Mapping the Online News Environment: Leveraging survey and web-tracking data for audience networks *

Hertie School
Master in Public Policy
2020

Master's thesis

May 25, 2020

Author:
Sebastian Ramirez Ruiz
Hertie School
s.ramirez-ruiz@mpp.hertie-school.org

Supervisor:
Prof. Dr. Simon Munzert
Hertie School
munzert@hertie-school.org

*The replication code for this Master's thesis can be found in this Github repository.

ACKNOWLEDGEMENTS

En primera instancia, quisiera agradecer a mi familia por su apoyo incondicional. Especialmente, a mis padres, a quienes admiro por su constancia, principios, y dedicación. Mami y papi, ustedes son mis modelos a seguir. I would also like to express my gratitude to my supervisor, Prof. Dr. Simon Munzert, for sparking my interest in methodology and encouraging me to always think critically. This work could have not been realized without his constant guidance. Further, I would like to thank the Volkswagen Foundation Computational Social Science Initiative that generously funded by the project from which the data employed in this work emanates. Finally, thank you Fra for being by my side every step of the way.

EXECUTIVE SUMMARY

In recent years, the intersection between digital news media and politics has been at the center of public and academic debate. A trend amplified as political polarization becomes a defining trait of the polity of an increasing number of states. The dynamics of content creation, diffusion, and production intrinsic to the digital environment have revolutionized the news media landscape. As a result of the rapid changes, scholarship has had to innovate, adjust, and adapt theoretically and methodologically.

Competing frameworks have regarded the possible repercussions of a high-choice news media environment. On one side, it is said that the existing digital media environment is conducive to selective exposure, echo-chambers, and filter bubbles. On the other, scholarship addresses the potential for fragmentation, but highlights the tendency for moderation. Most of the evidence provided by empirical work suggests that the state of consumption is one of overwhelming overlap between partisan and ideologically diverse audiences.

This study puts in perspective a segment of the previous work that has employed problematic methodological decisions in the study of audience overlap networks. I leverage the unprecedented granularity of linked individual surveys and digital trace data to demonstrate the importance of utilizing probabilistic methods for edge extraction to render reliable networks that present valid results. Further, thanks to the richness of the data, I build for the first time the consumer-end of an audience overlap network. I employ the consumer-end network to illustrate the opportunities that a network analytic approach offers to the study of questions related to media consumption patterns. I explore ideological homophily in the network through network statistical modeling and the predictive power of network-specific measures on vote choice and partisanship.

The analysis brings up questions about the extent to which the results of previous academic output are methodological artifacts. Additionally, the inspection of the consumer side hints that individuals at the ends of the political spectrum tend to be connected to ideological equivalents in the observed network. Further, the communities extracted from the network resulted in highly predictive of vote choice and to some extent party affiliation.

This study incorporates novel data sources to explore news media consumption from an alternative methodological frame. The analysis does not insinuate selective exposure or active avoidance of cross-cutting content. The results, however, hint that classic exposure measures may conceal dynamics behind the way political ideology relates to consumers' engagement with news media.

CONTENTS

Acknowledgements i

Executive Summary ii

Introduction 1

Literature Review 3

Data 6

Leveraging granular data 7

Method 11

Disparity filtering 12

Exponential Family Random Graphs (ERGMs) 12

Community detection 13

Results 14

Statistically significant edge extraction 14

Ideological homophily 18

Predictive clusters 23

Discussion 26

Conclusion 27

Statement of Authorship 34

Appendices i

Appendix A List of outlets i

Appendix B Statistically significant edge extraction iii

Appendix C Ideological homophily xi

Appendix D Predictive clusters **xii**

D.1 Consumer-end 0.01 backbone network xii

D.2 Consumer-end 0.05 backbone network xv

Appendix E Software Statement **xviii**

LIST OF FIGURES

1 The structure of audience overlap networks 8

2 Daily news site visits 9

3 Top 10 most visited news websites and ideological distribution of respondents 10

4 Degree centrality distribution of the networks under each approach 16

5 Graphical depiction of outlet-end backbone network at the 0.01 α significance level . . 18

6 Consumer-end backbone network at the 0.01 α level 20

7 Coefficient plot of the dyadic independence ERGMs for the two disparity filter specifications 22

8 Predicted probabilities of communities for 0.01 α consumer-end network 24

9 Predicted probabilities of communities for 0.05 α consumer-end network 25

B1 Node counts under different α specification for outlet-end network iii

B2 Edge counts under different α specification for outlet-end network iii

B3 Distribution of node-level degree, betweenness and eigenvector centrality iv

B4 Graphical depiction of the outlet-end unfiltered network v

B5 Graphical depiction of the outlet-end deviation from random duplication network . . vi

B6 Graphical depiction of outlet-end backbone network at the 0.05 α significance level . . vii

D7 Vote choice — Proportion of individuals per attribute held by each community xiii

D8 Party Identification — Proportion of individuals per attribute held by each community xiv

D9 Vote choice — Proportion of individuals per attribute held by each community xvi

D10 Party Identification — Proportion of individuals per attribute held by each community xvii

