As mentioned in the Method section, there is a potential risk for endogeneity, so a robustness test was performed. The test consists of calculating homophily and heterophily levels for the graph with r/The_Donald removed when calculating scores for Trump supporters and r/hillaryclinton removed when calculating scores for Clinton supporters (see Appendix 8). The results show that Trump supporters exhibit significantly positive levels of relative homophily and heterophily outside of r/The_Donald and Clinton supporters exhibit significantly positive levels of relative homophily and heterophily outside of r/hillaryclinton. Thus, even if endogeneity was present, the conclusion that both Trump and Clinton supporters have significantly positive levels of relative homophily and heterophily, still holds. Although, as mentioned in the Method section, that shoudn't be the case.

The second question was to look at political homophily in the run up and after the US presidential election took place. After generating homophily statistics using 2 week, 3 week and 1 month long periods, I found that generating political homophily rates with the first 2 period lengths produced noisy data, so a period length of 1 month was used. From Figure 3 it can be seen that for Trump supporters, relative homophily kept increasing from August, 2016 up until November of 2016, where it peaked, then settled at a slightly lower level below the November peak for the following months until February, 2017. For Clinton supporters, relative homophily also kept increasing up until November, then consistently fell in the following months. Both Trump and Clinton supporters experienced a rise in relative homophily in the run up to the election, having an increased preference to interact with users with similar beliefs. As well as, the results on homophily from the first research question held for every month (see Appendix 9, Appendix 10).

Limitations

This paper has a number of limitations, many of which, are not uncommon to researching in online communications. The general problem is the inability to collect all interactions involving political discourse on Reddit. Starting with the deleted comments, I was unable to find a simple solution, thus I have to rely on the fact that they make up only a small portion of the data, and given the overwhelmingly significant results for relative homophily and heterophily levels for Trump and Clinton supporters, there's a smaller chance that the comments of deleted users would have had an impact on the conclusions. A possible solution to this problem would be implementing machine learning methods to classify the comments by ideology or political leaning. From there, the academic community will gain insight into the probability distribution of these comments across ideological leaning. Along these lines is also the limitation of not having collected all political data, as well as, collecting data which isn't necessarily linked to political discourse or perhaps there was no exchange of useful information but rather a hateful and unproductive argument between disagreeing individuals. It is likely that not every single relevant subreddit or comment was included into this dataset since the list of political subreddits used for this research was not thoroughly audited by an accredited institution. However, even if the list of subreddits was accurately compiled, there will still be certain places on Reddit where a political conversation can occur. Finally, there are other forms of interactions on Reddit apart from comments; like, leaving up-votes and down-votes, or even just reading and deliberating someones comments without contributing. Such actions can not be observed from the data available at hand. However for the latter, these users have been identified as 'lurkers', and there is evidence from a Polish forum website that despite the majority of these users being silent, they share similar political

values (Sobkowicz & Sobkowicz, 2012). As for the limitation concerning the incompleteness of the political dataset; as with the deleted comments, machine learning techniques can be employed to extract all political comments made on Reddit. Such methods have already been put into practice on twitter in order to identify political comments (Preoţiuc, Liu, Hopkins, & Ungar, 2017; Colleoni et al., 2014).

Conclusion

In this paper I analysed political discourse across many subreddits, finding that Trump and Clinton supporters experience significant levels of relative political homophily and heterophily, and that Trump supporters have a higher degree of relative homophily whereas Clinton supporters have a higher degree of relative heterophily. In my second research question I explored a method of tracking levels of political homophily and heterophily for 2 ideologically opposing groups of users across time and found consistency of the levels found for political homophily across the months spanning August to February. Supporters of both Trump and Clinton show behaviours akin to 'chamber' with the high levels of homophily, yet at the same time, there are individuals that make an effort to engage in conversation with users they disagree with, as shown with levels of heterophily.

