Homophily and Antisocial Behavior in Online Social Networks

Samuel Hansen, Vincent Woo, Chenye Zhu
Stanford University

Social networks, algorithmic news feeds, and personalized search engines have given rise to "echo chambers," in which users "amplify ideological segregation" via selective exposure to conforming opinions [23]. Although prior work has identified associations between echo chambers and negative behaviors such as spreading fake news, little is understood about how homophily in online networks affects antisocial behaviors such as trolling, bullying, and cursing. Our analysis constructs network representations from comments scraped from Reddit.com - the self-proclaimed "Front Page of the Internet" - to elucidate the relationship between homophily across subreddits and antisocial behaviors within them. We find that subreddits'Onnela clustering coefficients, node strengths, and sizes are associated with significantly higher negative comment sentiment, suggesting a link between homophily and antisocial behavior on Reddit.

Additional Key Words and Phrases: Homophily, social networks, sentiment analysis, community detection, antisocial behavior, echo chambers.

1. INTRODUCTION

In the third century B.C., Aristotle observed, people "love those who are like themselves" [22]. Indeed, this notion of homophily, the tendency of people to associate with similar individuals, is an age-old emergent behavior. Today, homophily remains ubiquitous in the largest forums of human interactions: online social networks. For instance, researchers have observed homophilous clustering behavior in friendship formation [19; 29], as well as interracial marriages [21]. Further studies have identified links between homophily and negative behaviors such as aggression in adolescent groups [1; 7; 16], racial discrimination [13], and group polarization [3].

Prior literature suggests technological changes such as online social networks and algorithmic news feeds have given rise to "echo chambers," in which individuals primarily consume content in accordance with previous views [23]. Alarmingly, evidence suggests echo chambers help facilitate the spread of fake news and ideological extremism, yielding "a constant us-vs.-them mentality [that] depersonalizes the holders of alternative views" [8]. Psychological studies have shown that echo chambers, which represent homophilous ideological groups, "reinforce partisans' biased views of their opponents," suggesting a relationship between homophily and negative behaviors [12]. However, little is understood about how such homophily in online networks affects antisocial behaviors such as trolling, bullying, and cursing.

In turn, we investigate the question, "How does homophily in online groups affect antisocial behavior?" We examine this question using data from Reddit.com because it has a natural structure of online groups called "subreddits." To do so, we created network structures in which nodes represent subreddits and edges represent the proportion of users that overlap across subreddits. Our analysis compares network properties (such as weighted clustering coefficient) to aggregated user behaviors within each subreddit to examine the relationship between homophily across subreddits and antisocial behaviors within them.

We aim to establish whether the link between homophily and antisocial behaviors documented by off-line psychological studies scales to the level of large online groups. Doing so will hopefully contribute to ongoing efforts to understand the correlates of echo chambers and negative social behaviors.

2. RELATED WORK

Our study draws from research involving the psychology of echo chambers, sentiment analysis of online groups, and network analysis of users interacting on social websites.

Prior work in computational social science suggests homophily plays a role in a variety of negative human social behaviors. For instance, in *The Spreading of Misinformation Online* [4], Vicario et al. investigated the effect of echo chambers on cascades of conspiracy theory stories on Facebook.com.

Vicario et al. collected data from Facebook and investigated diffusion and cascades of science news and conspiracy theories. They found that homogeneity in social groups on Facebook drives content diffusion, resulting in homogeneous clusters of users. These digital "echo chambers" caused by homogeneity were found to be vulnerable to the spread of misinformation. Susceptibility to misinformation spread is clearly a negative consequence of echo chambers, which invites further inquiry into other forms of antisocial behavior linked to homophily in social networks.

In an off-line study of social networks, Espelage et al. examined the effects of homophily among adolescents in *Examination of Peer-Group Contextual Effects on Aggression During Early Adolescence* [9]. The authors collected data on hundreds of students from different genders and ages, and also their social behaviors such as bullying. Using the friendship links volunteered by the students, the authors utilized the NEGOPY software to construct social networks, and categorize students with respect to homophily. In particular, those involved in cliques were classified as more homophilous than others. Homophily was found to be associated with fighting, bullying, and other forms of aggression. Espelage at al. also remark that homophily likely extends to other types of aggression. In our study, we examine online aggression and similar antisocial behaviors.

Inspired by studies suggesting a relationship between homophily and antisocial behaviors, we turned to reliable ways to quantify both homophily and antisocial behavior in online groups.

In particular, in *Antisocial Behavior in Online Discussion Communities*, Cheng et al. analyzed antisocial behaviors in online discussion forums (CNN.com, Breitbart.com, and IGN.com) by investigating the linguistic and behavioral features of banned users. To objectively characterize antisocial users, Cheng et al. compared the behavior of users who were eventually banned from the online community, termed Future-Banned Users (FBUs), to those who were Never-Banned Users (NBUs). To briefly summarize their findings, Cheng et al. found that FBUs write comments that are less readable, less positive, and more profanity-laced compared to NBUs.

Cheng et al.'s analysis serves as a useful springboard for our own analysis for two reasons: first, the authors demonstrated that the

Linguistic Inquiry and Word Count (LIWC) system is a reliable tool to measure antisocial behavior by quantifying negative sentiment and profanity in user comments. Further studies regarding narcissism and attention-seeking in social networks [5] and predicting "Dark Triad" personality traits on Twitter [28] have also used the LIWC system to compute negative sentiment as a proxy for antisocial behavior.

Second, Cheng et al. provide inspiration to study the correlates of antisocial behavior beyond mainstream sites, stating, "antisocial behavior may also differ in communities other than the ones we studied (e.g. small special-interest communities)" [2]. Our analysis aims to tackle this question by examining differences in antisocial behavior across self-selected interest groups on Reddit.com. The drivers of antisocial commenting may depend more on specific characteristics of different online communities. The comment sections of CNN.com, Breitbart.com, and IGN.com represent loosely-organized online communities, whereas sites like Reddit.com have specifically-organized subreddits, with possibly unique commenting norms. Thus, community context may influence the severity of antisocial behaviors in online forums.

Our analysis relies on constructing a network representation from raw Reddit comments that can reflect homophily across subreddits. A natural way to construct a network from comments is to represent each user as a node, and each directed edge as representing whether user i commented on something posted by user j. Olson et al. used this approach in *Navigating the massive world of reddit: using backbone networks to map user interests in social media*, to demonstrate a method to navigate primary topics of interest in a social network using the backbone of the Reddit network [17].