LIST OF TABLES

1 Descriptive statistics of respondent characteristics, ... 7

2 Illustrative selection of types of network studies by level of analysis 11
3 Network statistics under each filtering specification 15
4 Nodes with higher degree, betweeness, and eigenvector centrality under each filtering method 17
5 Network statistics for the outlet-end network under each specification 19
6 Number and proportion of observations in each detected community 23
A1 List of included outlets with number of users and total reach i
B2 Communities obtained from backbone network after disparity filter at 0.01 α viii
B3 Communities obtained from backbone network after disparity filter at 0.05 α ix
C4 Model output of Exponential Random Graph Models for backbone consumer-end networks at the 0.01 and 0.05 significance levels. xi
D5 Vote for Donald Trump — Predicted probabilities of community models xii
D6 Vote for Hillary Clinton — Predicted probabilities of community models xii
D7 Self-identified Republican — Predicted probabilities of community models xii
D8 Self-identified Democrat — Predicted probabilities of community models xii
D9 Vote for Donald Trump — Predicted probabilities of community models xv
D10 Vote for Hillary Clinton — Predicted probabilities of community models xv
D11 Self-identified Republican — Predicted probabilities of community models xv
D12 Self-identified Democrat — Predicted probabilities of community models xv

INTRODUCTION

> "Modeling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. The first is that all models are wrong; some, though, are better than others and we can search for the better ones."
> P. McCullagh and J.A. Nelder (1989)

In recent years, questions regarding the relationship between digital news media and politics have been at the center of scholarly and public debates. The promise of the internet as a liberating tool is under inspection. The idea that information, often perceived as the citizen's currency in representative democracies, was going to increase its reach via the boundless net, appears to face resistance as political polarization becomes a defining feature of the polity of a growing number of states. Many theoretical and empirical research has focused on tackling questions of selective news exposure, echo chambers, and balkanized media environments. Competing theoretical arguments present mechanisms through which information-rich environments affect individual news consumption. On one hand, a stream of literature portrays the high-choice media environment as a facilitator to increasingly fragmented and reinforcement-seeking news consumption (e.g. Sunstein, 2001; Prior, 2007; Hindman, 2008; Garrett, 2009*a*; Sunstein, 2018). On the other hand, a different stream presents a picture where spite of the potential for fragmentation, there is an overlapping news consumption culture conducive to shared media experiences. (e.g. Gentzkow and Shapiro, 2011; Webster, 2014; Weeks et al., 2016; Guess, 2019).

The implications of the potential for fragmentation in the online news landscape as outlets specialize to find a competitive advantage for attention are vast. The underlying dynamics that reigned over information production are rapidly changing. The shared information space that characterized the pre-internet world, coined as the 'information commons' by Bennett and Iyengar (2008), is challenged by an environment where the audience acts as a mediator and co-producer of information. Normative questions arise concerning the future of public discourse in instances where citizens do not have a common reference point to engage in relevant social and political issues. The importance of the deliberative process in the public sphere for a functioning democracy is central to the arguments of various political theorists (Habermas, 1991; Neuman et al., 1992; Carpini and Keeter, 1996). The romanticized image of the 'informed citizen' could then be endangered by siloed information consumption.

Empirical identification strategies of ideological media enclaves have largely relied on self-reported media consumption (Mutz, 2001; Prior, 2007; Stroud, 2008; Weeks et al., 2016), as well as, market research and aggregated web-tracking data (Ksiazek, 2011; Webster and Ksiazek, 2012; Mukerjee et al., 2018*a*; Majó-Vázquez et al., 2019). Recently, with the rise of social science research leveraging individual level digital trace data, Guess (2019) successfully examined the distribution of media consumption against a measure of news media slant by quantifying the similarity of Democrat and Republican media diets. However, data gathered from web-tracking panels are not perfect. I propose that network analysis can provide an alternative perspective to engage the potential of trace data, less reliant on endogenous ideology measures to map the overlap of distributions and robust to random web behavior of web-tracking panel participants. Though the field of network science is not foreign to the literature of news media consumption, the majority of audience network and duplication research has utilized the problematic deviation from random duplication measure (Ksiazek, 2011; Webster and Ksiazek, 2012; Taneja and Webster, 2016; Nelson and Webster, 2017). The most recent approaches to audience overlap network modeling advocate for the use of network reduction techniques that account for the statistical significance of edge formation (see Mukerjee et al., 2018*a*; Majó-Vázquez et al., 2019).

I make use of individual level web-browsing and survey data collected from a sample of 1339 U.S. American respondents to illustrate the promises that the coupling of these data presents for the study of audience overlap and media consumption. I utilize the disparity filtering methodology proposed by Serrano et al. (2009) to extract the backbone structure of the online news audience network. I compare the news outlet-end backbone network structures extracted to its analogs under non-filtering and deviation from random duplication methods. Further, I construct for the first time the consumer-end of a news media audience network. I employ the consumer network to display the opportunities that the granularity of individual web-tracking data, in tandem with survey responses, present for inferential, and network statistical analysis. To demonstrate this, I apply an Exponential-Family Random Graphs Model (ERGM) to assess ideological homophily in digital news consumption. Additionally, I utilize the Louvain method for community detection Blondel et al. (2008) to explore the predictive power of the audience network modularity on party affiliation and vote choice.

The results suggest that backbone extraction through disparity filtering is more robust to heterogeneous weight and degree distributions than previous methodological approaches. With moderate centralization scores at the outlet-end of the network, the analysis portrays a digital media environ-

ment where audiences have a general tendency to overlap. Additionally, the outcome of the ERGM model on the consumer-end portion of the network indicates that participants on the more extreme ends of the ideological spectrum are more likely to attach to ideological analogs, which is not the case for moderates. Finally, the application of community extraction based on the density of the connections between the nodes exemplifies the opportunities of network statistics for inferential statistical analysis.

LITERATURE REVIEW

The concept of selective exposure has a long-standing tradition in social science research. The study of people's inclination to receive congenial information has been a cross-cutting concern of communication, sociology, psychology, and political science scholars. As early as the 1960s, research evaluating the role of individual beliefs and predispositions on information-seeking and processing had gained salience amongst academic circles to be considered worth synthesizing and critically inspected (see Sears and Freedman, 1967). The diverse body of literature, partly based on Festinger (1957) postulates of cognitive dissonance, has assessed the potential for selective exposure in a wide range of outcomes. The general academic interest on selective exposure has been magnified by the rapid transformation that technological developments have generated on the media landscape.