For future research, it would be interesting to explore the reason behind why Trump supporters have a significantly higher degree of relative homophily than Clinton supporters and correlate the differences with the properties of graph structures of the different subreddits since some subreddits may promote open discussions whereas other might dampen it. But, more generally, to what extent does Reddit contribute to the user behaviour and the high levels of relative homophily and heterophily, and to what extent are ideological groups are responsible for these observations? As a final suggestion for a possible extension, it would interesting to study trends in homophily and heterophily levels using the method demonstrated in this research and perform event studies to derive a causal effect of certain politically charged events like elections or policy decisions.

Appendix

Appendix 1

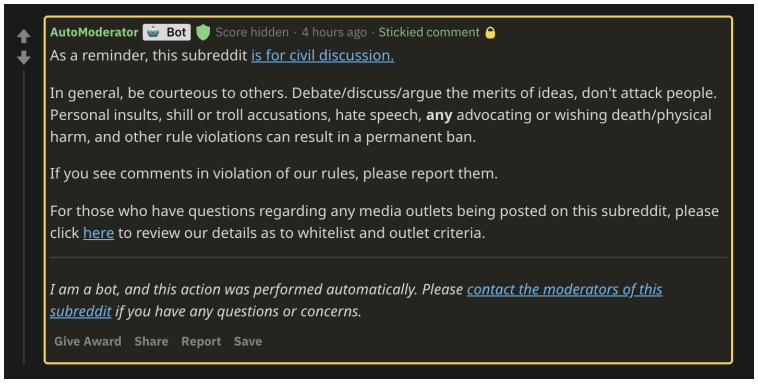
Subreddit	# Comments	# submissions	Total
The_Donald	7977562	990928	8968490
news	3688655	427713	4116368
politics	3723589	212957	3936546
ukpolitics	802791	20049	822840
PoliticalDiscussion	687326	13797	701123
hillaryclinton	552416	33848	586264
AskTrumpSupporters	359057	13242	372299
Libertarian	296262	16652	312914
Conservative	250720	22492	273212
MensRights	198801	10953	209754
EnoughTrumpSpam	113830	77855	191685
Political_Revolution	175416	15904	191320
AskThe_Donald	176858	9717	186575
LateStageCapitalism	176088	9552	185640
PoliticalHumor	145249	18553	163802
Anarcho_Capitalism	145827	9888	155715
HillaryForPrison	129485	21442	150927
socialism	134409	11336	145745
SandersForPresident	121152	5008	126160
WikiLeaks	117966	7250	125216
worldpolitics	93752	17831	111583
GaryJohnson	100823	9835	110658
Anarchism	82354	6772	89126
environment	52649	12189	64838
jillstein	45202	9241	54443
Bad_Cop_No_Donut	48632	4783	53415
NeutralPolitics	47278	2164	49442
geopolitics	44507	4242	48749
Republican	41848	4335	46183
democrats	27327	11666	38993
MarchAgainstTrump	35880	1115	36995
BasicIncome	30096	3515	33611

HillaryMeltdown	27825	2072	29897
COMPLETEANARCHY	26383	2804	29187
Fuckthealtright	26718	1559	28277
progressive	21117	6674	27791
Impeach_Trump	21785	4979	26764
gunpolitics	21370	1154	22524
Feminism	14577	4510	19087
esist	18039	996	19035
bidenbro	16963	1316	18279
Liberal	13738	4162	17900
communism	14645	2971	17616
conservatives	12607	4377	16984
new_right	1008	14558	15566
ThanksObama	14514	511	15025
BlueMidterm2018	13876	904	14780
Trumpgret	14359	394	14753
ChapoTrapHouse	13235	883	14118
Anarchy101	12503	989	13492
DebateCommunism	9712	466	10178
libertarianmeme	6650	943	7593
TinyTrumps	6422	1064	7486
HillaryForAmerica	4142	3127	7269
moderatepolitics	6165	621	6786
EndlessWar	2982	3082	6064
WomenForTrump	2143	2984	5127
GreenParty	3147	1628	4775
Objectivism	2035	934	2969
obama	1506	1265	2771
randpaul	2246	359	2605
labor	1192	1407	2599
DebateaCommunist	2391	39	2430
evolutionReddit	933	1426	2359
SocialDemocracy	1685	411	2096
alltheleft	1005	1080	2085
republicans	1200	639	1839
Green	1108	664	1772
leftcommunism	1326	312	1638