A similar concept was employed in in *User-Level Sentiment Analysis Incorporating Social Networks*, in which Tan et al. connected users in the Twitter network by their similarities, given by a combination of follows or direct references to other users. Their work attempts to predict user sentiment on a given topic by understanding the structure of the social network, and our investigation similarly attempts to predict negative behavior by understanding network structure.

Tan et al.'s idea that edges can encode similarity of interest relates to our own network representation, which treats edges between subreddits as the fraction of overlapping users between them (see methods section).

Taken together, prior literature regarding patterns of homophily in social groups, tools for quantifying antisocial behaviors, and representations of networks informed the methods we used for this study.

3. METHODS

3.1 Data Collection

Our dataset consists of summary statistics of Reddit comment data spanning 2014 [11]. Because many subreddits have small comment counts, only the top 2,048 subreddits (as ranked by total comment count) were included for analysis. Further, only active users (those that commented at least 50 times over the year) were included in the data set.

After applying these filters, Reddit comments were analyzed according to the Linguistic Inquiry and Word Count (LIWC) system, which counts the frequencies of words in a comment that belong to various lexical categories. Specifically, for each comment, the LIWC outputs "4 general descriptor categories (total word count, words per sentence, percentage of words captured by the dictionary, and percent of words longer than six letters), 22 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 32 word categories tapping psychological constructs (e.g., affect, cognition, biological processes), 7 personal concern categories (e.g., work, home, leisure activities), 3 paralinguistic dimensions (assents, fillers, non-fluencies), and 12 punctuation categories (periods, commas, etc.)" [29].

In turn, our final data set consisted of two parts:

(1) lists of summary statistics of lexical features derived from comments for each subreddit;
(2) lists of user IDs denoting which users commented in which subreddits over 13 four-week periods spanning 2014.

For each subreddit, the 2007 Linguistic Inquiry and Word Count (LIWC) system computed the fraction of words in comments from the 13 periods belonging to 64 linguistic categories, including positive and negative sentiment. We used the fraction of negative words in a subreddit as the metric representing antisocial behavior.

Lastly, Reddit flags its subreddits as either "safe for work" (SFW) or "not safe for work" (NSFW) according to whether it contains content inappropriate for workplace settings, such as pornography or lewd images and text. We hypothesized that patterns of negative behavior would differ between SFW and NSFW subreddits, so we scraped the list of all 718 NSFW subreddits from http://redditlist.com/nsfw to join into our final dataset.

3.2 Network Representation

To measure the effect of homophily on negative sentiment, we constructed network representations in which nodes are subreddits, and edges are weighted by the Jaccard similarity between subreddits. Jaccard similarity is defined as:

$$J(M,N) = \frac{|M \cup N|}{|M \cap N|} \tag{1}$$

where M and N are sets of users between two different subreddits. Intuitively, if the user set of subreddit M overlaps substantially with the user set of subreddit N, subreddits M and N are more similar. Prior literature supports this application of Jaccard similarity, which has been used to represent similarity of collaborative tags by users [14], proximity of co-authors in collaboration networks [15], and relatedness of web pages [26]. We apply Jaccard similarity to create two different network representations using subreddit user IDs.

3.2.1 *Network 1: Jaccard Similarity over the Year.* The first network representation includes edges representing the fraction of users that commented in two subreddits during any of the 13 periods. Let $X_{i,A}$ represent the set of active users for the i^{th} 4-week period for subreddit A. Then, the edge weight between subreddits A and B is:

$$weight(A,B) = \frac{|(\cup_{i\in[1,13]} X_{i,A}) \cap (\cup_{i\in[1,13]} X_{i,B})}{(\cup_{i\in[1,13]} X_{i,A}) \cup (\cup_{i\in[1,13]} X_{i,B})} \tag{2}$$

This metric defines the edges of our network: for any pair of subreddits where $weight(A,B) > 0$, an edge exists with edge weight $weight(A,B)$. This edge building process is computationally expensive, as it required $\binom{n}{2} \in O(n^2)$ operations, each requiring set operations, where n is the size of the set of subreddits.

The resulting network has a large number of edges with low edge weights, likely due to a small set of hyperactive users who are active on many subreddits (see Table 1 and Fig. 8).

3.2.2 *Network 2: Jaccard Similarity Averaged by Month.* For a more granular representation, we also generated a network based on monthly activity instead of activity over the year. Edge weights thus represent monthly average Jaccard similarity between two subreddits over the 13 periods. Whereas Network 1 captures subreddit similarity over the entire year, Network 2 reflects subreddit similarity averaged over monthly periods, which is perhaps more stable. Edge weights in Network 2 are defined as:

$$weight'(A,B) = \frac{\sum_{i\in[1,13]} \frac{|X_{i,A}\cap X_{i,B}|}{|X_{i,A}\cup X_{i,B}|}}{13} \tag{3}$$

This version of the weight calculation was computationally heavier because the 13 periods could not be preprocessed, and each computation was up to 13 times more expensive. To generate this network representation faster, we wrote scripts to split computation into several processes (up to the number of CPU cores on the machine) to parallelize the weight calculation. This is an embarrassingly parallel processing problem because weights between subreddits are only dependent on the data from the two subreddits, and the edges can be written in any order.

3.3 Metrics & Algorithms

A key step in our analysis was identifying a reliable way to measure homophily in our networks. Intuitively, we want to capture the degree to which each subreddit participates in tightly clustered groups of subreddits. If a subreddit is tightly clustered amongst other subreddits, we may infer its users are drawn to subreddits of a specific theme, rather than a wider ranges of themes. In short, tightly clustered subreddits can be considered more homophilous.

A natural measure of a nodes clustering is its clustering coefficient. However, because our network representations include weighted edges, we must consider formulations of clustering coefficients that take edge weights into account. After consulting previous literature [25] we identified two weighted clustering coefficient metrics that can represent homophily in our networks.

First, Barrat et al. [25] proposed the following weighted clustering coefficient metric:

$$\hat{C}_{i,B} = \frac{1}{k_i(k_i-1)} \sum_{j,k} \frac{1}{\langle w_i\rangle} \frac{w_{ij}+w_{ik}}{2} a_{ij}a_{jk}a_{ik} \tag{4}$$

where w_{ij} represents the edge weight between nodes i and j, $a_{ij} = 1$ if there is an edge between i and j (and 0 otherwise), $\langle w_i\rangle = \sum_i \frac{w_{ij}}{k_i}$, and k_i represents the degree of node i. The Barrat coefficient reflects how much of vertex strength is associated with adjacent triangle edges [25]. A subreddit with a high Barrat clustering coefficient has strong edge weights to neighbors with which it forms triangles. In the case of our network representations, we would interpret a tightly clustered subreddit as one that has high overlap in the fraction of users between neighboring subreddits. To obtain this metric for each subreddit, we wrote scripts to compute the degree and normalizing constant $\langle w_i\rangle$ for each subreddit, iterated over all edges in which the subreddit participated in triangles, and recorded the Barrat clustering coefficient as defined above.