The profusion of media choices, especially through digital channels, has generated concern about political opinion formation, polarization, and fragmentation (e.g., Prior, 2007; Sunstein, 2001, 2018). The dynamics of online content creation and dissemination create an environment in which the opportunity for selective exposure is larger than it has ever been. The academic debate remains respecting the mechanisms at play: whether inclination for congenial information, active avoidance of dissonant content, or active seeking for reinforcing content Garrett (2009*b*). Some evidence suggests that media consumers tend to approach political news with confirmation bias (Knobloch-Westerwick et al., 2020). Gentzkow and Shapiro (2011), whilst studying the ideological fragmentation of the American electorate, find that although online news consumption segregation is higher than its offline counterpart, it is substantially lower than ideological segregation of face-to-face interactions. In the social media setting, Bakshy et al. (2015), observe that though conservatives and liberals exhibit substantial homophily in their Facebook friend networks, they are being exposed to and engaging with, cross-

cutting news content. Similar findings come from Twitter, where it is highlighted that users do come into contact with information from diverse ideological grounds due to the dynamism and flexibility of communication structures (Barberá et al., 2015). Guess (2019) finds that most U.S. Americans across the political spectrum have ideological moderate media consumption diets by leveraging individual surveys and web-tracking data. On the other hand, Stroud (2008) suggests that people's partisan beliefs motivate their media usage behavior by analyzing cross-platform media consumption during the 2004 US presidential election. After exploring the browsing histories of 50,000 internet users, Flaxman et al. (2016) find that social networks and search engines are associated with an increase in the mean ideological distance of individuals; however, these are also associated with increased exposure to content from the other side of the ideological spectrum. Recently, Stier et al. (2020) conducted a cross-country study utilizing web-tracking and survey data to assess the relationship between selective exposure and populist views. The authors find that populist attitudes touch people's media consumption habits; however, these results are dependent on the supply end of the country's media structure.

The questions raised by the increased opportunity to engage in ideologically siloed media consumption in the high-choice environment are far from settled. A subset of scholars has leveraged the insights from early audience duplication research (see e.g., Goodhardt, 1966; Goodhardt and Ehrenberg, 1969; Webster, 1985) to assess the degree to which audiences are fragmented or centralized. Ksiazek (2011) suggested a network analytic approach to understand cross-platform audience behavior, where media outlets represent the nodes and the existence, and extent, of duplication between their audiences the edges of the network. The network approach to audience duplication has been utilized to measure siloed and cross-cutting media consumption, often relying on aggregated consumption data from media measurement and analytics companies, such as Nielsen and ComScore (Taneja, 2016). Following the proposed conceptualization of the network analytic approach, Webster and Ksiazek (2012) formalized the framework for the study of polarization in media diets, pointing at the limits of alternative tools to study media consumption fragmentation and highlighting the benefits for an audience-centric approach. The authors find evidence of a largely overlapping media consumption pattern based on the audience network centralization measure. The overlapping patterns of news consumption are described as "remarkable for how little variation there is. Almost every outlet shares an audience with every other outlet, above and beyond the level of duplication that would be expected by chance alone"

(Webster, 2014, p. 122). More recently, Taneja and Webster (2016) explored the cultural and technological dynamics at play in audience network formation at a global scale, finding that the network was highly centralized, as well as, the importance of cultural proximity informing the global media use network. Nelson and Webster (2017) presents a portrait of the online political news audience based on online audience data provided by ComScore. The authors find that large political news sites attract ideologically diverse audiences and that they share audiences with most smaller and ideologically extreme outlets.

All the previously mentioned studies that utilize a network analytic approach employ the deviation from random duplication method to assess edge significance proposed by Ksiazek (2011). The method assumes that the difference between the expected and observed audience overlap serves to test the significance of the ties. Deviation from random duplication mimics the logic behind the chi-squared test of independence. According to Ksiazek, if the difference between observed duplication and expected duplication between two nodes is greater than zero, the overlap deviates from random duplication. Observed duplication refers to the percentage of audiences that visited both sites *i* and *j*, whereas the expected duplication is the dot product of the percentage of the audience that visited *i* and the percentage of the audience that visited *j*. For example, if 50% of the overall audience visited outlet *i* and 50% visited outlet *j*, we would expect 25% of overlap between the two outlets by chance. Under this logic, if in the network the outlets *i* and *j* have an observed overlap greater than 25%, the occurrence of the tie would pass the test.

The research based on deviation from random duplication showcases the potential that a network analytic approach can offer to the literature on media exposure, however, Mukerjee et al. (2018*a*) point out that the method is hardly a statistical significance test. Though deviation from random duplication mimics the logic behind the chi-squared statistic calculation, it does so based on percentages and not frequencies, which ignores the differences in reach that the outlets have. Beyond that, as opposed to the chi-squared test, which offers a convention based on the degrees of freedom, deviation from random duplication takes any value greater than zero as significant (see Mukerjee et al., 2018*a*; Webster and Taneja, 2018; Mukerjee et al., 2018*b*, for a more detailed assessment of, and discussion about, deviation from random duplication). The authors posit a paramount concern, "audience behavior is noisy (i.e., audience metrics do not completely filter out random behavior), having a probabilistic method at hand to eliminate the weakest overlapping ties (weakest in a statistical sense) is important" (Mukerjee

et al., 2018*a*, p. 33). They present a novel approach based on the *phi coefficient* operating at the dyadic level. Though Mukerjee et al. (2018*a*) showcase that the overall structure and centralization score of the networks substantially change, they do not go into detail on how deviation from random duplication affects different outlets along the weight distribution. In the following sections, I employ the more established technique in network science of disparity filtering (Serrano et al., 2009)[1]. Disparity filtering, also known as backbone extraction, has already been implemented in the camp of audience duplication networks in Majó-Vázquez et al. (2019). I utilize the probabilistic filtering technique to explore the substantive effects of utilizing deviation from random duplication beyond the network structural attributes. Further, given that network analysis of audience duplication has relied on aggregated data, I leverage the granularity of individual trace data to construct for the first time the consumer-end of the audience network.