badgovnofreedom	366	1212	1578
ronpaul	1088	389	1477
monarchism	1145	215	1360
Good_Cop_Free_Donut	942	274	1216
antiwar	396	594	990
KasichForPresident	766	177	943
AnarchistNews	326	586	912
demsocialist	565	343	908
mcmullin	631	244	875
democracy	455	301	756
Agorism	519	171	690
CornbreadLiberals	171	483	654
futuristparty	415	175	590
LibertarianLeft	385	204	589
LibertarianSocialism	376	176	552
The_DonaldBookclub	393	112	505
paleoconservative	97	308	405
Green_Anarchism	177	174	351
LibertarianDebates	271	26	297
Trueobjectivism	219	73	292
christian_ancaps	235	52	287
BullMooseParty	225	61	286
greed	128	151	279
TedCruz	107	97	204
voluntarism	85	112	197
centrist	129	51	180
firstamendment	74	91	165
MartinOMalley	60	88	148
fairtax	117	25	142
Marco_Rubio	69	67	136
PirateParty	46	38	84
EnoughPaulSpam	54	19	73
nazbol	48	6	54
Romney	24	25	49
neoprogs	7	37	44
SelfAwarewolves	21	22	43
EnoughObamaSpam	2	15	17

Paul	4	12	16
BenCarson	1	14	15
ACTA	1	7	8
AnarchObjectivism	1	5	6
FlushTheTPP	0	3	3

Appendix 2



An example of an automated reply to a user submission made by the r/politics moderator bot

Appendix 3

	r/The_Donald	r/hillaryclinton
Deleted	4.87	1.99
Present	9.74	6.81

Average comment scores for deleted and present users

Appendix 4

Type of Comment	r/The_Donald	r/hillaryclinton
$Trump \rightarrow Clinton$	-	1.77
$Deleted \rightarrow Clinton$	-	1.84
$Clinton \rightarrow Trump$	5.39	-
$Deleted \rightarrow Trump$	5.39	-

Average comment scores for given comment type

Appendix 5

	Test Statistic	p-value
r/hillaryclinton	-0.282951899	0.77721471
r/The_Donald	-0.012380937	0.990121697

t-test for average score of $Trump \rightarrow Clinton = Deleted \rightarrow Clinton$ in r/hillaryclinton and $Clinton \rightarrow Trump = Deleted \rightarrow Trump$ in r/The_Donald

Appendix 6

	Test Statistic	p_value
r/hillaryclinton	0.170985928	1.03564E-61
r/The_Donald	0.205716063	2.0581E-49

Kolmogorov-Smirnov test for score of ($Trump \rightarrow Clinton$) having same dist. as ($Deleted \rightarrow Clinton$) in r/hillaryclinton and ($Clinton \rightarrow Trump$) having same dist as ($Deleted \rightarrow Trump$) in r/The_Donald

Appendix 7

	Z score
(Trump Hom - Clinton Hom)	112.3561526***
(Trump Het - Clinton Het)	1.567100052*

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Appendix 8

	Score	Baseline	Z score
Trump Homophily	0.117	0.046	30.788***
Trump Heterophily	0.039	0.003	69.091***
Clinton Homophily	0.072	0.011	32.58***
Clinton Heterophily	0.116	0.04	32.41***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Robustness tests: In calculating Trump Homophily and Heterophily levels, r/The_Donald was removed from the graph so as to look at user behaviour outside of that community. Similarly for Clinton Supporters, when measuring their scores, r/hillaryclinton was removed from the graph