Second, Onnela et al. [25] proposed an alternate clustering coefficient formulation:

$$\hat{C}_{i,O} = \frac{1}{k_i(k_i-1)} \sum_{j,k} (\hat{w_{ij}}\hat{w_{jk}}\hat{w_{ik}})^{\frac{1}{3}} \tag{5}$$

where $\hat{w}_{ij} = \frac{w_{ij}}{max(w)}$. Unlike the Barrat clustering coefficient, the Onnela metric reflects how large triangle weights are compared to network maximum [25]. In turn, the Onnela clustering coefficient encodes how clustered a subreddit is in relation to the whole network, whereas the Barrat coefficient reflects how clustered a subreddit is locally. We computed the Onnela coefficient using the weighted clustering coefficient function in the NetworkX library.

Furthermore, to understand the connectedness of each subreddit in the network, we computed an adapted version of degree centrality that accounts for weighted edges called "node strength." Some formulations of node strength simply describe the sum of its edge weights. However, we used the following metric proposed by Opsahl et al. [18] that accounts for both the degree and average edge weight of each node:

$$C_D^w = k_i^{(1-\alpha)} \times s_i^{\alpha} \tag{6}$$

where k_i is the degree of node i, s_i is the average edge weight of node i, and α is a tuning parameter. When $\alpha = 0$, the equation simply reflects the node's unweighted degree. If $\alpha = 1$, the equation only reflects the average edge weight. We used $\alpha = 0.5$ to obtain equal relative importance of the number of edges compared to edge weights.

Lastly, because past literature suggests a positive relationship [6] between group size and antisocial behavior, we computed each subreddit's "size" as the number of unique users that commented in that subreddit during any of the 13 four-week periods.

3.4 Community Detection

Exploratory data analysis revealed specific clusters of subreddits may possess different levels of negative sentiment. For instance, subreddits devoted to sports teams appeared frequently in the most negative subreddits (see results section). We hypothesized we would observe distinct communities in our network representations with different levels of negative sentiment. For instance, we may expect to observe a community of NSFW subreddits (e.g. r/ImGoingToHellForThis, r/watchpeopledie, r/MorbidReality) with higher negative sentiment than a community of SFW subreddits (e.g. /fitness, r/food, r/cats).

To identify these modular clusters within our networks, we executed two community detection approaches tailored to weighted graphs: 1) the Cluster Affiliation Model for Big Networks (BigCLAM) [30] , and 2) the Girvan-Newman algorithm [10].

3.4.1 *BigCLAM.* The Cluster Affiliation Model for Big Networks (or BigCLAM algorithm) is an affiliation graph model (AGM) based algorithm for detecting overlapping communities. It accounts for the fact that a subreddit can connect with multiple interest groups, thereby attracting different types of users. For instance, r/soccer connects with both sport-themed subreddits that attract sport enthusiasts, as well as with non-U.S. country-themed subreddits that attract non-American users. By executing the BigCLAM algorithm, we aimed to detect the community structures present in our network representations.

In addition to using the BigCLAM algorithm package in the Stanford Network Analysis Platform (SNAP) C++ library, we also re-implemented the package ourselves to dive deeper into the analysis.

3.4.2 *Girvan-Newman.* To ensure robustness in our community detection results, we also employed the Girvan-Newman algorithm, which is a hierarchical method used to detect communities in complex systems. It recursively divides up groups into sub-

Table I. : Summary of Node and Edge Counts by Network Representation

Network Representation	Defaults Filtered	Number of Nodes	Number of Edges
Network 1	No	2048	2,057,436
Network 1	Yes	1998	1,996,477
Network 2	No	2048	1,984,762
Network 2	Yes	1998	1,923,834

Network 1 corresponds to the representation based on Jaccard similarity computed over the year, whereas Network 2 corresponds to the representation based Jaccard similarity averaged by month.

groups by repeatedly removing edges with the highest "edge betweenness". The result of the algorithm is a dendrogram that indicates arrangement of clusters within the network. We selected this method because it accounts for undirected weighted edges by incorporating them into edge betweenness calculations.

We computed the clusters using the girvan_newman method in the NetworkX library that included a custom function to recursively remove edges with the largest weights.

4. RESULTS

Reddit categorizes 50 subreddits as "defaults," which are featured at the top of the main landing page of the site. New users are automatically subscribed to default subreddits and must explicitly search for and subscribe to non-defaults subreddits.

We excluded default subreddits in the computation of all aforementioned metrics out of concern they would yield skewed measures of homophily. For instance, the fraction of overlapping users between r/funny (a default) and r/startups (a non-default) does not capture the same notion of similarity given by the comparison of two non-default subreddits because users must explicitly opt-in to non-defaults.

We performed our analyses using the versions of Networks 1 and 2 that excluded default subreddits. In the following results, "Yearly Graph" refers to the version of Network 1 without defaults, and "Monthly Graph" refers to the version of Network 2 without defaults.

4.1 Clustering coefficient vs. Negative Sentiment

Given previous psychological evidence [1; 3; 7; 13; 16] supporting a link between homophilous groups and antisocial behavior, we hypothesized we would observe a positive relationship between our metrics of homophily (Barrat and Onnela clustering coefficients) and negative sentiment. Further, because NSFW subreddits are typically centered around lewd themes (e.g. r/watchpeopledie, r/JusticePorn), we expected to observe a stronger relationship between homophily and negative sentiment in NSFW subreddits.

First, to examine the relationship between the Barrat clustering coefficient and negative sentiment, we bucketed Barrat coefficient values into top and bottom quartiles and evaluated whether these quartiles exhibited significant differences in average negative sentiment. Mann-Whitney U-tests revealed no significant differences between Barrat coefficient quartiles in either the yearly or monthly graphs (see Discussion section for further interpretation).

Because we did not observe significant differences across quartile comparisons, we focused our analysis on the relationship between the Onnela clustering coefficient and negative sentiment.