DATA

This study relies on a combination of individual level survey and web-tracking data. The data were collected by the market research firm *YouGov* for the *Paying Attention to Attention: Media Exposure and Opinion Formation in an Age of Information Overload* project [2]. The subject sample, gathered from the U.S. American *YouGov* pulse panel, is composed of $n = 1339$ panelists and is designed to constitute a representative set of the electorate. The panelists were incentivized to install a passive web-tracking software that logged their site visits on their personal computers and mobile devices. Participants were presented with the information about the character of the data collection effort and were asked for explicit consent. Additionally, the panelists could decide to stop the passive tracking for periods of 15 minutes or opt-out of the tracking altogether at any point in time. The web-tracking log links the user identification number of the panelist to the URL of the site visited, the time, date, and duration of the visit, as well as the type of device where the site was accessed from. In this paper, the web traffic is condensed at the URL domain level. A visit is counted as a page view. The passive tracking does not disaggregate by engagement with the website.

[1] Mukerjee et al. (2018*a*) note the promises of disparity filtering, yet decide to utilize the *phi correlation* approach to illustrate the drawbacks of deviation from random duplication, as both derive the null at the dyadic level, as opposed to the ego-network level

[2] Further information regarding the project by Pablo Barberá, Andrew Guess, Simon Munzert, and JungHwan Yang can be found here

TABLE 1: Descriptive statistics of respondent characteristics,

Attribute	n
Sex	
Male	644
Female	695
Age	
18-34	138
35-44	190
45-54	202
55+	809
Region	
Midwest	335
Northeast	220
South	479
West	305
Education	
No HS	18
High school	205
Some college	293
2-year	195
4-year	344
Post-grad	284

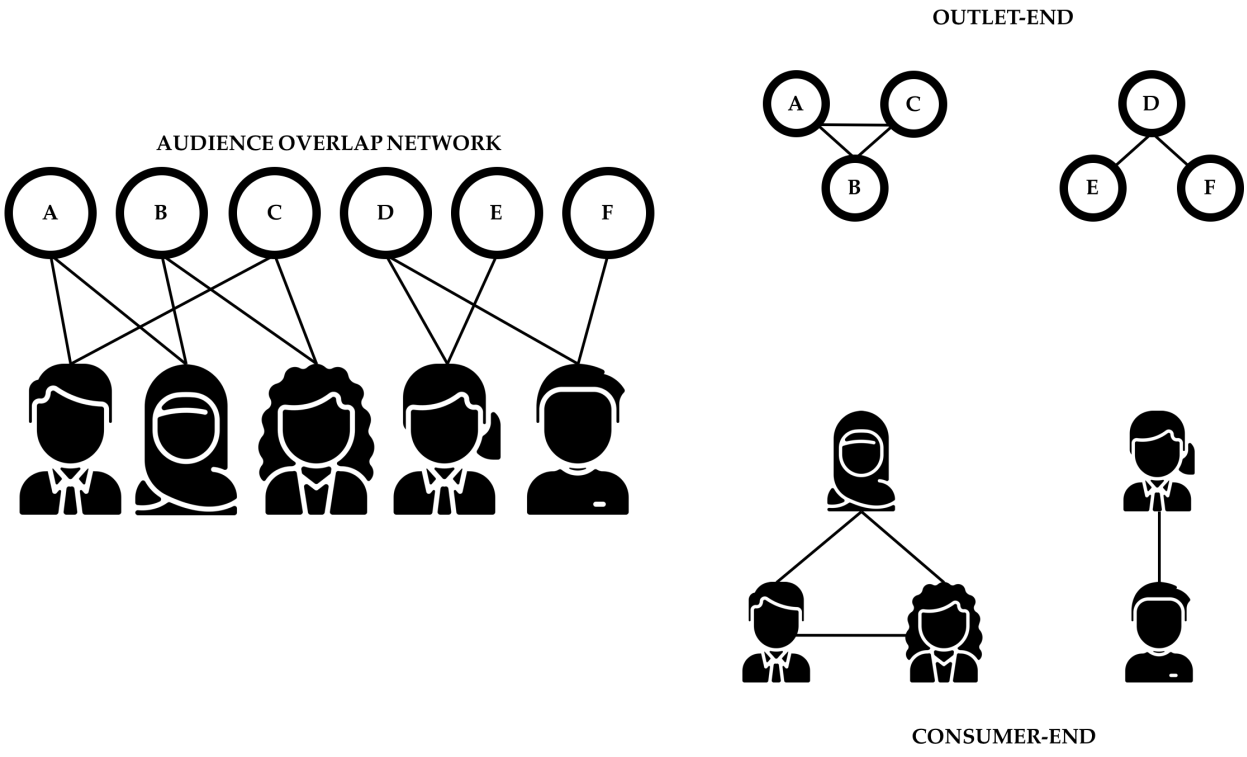
In addition to the web-tracking logs, I make use of survey responses linked to the digital trace data through an anonymous unique user identification number. Central to the analysis are the survey responses to self-reported media consumption, political ideology, and demographic information. Table 1 showcases some descriptive statistics of the attributes of the panelists. The *Paying Attention to Attention: Media Exposure and Opinion Formation in an Age of Information Overload* project fielded eight survey waves ranging from April 2018 to October 2019. As part of the project, during the third survey wave —fielded from October 05 to October 29, 2018— *YouGov* panelists were assigned to an experiment that could affect media consumption patterns. With this in mind, this study only makes use of the responses emerging from the first fielded survey wave from April 23 to June 07, 2018. Additionally, the web-tracking data employed for this study is contained to the navigation history of the panelists from August 1 to October 4, 2018 with a total of $26'227,508$ URL entries.