Appendix 9

	Trump Hom			Clinton Hom		
	Value	Baseline	Z score	Value	Baseline	Z score
Aug '16	0.588	0.070	186.428***	0.248	0.014	101.290***
Sep '16	0.606	0.078	122.821***	0.235	0.014	83.932***
Oct '16	0.720	0.090	130.823***	0.328	0.017	90.169***
Nov '16	0.752	0.101	124.319***	0.343	0.017	108.084***
Dec '16	0.739	0.108	176.630***	0.259	0.014	69.793***
Jan '17	0.723	0.093	181.410***	0.174	0.013	43.087***
Feb '17	0.732	0.095	171.470***	0.139	0.013	37.626***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Appendix 10

	Trump Hom - Clinton Hom
	Z scores
Aug '16	78.178***
Sep '16	58.626***
Oct '16	56.245***
Nov '16	53.838***
Dec '16	74.934***
Jan '17	94.437***
Feb '17	94.262***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Tests for $TrumpHomophily - ClintonHomophily = 0$

References

Barros, J. M., Buitelaar, P., Duggan, J., & Rebholz-Schuhmann, D. (2019). Unsupervised classification of health content on reddit. In *Proceedings of the 9th international conference on digital public health* (p. 85–89). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3357729.3357745` doi: 10.1145/3357729.3357745

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (n.d.). *The pushshift reddit dataset.*

Becker, G. S. (1958). Competition and democracy. *Journal of Law and Economics*, *1*, 105-109.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*(2), 317–332.

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., & Menczer, F. (2016). Political polarization on twitter..

Danah, B., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1).

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298-320.

Gentzkow, M., & Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, *126*(4), 1799–1839.

Honeycutt, C., & Herring, S. C. (2019). Beyond microblogging: Conversation and collaboration via twitter..

Kang, B. N. Y. (2019). A quantitative exploration of repeated advice-seeking on reddit's r/relationships. *Social sciences::Psychology*, *10*(2).

Kwak, H., An, J., Posegga, O., & Jungherr, A. (2019). Political discussions in homogeneous and cross-cutting communication spaces..

Kwak, H., Lee, C., Park, H., & Moon, S. (2019). What is twitter, a social network or a news media?.

McPherson, M., Lovin, L. S., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415-444.

Medaglia, R., & Yang, Y. (2016). Online public deliberation in china: evolution of interaction patterns and network homophily in the tianya discussion forum. *Information, Communication Society*, *20*(5), 733-753.

Nagel, E. (2013). Faceless bodies: Negotiating technological and cultural codes on reddit gonewild. *Scan: Journal of Media Arts Culture*, *10*(2).

Nobles, A. L., Leas, E. C., Althouse, B. M., Dredze, M., Longhurst, C. A., Smith, D. M., & Ayers, J. W. (2019, 11). Requests for Diagnoses of Sexually Transmitted Diseases on a Social Media Platform. *JAMA*, *322*(17), 1712-1713. Retrieved from `https://doi.org/10.1001/jama.2019.14390` doi: 10.1001/jama.2019.14390

Pennacchiotti, M., & Popescu, A. (2011). Democrats, republicans and starbucks afficionados: User classification in twitter..

Peterson, T. (2019). *Reddit reports 30% increase in monthly active users, eclipsing twitter and pinterest.* https://marketingland.com/reddit-reports-30-increase-in-monthly-active-users-eclipsing-twitter-and-pinterest-272295.

Preoţiuc, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of twitter users..

Sobkowicz, P., & Sobkowicz, A. (2012). Properties of social network in an internet political discussion forum. *Advances in Complex Systems*, *15*(16).

Sponcil, M., & Gitimu, P. (2012). Use of social media by college students: Relationship to communication and self-concept. *Journal of Technology Research*.

Wasserman, S., & Faust, K. (2012). *Social network analysis: Methods and applications*. Cambridge University Press.