The scatterplots in Fig. 1 include locally weighted regression (LOESS) smoothers depicting how negative sentiment changes for different Onnela clustering coefficient values and SFW-statuses. We observe that Onnela clustering coefficient values have a increasing relationship with negative sentiment in both the yearly and

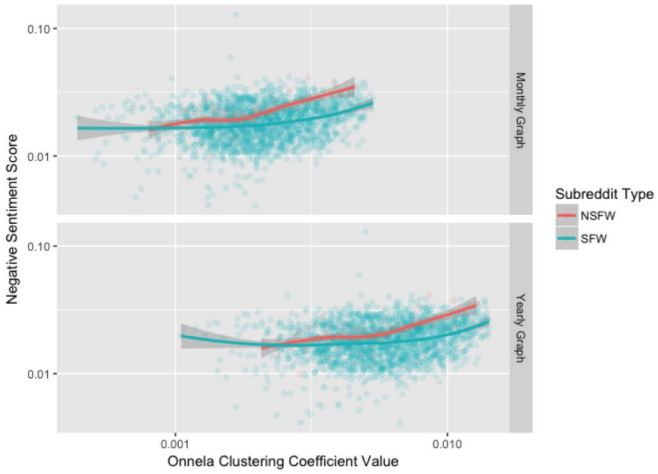


Fig. 1: Onnela clustering coefficient versus negative sentiment by graph type. "Yearly Graph" has edges representing the fraction of users who ever commented in two subreddits, and "Monthly Graph" has edges representing the monthly averaged fraction of overlapping users across subreddits.

monthly graphs. Because the LOESS smoothers are approximately linear, we fit linear regression models to determine whether this relationship is statistically significant. In the yearly graph, a 10% increase in Onnela clustering coefficient is associated with a 1.66% increase in negative sentiment ($p < 2 \times 10^{-16}$), and in the monthly graph, a 10% increase in Onnela clustering coefficient is associated with a 1.96% increase in negative sentiment ($p < 2 \times 10^{-16}$), on average. Furthermore, SFW subreddits are associated with significantly lower negative sentiment scores in both the yearly ($p < 5.11 \times 10^{-6}$) and monthly graphs ($p < 8.19 \times 10^{-7}$).

Although Fig. 1 reveals similar LOESS trends, clustering coefficient values are shifted rightward in the yearly graph, as compared to the monthly graph. We expect that this difference is due to how the graphs were constructed: the yearly graph represented edges as the fraction of users that *ever* commented in subreddits A and B, whereas the monthly graph represented edges as the fraction of users that commented in subreddits A and B during each 4 week period, averaged over all 13 periods. In turn, the yearly graph would have greater edge weights, on average, thereby shifting the distribution of Onnela clustering coefficients into higher values.

We supplemented our regression analysis by bucketing Onnela clustering coefficient values into top and bottom quartiles and evaluating whether these quartiles exhibited significant differences in average negative sentiment. Fig. 2 depicts the differences in average negative sentiment between the top and bottom quartiles of Onnela clustering coefficient, faceted by SFW status and network representation. Mann-Whitney U-tests revealed the top Onnela coef-

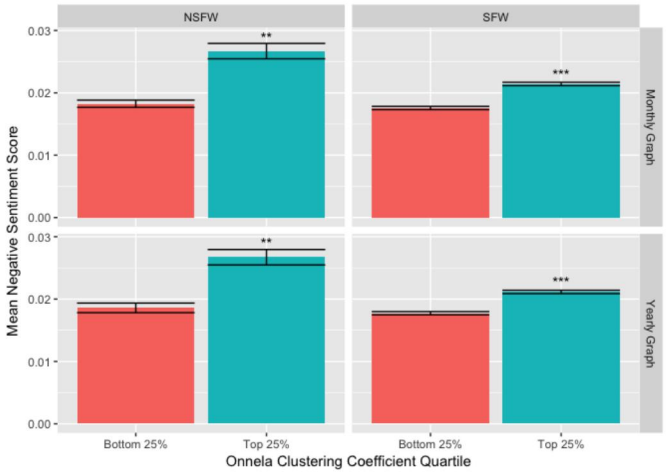


Fig. 2: Clustering coefficient quartile versus negative sentiment by graph type and SFW status. Statistical comparisons are across quartiles within each graph type and SFW status. ** = $p < 1 \times 10^{-7}$; *** = $p < 2 \times 10^{-16}$.

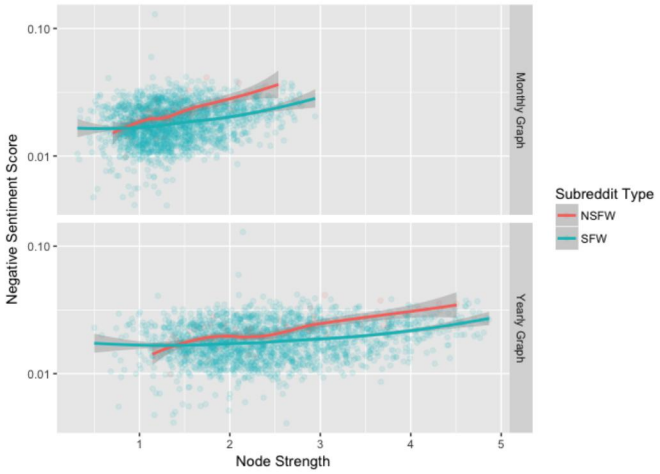


Fig. 3: Node strength versus negative sentiment by graph type.

ficient quartile has significantly higher negative sentiment than the bottom quartile for NSFW subreddits ($p < 6.09 \times 10^{-8}$) and SFW subreddits ($p < 2 \times 10^{-16}$) in the monthly graph and for NSFW subreddits ($p < 2.26 \times 10^{-9}$) and SFW subreddits ($p < 2 \times 10^{-16}$) in the yearly graph. Overall, these results suggest that higher Onnela clustering coefficients are associated with higher negative sentiment, especially for NSFW subreddits ($p < 1.21 \times 10^{-4}$).

4.2 Node Strength vs. Negative Sentiment

Node strength, a weighted version of degree centrality, captures how connected a subreddit is to others in a network. Building upon our analysis of clustering coefficient, we examined the relationship between node strength and negative sentiment in both the yearly and monthly graphs.

First, as depicted in Fig. 3, we observed an increasing relationship between node strength and negative sentiment in both network representations. The LOESS smoothers appear linear, so we fit linear regression models to evaluate the significance of this trend. In turn, we found a 10% change in node strength is associated with 2.28% and 2.58% average increases in negative sentiment in the yearly graph and monthly graphs, respectively (both p's$< 2 \times 10^{-16}$). Furthermore, SFW subreddits are associated with lower negative sentiment in the yearly ($p < 2.41 \times 10^{-5}$) and monthly ($p < 3.69 \times 10^{-5}$) graphs, implying the relationship between node strength and negative sentiment is stronger amongst NSFW subreddits.

Fig. 3 further reveals that the yearly graph is associated with wider spread of node strength values, as compared to the monthly graph. Again, this difference is likely to due the fact that the yearly graph is constructed in such a way that edge weights are larger than those in the monthly graph, on average. Because node strength is a function of edge weight, the larger spread of node strength values in the yearly graph is to be expected.