LEVERAGING GRANULAR DATA

Traditionally, audience duplication research has relied on aggregated data at the outlet level. The data often reflect two dimensions of consumption. First, at the individual outlet level, the proportion of

the sample that consumed outlet *i*. Second, at the dyadic relation, the proportion of the sample that accessed both outlets *i* and *j*. With the latter measure being the extent of the overlap. As it is presented in Figure 1, the measures utilized to construct the audience overlap network can be broken down into smaller sub-units. The overlap previously described is presented as the outlet-end of the network, where the nodes of the network represent the news outlet and the ties are created by consumers. This is a prime example of the opportunities that the integration of survey and digital trace data at the individual level can offer for social science research (see Stier et al., 2019, for an overview). The granularity of the disaggregated trace data allows for the first time to produce what is presented as the consumer-end of a news media audience network. The nodes represent individuals and the ties are formed by the co-occurrence of consumption of a news outlet.

To exploit the potential offered by the nuance of the combined data set, pre-processing and data cleaning decisions were required. Since the trace data comprise the complete browsing behavior of the panelists, there were no pre-determined mappings of digital news outlets. I utilized the list of outlets



Note. This graph is partly based on a graphic rendition by Mukerjee et al. (2018*a*)

FIGURE 1: The structure of audience overlap networks

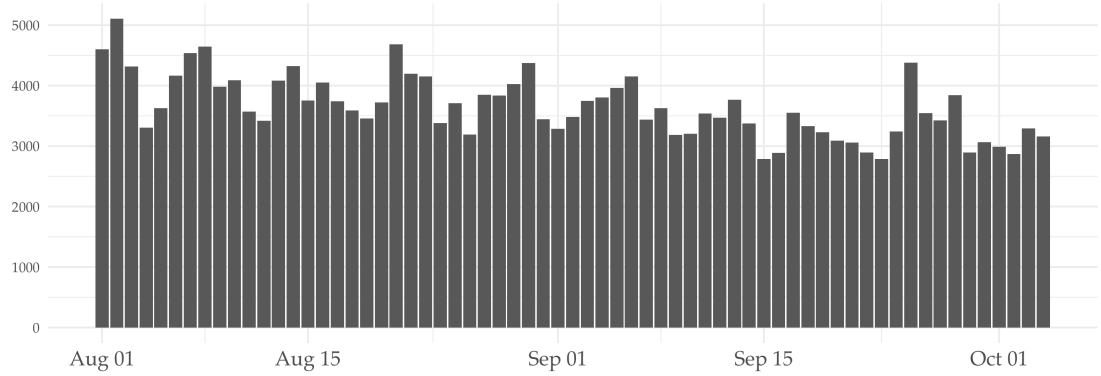


FIGURE 2: Daily news site visits

provided by the *AllSides.com* project, which relies on blind surveys, editorial reviews, and academic analyses to rate news outlets along the left-right spectrum (All Sides, 2018). The ideological ratings are disregarded for this study; however, as an artifact of the task that the *AllSides.com* project undertakes, the collection of outlets captures less prominent sites, not present in the ComScore and Nielsen data, crucial to understanding the assortment of media diets. The full list of outlets contains 296 single digital news entities. The site entries that contained visits to the news outlets' domain in the digital trace data set were extracted. A total of 210 outlets were mapped and included in the analysis. The full list of the news outlets and their observed reach can be seen in Appendix A.

The extraction process of news visits from the web-tracking data rendered a total of $243,509$ entries by 1117 of the panelists. Figure 2 presents a graphic rendition of the aggregated daily news consumption trends of the panelists. The daily news site visits ranged from 5106 to 2785 during the two-month period, with a median of 3569 visits associated with the sites. The extracted news visits were used to generate the adjacency matrices for the consumer- and outlet-end networks. Further, Figure 3 showcases the ten most visited news sites by raw visit count measures and the ideological distribution of the respondents. Though the sample's political ideology follows a normal distribution across the spectrum, the graphic portrays outlet traffic being disproportionately driven by different ideologically aligned users. More than 85% of any of the top 10 sites' traffic is accounted for by a collection of respondents on a single side of the ideological spectrum plus moderates.