Given that segments of the LOESS smoothers in Fig. 3 reflect nonlinearities, we sought to make the our regression inferences more robust by bucketing node strength into top and bottom quartiles and evaluating whether these quartiles exhibit significant differences in average positive and negative sentiment.

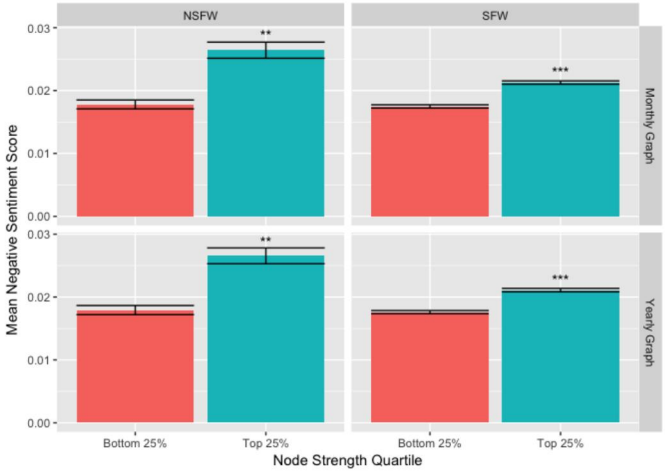


Fig. 4: Node Strength quartile versus negative sentiment by graph type. Statistical comparisons are across quartiles within each graph type and SFW status. ** = $p < 1 \times 10^{-7}$; *** = $p < 2 \times 10^{-16}$.

As shown in Fig. 4, Mann-Whitney U-tests confirmed negative sentiment is significantly higher in the top quartile of node strength in SFW ($p < 2 \times 10^{-16}$) and NSFW ($p < 6.09 \times 10^{-8}$) subreddits in the monthly graph and in SFW ($p < 2 \times 10^{-16}$) and NSFW ($p < 2.365 \times 10^{-8}$) subreddits in the yearly graph. As with the regression results, the difference between mean negative sentiment is larger in NSFW subreddits than in SFW subreddits, suggesting a significant interaction between SFW status, node strength, and negative sentiment ($p < 2.11 \times 10^{-4}$). Taken together, these results suggest that subreddits that are more strongly connected to other subreddits are associated with significantly higher negative sentiment.

4.3 Subreddit Size vs. Negative Sentiment

Exploratory data analysis suggested clustering coefficient and node strength may also depend on subreddit size, defined as the count of unique users who ever commented in a subreddit during the year

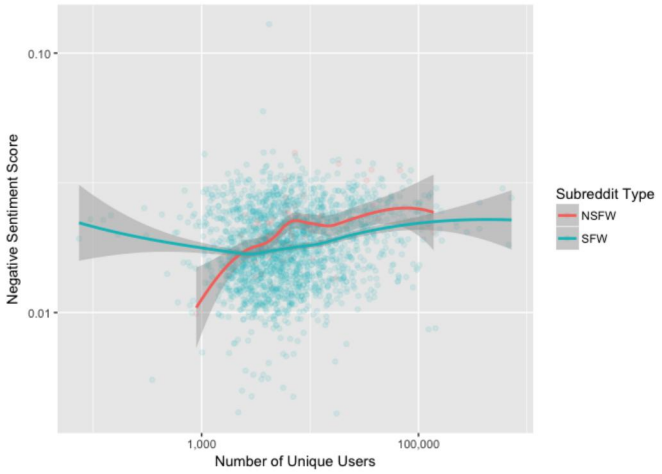


Fig. 5: Subreddit size versus negative sentiment by SFW status.

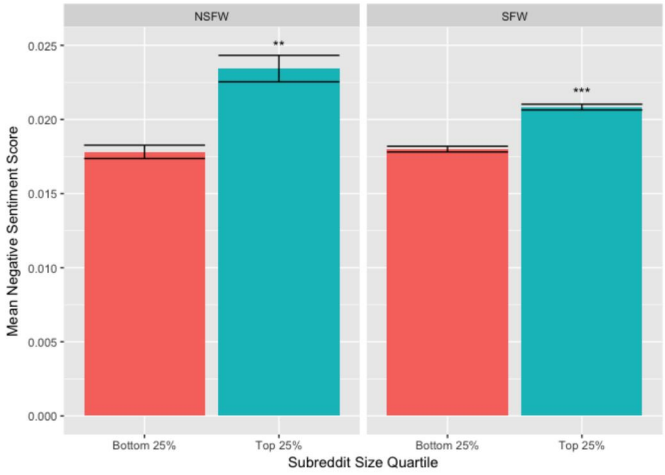


Fig. 6: Subreddit size quartile versus negative sentiment by SFW status. Statistical comparisons are across quartiles within each SFW status. ** = $p < 6.962 \times 10^{-5}$; *** = $p < 1.156 \times 10^{-13}$.

for which data were collected. In turn, we also evaluated the relationship between subreddit size and negative sentiment.

First, we examined locally-weighted regression smoothers in Fig. 5, which revealed increasing, yet nonlinear relationships between subreddit size and negative sentiment. Fig. 5 suggests negative sentiment increases slightly faster for NSFW subreddits than for SFW subreddits in middle values of size.

Since the relationships in Fig. 5 are nonlinear, we forewent a linear regression analysis and instead bucketed subreddit size into top and bottom quartiles to assess differences between interquartile mean negative sentiment scores.

Mann-Whitney U-tests revealed the top size quartile is associated with significantly higher negative sentiment in NSFW ($p < 6.962 \times 10^{-5}$) and SFW ($p < 1.156 \times 10^{-13}$) subreddits. The difference in mean negative sentiment between the top quartiles of SFW and NSFW subreddits is not significant ($p = 0.136$), which is consistent with the fact that the error regions of the LOESS smoothers overlap in the top and bottom quartiles of subreddit size in Fig. 5.

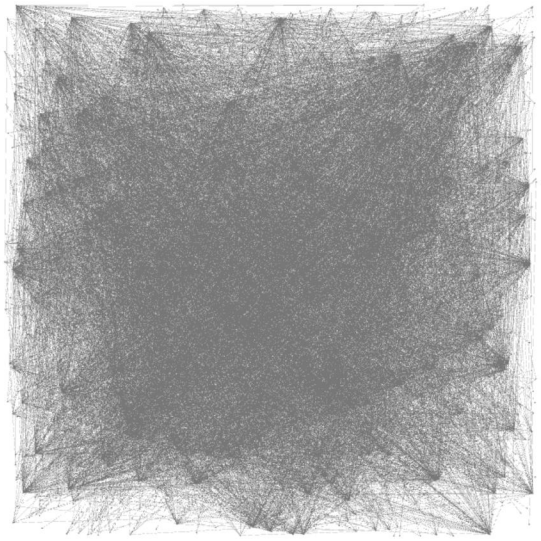


Fig. 7: Community representation detected by the BigCLAM algorithm.