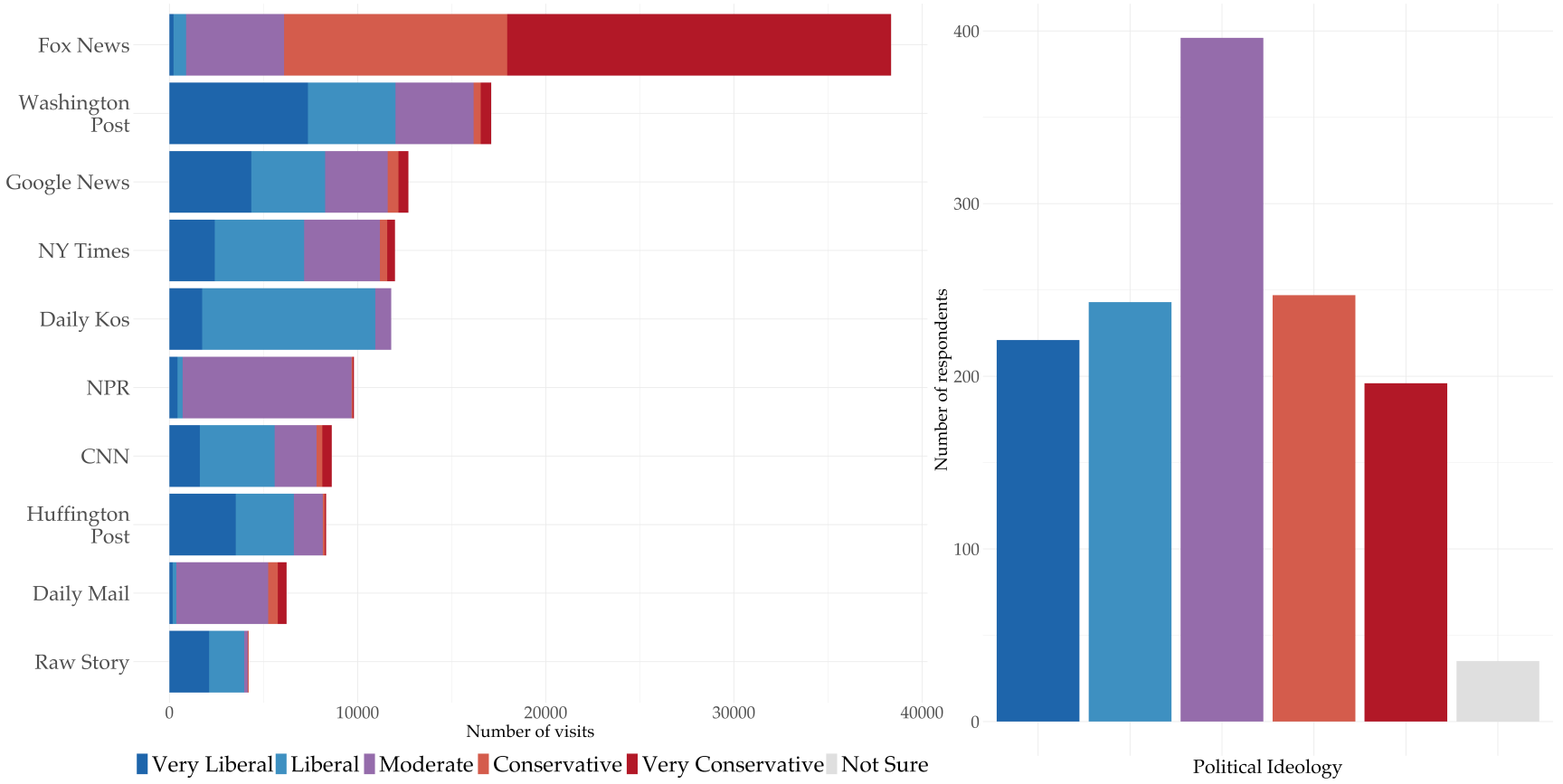


Figure 3: Top 10 most visited news websites and ideological distribution of respondents

METHOD

The analysis of network data has become increasingly salient in social science research. I propose that the insights generated from network descriptive analysis, in addition to inferential and predictive statistical network modeling, can provide an alternative perspective to engage the full potential of individual level digital trace data for the study of media consumption. So far, due to data limitations, audience overlap research has been bound to mostly network- and node-level analysis of news outlets. The richness of novel data sources allows for diverse avenues of study. Table 2 illustrates some of the possible relationships that can be analyzed under a network approach.

TABLE 2: Illustrative selection of types of network studies by level of analysis

Level	Network attributes as explanatory features	Network attributes as outcome features
Node	Degree centrality of an individual as predictors of correct political knowledge answers in the survey	Self-reported news consumption of individuals or type of news outlet (e.g., legacy or digital born) to predict centrality
Network	Community affiliation as a predictor of individual vote choice and partisanship	Type of news media environment (e.g., legacy or digital-born dominated) to predict the structure of the network
Dyad	Edge between a pair of individuals as a predictor of political public opinion answers in the survey	Similarity of individuals' demographic attributes to predict edge formation (e.g., ideological homophily)

Note. The table excludes instances where network attributes can be both explanatory and outcome features.

The examples of the table are conceived for undirected networks; however, digital trace data can be used to reconstruct web-browsing paths, which would allow for the study of audience and information flows to understand the way individuals navigate the media environment. Since, this examination is centered on audience overlap networks, the assessment of flows is outside of the scope of this paper; nevertheless, the insights that could be generated by exploring a directed network iteration are open for future research. My focus in this study is two-fold. First, to demonstrate the importance of reliable statistical significance testing to render dependable networks for analysis. Second, to illustrate the potential of a network analytic approach to study questions related to media consumption patterns. In the following sections, I provide a brief overview of three established network science methods utilized in this paper: Disparity filtering (Serrano et al., 2009), the Louvain method for community

detection (Blondel et al., 2008), and Exponential Family Random Graph Models (ERGMs) (Wasserman and Pattison, 1996).