4.4 Community Detection

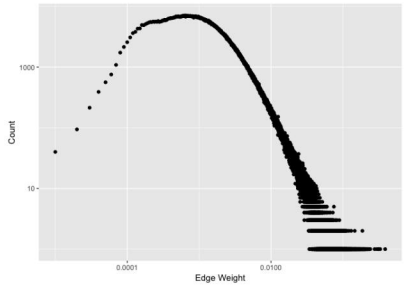
To perform community detection, we executed both the BigCLAM and Girvan-Newman algorithms using the yearly graph filtered to include edge weights greater than or equal to the 90th percentile. Both algorithms detected a single gigantic cluster in the network (with few smaller clusters), suggesting dense interconnectedness between subreddits overshadows weak, if any, sub-modular communities.

4.4.1 *BigClam.* We ran the BigCLAM community detection algorithm on the yearly graph filtered to include edges in the top 10% of edge weights. Surprisingly, both our own implementation of the BigCLAM algorithm and the associated SNAP C++ package resulted in a massive interconnected component (see Fig. 7).

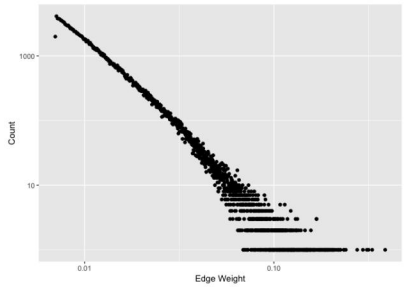
Even though the BigCLAM algorithm detected 374 communities within the network (which is lower than the upper bound of 400 expected communities), the strength of membership for any subreddit to be in any cluster is significant[1], as shown in the Fig. 7. In fact, the average subreddit belongs to more than 300 communities, which does not support the existence of distinct modular clusters of subreddits. We would describe this network as strongly interconnected with weak signs of sub-modular interest groups.

We interpret this result as the consequence of Reddit's massive user base. With 243.63 million monthly active users on Reddit [24], for any pair of subreddits, it is almost always possible to find a small group users who have co-commented in both. This is particularly true among the 1,998 largest non-default subreddits in our network. Therefore, the edge weight distribution in the networks are skewed towards the lower end of the spectrum (see Fig. 8b), with many edges possessing weights less than 0.01. Additionally, the distribution of the top 10 percent of edge weights effectively follows a power law distribution in Fig. 8b, with a significant count of edges near the lower spectrum of edge weights. These loosely connected edges dominate the graph structure and overshadow the existence of sub-modular interest groups in the network.

[1] In fact, the average score for F_{uv} is above 600.



(a) Network includes all edges present in the yearly graph.



(b) Network only includes edges with weights greater than or equal to the 90^{th} percentile.

Fig. 8: Histogram of edge weights computed from the yearly graph, which encode the fraction of users that ever commented in two subreddits during the year.

4.4.2 *Girvan-Newman.* Similarly, the results from the Girvan-Newman algorithm support our theory that densely interconnected nodes overshadow the weak structure of interest groups. By analyzing the dendrogram produced by the Girvan-Newman algorithm, we found that the yearly graph filtered to include edge weights above the 90^{th} percentile consists of one giant cluster that covers more than 85% of the subreddits (with 1,743 nodes in total), surrounded by several small clusters.

Ultimately, since both community detection algorithms did not identify distinctive sub-modular structures in our network, we could not proceed with community-specific analyses of negative sentiment.

5. DISCUSSION

5.1 Interpretation of Results

Our results tell a cohesive story: Onnela clustering coefficient, node strength and subreddit size are associated with higher negative sentiment. In other words, negative subreddits tend to be more tightly clustered, strongly connected to other subreddits (in terms of user overlap), and contain more users. Further, we observe an interaction effect with NSFW subreddits such that the relationship between Onnela clustering coefficient and node strength is significantly stronger in NSFW subreddits as compared to SFW subreddits. Taken together, these results support our hypothesis that homophilous subreddits (i.e. those with tighter clustering) are associated with more antisocial behavior, as measured by negative sentiment.

These patterns of negative behavior may be attributable to the online disinhibition effect (ODE), coined by cyberpsychologist John Suler as the tendency of online users to "act out more frequently or intensely than they would in person" [27].

Theories of deindividuation attribute "loss of individuality [to] submergence in the crowd" [20], suggesting that a variety of group properties such as size and cohesion may affect antinormative behavior. In Diener et al.'s study, *Effects of Deindividuation Variables on Stealing Among Halloween Trick or Treaters* , the authors demonstrated that the interaction between group size and anonymity produced the highest stealing rates of candy from bowls labeled "take one only" [6]. In turn, the researchers suggested self restraint decreases in larger anonymized groups.

Reddit is an ecosystem of pseudoanonymous individuals (masked by non-personally revealing user names) posting, commenting, and voting within groups. In turn, the positive relationship between group size and negative sentiment is consistent with predictions from psychological literature: larger groups afford more anonymity to users, who in turn, are more likely to behave in disinhibited, negative ways.

Our analysis excludes the default subreddits, meaning this association between subreddit size and negative sentiment is not an artifact of default groups, but rather of those to which users voluntarily subscribe.

Standard Pearson correlation analyses revealed subreddit size is correlated with the Onnela clustering coefficient ($\rho = 0.351$) and node strength ($\rho = 0.416$), suggesting their increasing relationships with negative sentiment may be partly due to the tendency of larger subreddits to be more strongly connected to and tightly clustered amongst other subreddits.

However, these observations are correlational, not causal, implying further research is needed to tease apart the interplay between how a subreddit's clustering, node strength, and size affect its negative sentiment.

The interaction between clustering coefficient, node strength, subreddit size, and negative sentiment may also be mediated by another factor: subreddit-theme. A simple inspection of the top 10 most negative NSFW and SFW subreddits (see Fig. 9) suggests specific themes may contribute to negative sentiment. For instance, the top NSFW subreddits in Fig. 9 tend to be associated with graphic images and videos (r/bestofworldstar, r/watchpeopledie, r/MorbidReality.) and pornography (r/ConfusedBoners, r/celebnsfw, r/MassiveCock). The top SFW subreddits tend to be associated with sports (r/rangers, r/falcons, r/miamidolphins.) and antisocial themes (r/Anxiety, r/PublicFreakout, r/fifthworldproblems).

Our analysis attempted to account for subreddit theme by examining the interaction between SFW status, negative sentiment, and the network metrics we computed. However, the binary variable of SFW status could be broken into granular subreddit themes, such as sports, pornography, and violence, among others. Doing so would provide more nuanced insight into the interaction between subreddit theme, clustering (our measure of homophily), and negative sentiment.

5.2 Tests of Robustness

To ensure that our results were not sensitive to the exclusion of default subreddits, we repeated all comparisons of negative sentiment against Onnela clustering coefficient, node strength, and subreddit size using the yearly and monthly network representations that included default subreddits. All p-values and regression coefficients were within the same order of magnitude as those produced

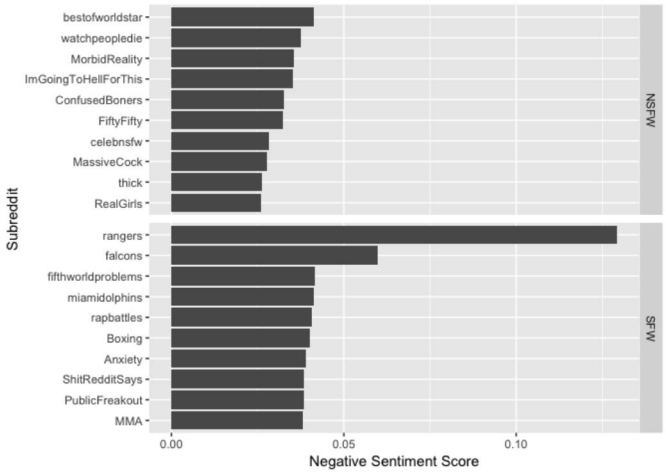


Fig. 9: Lists of top 10 most negative NSFW and SFW subreddits.

by the analysis that excluded defaults. Overall, our inferences do not change when defaults are included.

Furthermore, we created two network representations (i.e. the "yearly" and "monthly" graphs) to reduce the sensitivity of our results to arbitrary network constructions. Although we observe minor differences in regression coefficients and p-values, the significant patterns are the same across both graph types.

5.3 Sources of Error

As an important caveat, we did not observe a significant relationship between the Barrat clustering coefficient and negative sentiment. This is likely because the Barrat and Onnela coefficients capture different notions of clustering: the Barrat coefficient "reflects how much of vertex strength is associated with adjacent triangle edges," whereas the Onnela coefficient "reflects how large triangle weights are compared to network maximum" [25]. In other words, the Barrat coefficient encodes local clustering by focusing on the weights of a node's adjacent edges. However, the Onnela coefficient normalizes adjacent edge weights by the network's maximum, thereby representing each node's clustering in relation to the whole network.

Local clustering captured by the Barrat coefficient is likely skewed by the densely overlapping edges in our network representations. Since we only included users that commented at least 50 times during the year, it is possible that the Barrat coefficient does not reveal a relationship with negative sentiment due to exclusion of infrequent users. Further, since we included the top 1,998 subreddits for analysis, it is possible that uncommon subreddits may express different patterns between the Barrat clustering coefficient and negative sentiment. But despite these possible sources of error, the Onnela coefficient, which encodes the clustering of a subreddit relative to the whole network, is a more appropriate metric of homophily in our study.

Another potential source of error may stem from the fact that we used the yearly graph with filtered edges to perform community detection. Because the complete yearly graph has 2,057,436 edges, we were limited by computational power against executing the Girvan-Newman algorithm (which is implemented in non-optimized NetworkX) on the unfiltered graph. We chose to execute both community detection algorithms on the same graph for adequate comparison. However, these algorithms may identify more distinct communities in network representations without filtered edges.

5.4 Future Work

Although different mechanisms may be giving rise to our results (such as the online disinhibition effect or subreddit theme-specific variation), our analysis establishes a significant relationship between homophily and antisocial behavior on Reddit.com. This finding is consistent with prior work that has documented a relationship between homophily and negative behavior in off-line groups [1; 7; 9; 16]. Our analysis could be extended in the following ways:

5.4.1 *Topic Modeling.* First, researchers could perform topic modeling on raw comments to extract overarching themes of subreddits, and subsequently assess how the relationship between homophily and negative sentiment changes across different themes. In effect, this would build upon our interaction analysis of SFW status.

5.4.2 *Cosine Edge Weights.* Second, future researchers could redefine the edge weights in the graph to obtain a different network representation. For instance, edge weights could incorporate cosine similarity, where:

(1) Each subreddit S is represented by a long vector v^S of length N, where N stands for the total number of subreddit users.
(2) The i^{th} entry of the vector v^S equals 0 if user i has never commented on subreddit S, and equals $\frac{1}{n_i}$ otherwise, where n_i represents the total number of subreddits in which user i ever commented.
(3) The edge weight between two subreddits A and B is defined as the cosine similarity of the two corresponding vectors v^A and v^B:

$$w(A,B) = \frac{v^A \cdot v^B}{|v^A| \cdot |v^B|} \tag{7}$$

This formulation of edge weights is better suited with the hypothesis that a person who is exposed to a broader range of topics (represented by the number of subreddits on which he or she has commented) should contribute less to homophilous behaviors in the Reddit community. By properly discounting each person's contribution by the number of subreddits he or she commented on (n_i), we may better represent user commenting behavior in the construction of Reddit network representations.

In conclusion, it is our hope that this analysis of homophily and antisocial behavior in online social networks sparks further inquiry concerning the mechanics of echo chambers in digital communities.

6. INDIVIDUAL CONTRIBUTIONS

This section lists the individual contributions of the group members to this project.

6.1 Samuel Hansen

—Formulated research problem.
—Conducted literature review for project motivation (1) and interpretation of results (5.1) sections.
—Wrote all sections of final report except sections 4.4, 5.4.2, half of sections 2, 3.2, 3.4, and references.
—Researched and chose appropriate network metrics to compute.

—Performed all data analysis, including regression, correlation, and quartile analyses in R.

—Plotted all graphs (except BigCLAM output).

—Wrote script to compute node strength.

—Executed Girvan-Newman community detection algorithm.

—Performed sensitivity analysis on graphs that included default subreddits.

—Helped edit poster.

6.2 Vincent Woo

—Wrote half of section 2 and section 3.2.

—Wrote script to create edge representation from initial data, for yearly and monthly data.

—Wrote script to perform various filtering of yearly and monthly data.

—Wrote script to compute Onnela coefficients for all graphs.

—Created the presentation poster.