DISPARITY FILTERING

As it has been previously established, the reach of news outlets varies by different orders of magnitude. Deviation from random duplication research disregarded the node and edge weight distribution. I utilize the disparity filtering method proposed by Serrano et al. (2009), also known as backbone extraction, to retrieve the ties that constitute the overlap that is statistically significant. This approach was introduced to the audience overlap research by Majó-Vázquez et al. (2019). The backbone extraction method defines its null model at the ego-network level, as opposed to the dyadic level. In other words, the disparity filtering technique derives statistical significance relative to the node being considered. This is particularly important given the long-tail distribution that characterizes media consumption. The technique accounts for the disparity in the weight distribution to establish statistical significance. The null model for undirected weighted networks is dependent of the degree k of the node being examined, where the probability density function of values being x is:

$$p(x)dx = (k-1)(1-k)^{k-2}dx \tag{1}$$

Based on the null model, the backbone extraction technique identifies the edges that are statistically significant. For example, in the dyad of nodes *i* and *j*, the method assesses the probability α that the normalized weight p_{ij} is compatible with the null hypothesis. The latter can be formalized as:

$$\alpha_{ij} = (k-1)\int_0^{p_{ij}}(1-k)^{k-2}dx < \alpha \tag{2}$$

The test renders interpretable *p-values* as mainstream statistical significance tests. All the edges that have a probability value larger than the specified α are retained as significant edges. The method ensures that the fluctuation of the strengths of the nodes in the news consumption network system is not disregarded. Disparity filtering renders a network that does not penalize either small, nor large nodes.

EXPONENTIAL FAMILY RANDOM GRAPHS (ERGMS)

Exponential random graphs models, also known as *p** models, are conceived to examine the collection of network micro-configurations that lead to the observed network (Wasserman and Pattison, 1996; Pattison and Wasserman, 1999). In other words, ERGMs are statistical models concerned with modeling the random behavior of the adjacency matrix Y conditional on the covariate matrix X. Due to the nature of network formation, network data intrinsically violates the independence assumptions of mainstream statistical modeling strategies. ERGMs are consistent with the existence of dependency between network edges. The *p** models hold similarities with generalized linear models; however, the network models present the 'conditional probability of a tie', based on what is observed in the network. The general form of the model can be formalized as:

$$Pr(Y = y|X) = k^{-1}exp\{n^t g(y, X)\} \tag{3}$$

where $g(Y, X)$ is a user-defined theoretically motivated vector conceived to capture the social processes of the network and k is a normalizing constant defined by:

$$k = \sum_{w} exp\{n^t g(y, X)\} \tag{4}$$

Given the underlying aim of this portion of the study, namely, illustrate the opportunities attached to the network analytic approach, I focus on dyadic independence models for simplicity and to avoid degeneracy from MCMC maximum likelihood estimation (see Hunter, Handcock, Butts, Goodreau and Morris, 2008; Hunter, Goodreau and Handcock, 2008; Lusher et al., 2013). The limitations of relying on independence, rather than dyadic dependence modeling are examined in the following section.

COMMUNITY DETECTION

Given the formulation of academic questions of selective exposure and media consumption patterns in general, the field of community detection within network science constitutes a pertinent tool for the analysis of audience networks. Community detection consists in separating the network into sets of densely connected nodes, only sparsely tied to nodes belonging to other sets. In this paper, I utilize the largely utilized Louvain method for community detection proposed by Blondel et al. (2008). The

technique extracts communities by optimizing modularity, which measures the density of edges inside communities compared to edges amongst communities.

RESULTS

In this section I focus on three specific tasks. First, I assess comparatively the resulting outlet-end network extracted from the digital trace data under different filtering specifications: unfiltered, deviation from random duplication, and disparity filtering at the 0.05 and 0.01 α significance levels. Second, I explore the integration of network statistical analysis to the study of media consumption by leveraging a dyadic independence ERGM to assess ideological homophily in the consumer-end network. Finally, I inspect the predictive power of network-specific objects, in this case the extracted communities from the consumer-end network, to model individuals' vote choice and party affiliation.

All analyses for this study were performed in R (R Core Team, 2018). The network objects were created based on the $igraph$ and $statnet$ packages (Csardi and Nepusz, 2006; Butts, 2015). The applications of disparity filters, Louvain community detection and ERGMs were based on the $skynet$, $igraph$, and $ergm$ packages respectively (Teixeira, 2018; Csardi and Nepusz, 2006; Handcock et al., 2019). A software statement for this study can be found in Appendix E.

STATISTICALLY SIGNIFICANT EDGE EXTRACTION

I make use of the outlet-end adjacency matrix to assess the impact that methodological choices for statistical significant edge extraction have on structural and substantive features of the media consumption networks. Though in their critique to deviation from random duplication Mukerjee et al. (2018*a*) present an overall image of how the network structure changed due to their proposed technique, the authors did not provide a parallel with deviation from random duplication to put previous research in perspective. I perform the analysis based on four modeling choices: unfiltered, deviation from random duplication, and disparity filter at the 0.05 and 0.01 α levels of significance. I extracted network- and node-level statistics to illustrate how methodological choices in appraising significant overlap could render divergent impressions of media consumption. Table 3 presents a comparison of the structural characteristics of the networks rendered by each approach. Network centralization provides an index of equality in the degree distribution, where 0 indicates that all nodes have the same

degree and 1 where a sole node dominates the whole network. Density presents the proportion of actual against potential edges in the observed network. Transitivity, also known as the clustering coefficient, refers to the ratio of triangles and connected triples in the network. Finally, maximum and minimum degree refers to the number of adjacent edges of the more and less connected nodes in the network.