—Collected all references and inserted bibliography and citations.

6.3 Chenye Zhu

—Wrote section 3.4 and section 4.4 on Community Detection.

—Wrote section 5.4.2 on Cosine Edge Weights.

—Implemented python packages to compute Barrat clustering coefficients described by equation (4), with extensive support for parallel computing.

—Calculated the Barrat clustering coefficients for all graphs

—Implemented packages for BigCLAM algorithm in python and validated the results with SNAP C++ BigCLAM implementation.

—Analyzed the results for community detection. Hypothesized and validated the possible explanation of the observation.

—Conducted explorative research work with cosine edge weights using equation (7).

ACKNOWLEDGMENTS

We are grateful to the following people for resources, discussions and suggestions: Will Hamilton, Jure Leskovec, and Leon Yao.

REFERENCES

[1] R. B. Cairns, B. D. Cairns, H. J. Neckerman, S. D. Gest and J.-L. Gariepy, "Social Networks and Aggressive Behavior: Peer Support or Peer Rejection?," Developmental Psychology, vol. 24, pp. 815-823, 1988.

[2] Cheng, J., Danescu-niculescu-mizil, C., & Leskovec, J. (2015). "Antisocial Behavior in Online Discussion Communities."

[3] P. Dandekar, A. Goel and D. T. Lee, "Biased assimilation, homophily, and the dynamics of polarization," Proceedings of the National Academy of Sciences, vol. 110, no. 15, pp. 5791-5796, 2013.

[4] Del, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., & Caldarelli, G. (2015). "The spreading of misinformation online." https://doi.org/10.1073/pnas.1517441113

[5] Dewall, C. N., Buffardi, L. E., Bonser, I., & Campbell, W. K. (2011). "Narcissism and implicit attention seeking : Evidence from linguistic analyses of social networking and online presentation. Personality and Individual Differences", 51(1), 5762. https://doi.org/10.1016/j.paid.2011.03.011

[6] Diener, Edward et al. Effects of Deindividuation Variables on Stealing among Halloween Trick or Treaters. Journal of Personality and Social Psychology 33.2 (1976): 178183. Web.

[7] T. J. Dishin, G. R. Patterson, M. Stoolmiller and M. L. Skinner, "Family, School, and Behavioral Antecedents to Early Adolescent Involvement With Antisocial Peers," Developmental Psychology, vol. 27, pp. 172-180, 1991.

[8] Emba, Christine (2016, July 14). "Confirmed: Echo chambers exist on social media. So what do we do about them?" Retrieved from https://www.washingtonpost.com/news/in-theory/wp/2016/07/14/confirmed-echo-chambers-exist-on-social-media-but-what-can-we-do-about-them/

[9] Espelage, D. L., Holt, M. K., & Henkel, R. R. (2003). "Examination of Peer-Group Contextual Effects on Aggression During Early Adolescence", 74(1), 205220.

[10] Girvan, M., & Newman, M. E. J. (2002). "Community structure in social and biological networks", 99(12).

[11] Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora."

[12] Iyengar, S. (2012). "Affect , Not Ideology A Social Identity Perspective on Polarization." https://doi.org/10.1093/poq/nfs038

[13] N. Jacquemet and Y. Constantine, "Indiscriminate Discrimination: A Correspondence Test for Ethnic Homophily in the Chicago Labor Market," Labour Economics, vol. 19, no. 65, pp. 824-832, 2012.

[14] Markines, Benjamin et al. Evaluating Similarity Measures for Emergent Semantics of Social Tagging. Proceedings of the 18th international conference on World Wide Web, WWW 09, ACM (2009): 641650. Web.

[15] McKerlich, Ross, Cindy Ives, and Rory McGreal. Measuring Use and Creation of Open Educational Resources in Higher Education. International Review of Research in Open and Distance Learning 14.4 (2013): 90103. Web.

[16] K. C. Monahan, L. Steinberg and E. Cauffman, "Affiliation With Antisocial Peers, Susceptibility to Peer Influence, and Antisocial Behavior During the Transition to Adulthood," Developmental Psychology, vol. 45, no. 6, pp. 1520-1530, 2009.

[17] Olson RS, Neal ZP. "Navigating the massive world of reddit: using backbone networks to map user interests in social media". PeerJ Computer Science(2015) 1:e4 https://doi.org/10.7717/peerj-cs.4

[18] Opsahl, Tore, Filip Agneessens, and John Skvoretz. Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. Social Networks 32.3 (2010): 245251. Web.

[19] H. H. Park, R. K. Rethemeyer, K. Bryce, D. Andersen and H. Kim, "Making Friends and Influencing Careers: Social Integration, Homophily, and Cohort-Wide MPA Courses," Journal of Public Affairs Education, pp. 417-445, Summer 2011.

[20] Postmes, Tom, and Russell Spears. Deindividuation and Antinormative Behavior: A Meta-Analysis. Psychological Bulletin 123.3 (1998): 238259. Web.

[21] G. Potarca and M. Mills, "Partner Preferences and Racial Homophily in Online Dating: A Cross- National Comparison," in Population Association of America 2012 Annual Meeting, San Francisco, 2012.

[22] A. and H. Rackham, "The Nicomachean ethics", Cambridge, MA.: Harvard University Press, 1962.

[23] Rao, J. M. (2016). "FILTER BUBBLES , ECHO CHAMBERS , AND ONLINE NEWS CONSUMPTION", 80, 298320. https://doi.org/10.1093/poq/nfw006

[24] Reddit. (n.d.). Number of unique visitors to Reddit from July 2012 to April 2016 (in millions). In Statista - The Statistics Portal. Retrieved December 11, 2016, from https://www.statista.com/statistics/443332/reddit-monthly-visitors/.

[25] Saramki, Jari et al. Generalizations of the Clustering Coefficient to Weighted Complex Networks. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 75.2 (2007): 25. Web.

[26] Strehl, Alexander, Joydeep Ghosh, and Raymond Mooney. Impact of Similarity Measures on Web-Page Clustering. In Workshop on Artificial Intelligence for Web Search (AAAI 2000) (2000): 5864. Web.

[27] Suler, John. "The Online Disinhibition Effect." CyberPsychology & Behavior 7.3 (2004): 32126. Web.

[28] Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets", 386393. https://doi.org/10.1109/ICMLA.2012.218

[29] S. Utz and J. Jaroslaw, "Making "Friends" in a Virtual World The Role of Preferential Attachment, Homophily, and Status," Social Science Computer Review, pp. 1-21, 2015.

[30] Yang, J. (2013). "Overlapping Community Detection at Scale : A Non-negative Matrix Factorization Approach."