TABLE 3: Network statistics under each filtering specification

	Unfiltered	Dev-from-random duplication	Disparity filter at 0.05	Disparity filter at 0.01
Number of Nodes	210	210	157	55
Number of Edges	16432	16061	1222	105
Centralization	0.25	0.25	0.8	0.67
Density	0.75	0.73	0.1	0.07
Transitivity	0.87	0.85	0.26	0.12
Max Degree	208	206	141	40
Min Degree	2	2	1	1

The characteristics of the network emergent from the deviation from random duplication approach are not very different from those of the raw unfiltered network. The differences in centralization when compared to the backbone resultant networks are worth assessing since they have been dominant in audience overlap research. Under deviation from random duplication, the resultant network has a more equal degree distribution amongst outlets — 0.25, compared to 0.8 and 0.67 from the backbone networks. In Ksiazek (2011), the author indicates that audience behavior has an inclination to fragmentation based on a centralization score of 0.17, whilst in Webster and Ksiazek (2012), the authors counter the previous results and highlight 'the myth of enclaves' based on a centralization score of 0.86. The observed results from this network are not directly transferable to other networks. Still, they illustrate the possibility that previous findings could be partly driven by a methodological artifact of the deviation from random duplication approach. It is possible that the centralization scores presented through the method follow the distribution of the mapping from the raw network, not that of the significant overlap.

Furthermore, Table 4 provides an overview of the impact of the different techniques at the node level. The table contains the five outlets with the highest degree, betweenness, and eigenvector centrality measures under each method. As it was established previously, degree centrality refers to the number of ties a node has. Additionally, betweenness points at how many times a node falls on the

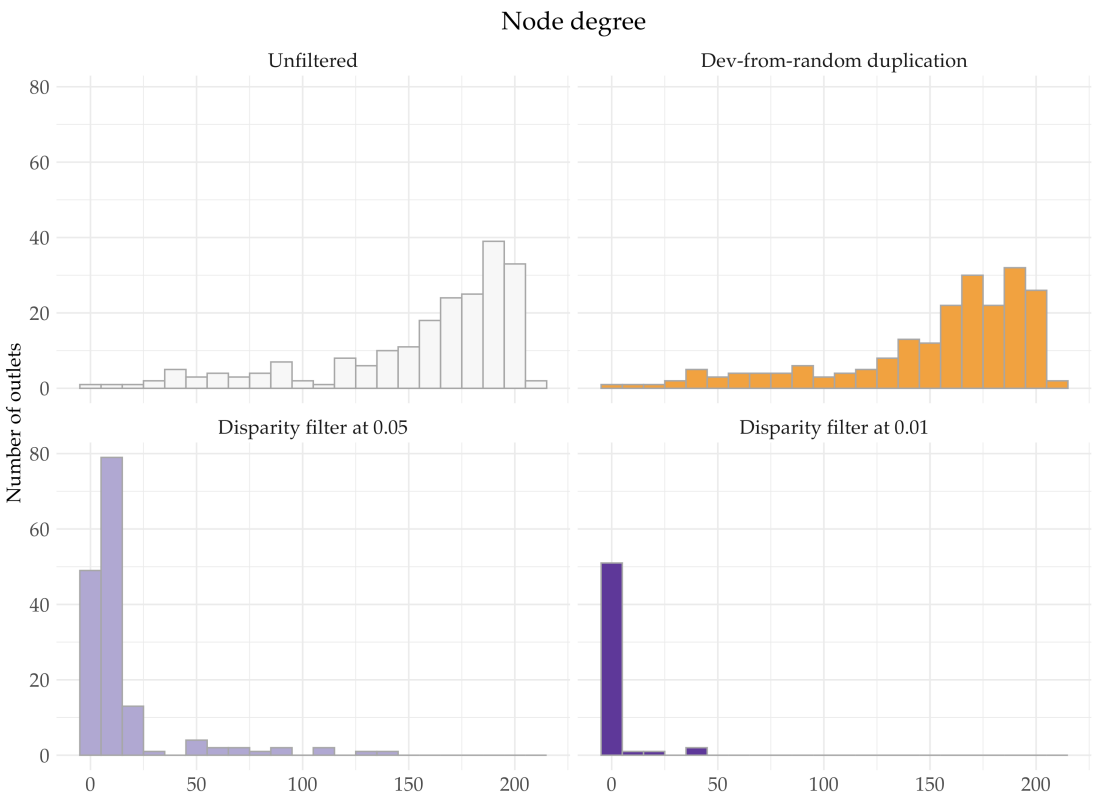


FIGURE 4: Degree centrality distribution of the networks under each approach

shortest path between two other nodes and eigenvector centrality presents a node's centrality proportional to the sum of centralities of the nodes that it is connected to (Borgatti et al., 2018). In other words, degree centrality presents a raw count of ties, betweenness showcases a node's role as a bridge, and eigenvector centrality provides a measure of how popular a node is accounting for how well connected its adjacent nodes are. The top outlets on degree and eigenvector centrality render intuitive results for all four approaches; however, as Figure 4 presents, the distribution is negatively skewed for the baseline and deviation from random duplication networks. The distribution counters the characteristic long-tail positively skewed distribution of media consumption, rendered in the backbone networks. The top outlets remain intuitive because of the inclination of popular outlets to be better connected even without assessing statistical significance. The latter is not the case for betweenness centrality, where outlets like *www.commentarymagazine.com* become important connectors. Since disparity filtering assesses significance at the ego-level, it captures noise introduced by the phenomenon that may be behind this, namely, a very small number of consumers with large media diets.

Table 4: Nodes with higher degree, betweeness, and eigenvector centrality under each filtering method

	k		*Betweeness*		*Eigenvector centrality*
Unfiltered		*Unfiltered*		*Unfiltered*	
www.washingtonpost.com	208	news.yahoo.com	1018	www.washingtonpost.com	1
www.nytimes.com	207	www.commentarymagazine.com	963	www.nytimes.com	0.98
thehill.com	205	www.intellectualconservative.com	798	www.cnn.com	0.87
www.foxnews.com	205	www.myrecordjournal.com	681	www.foxnews.com	0.83
www.cnn.com	204	www.civilbeat.org	511	www.usatoday.com	0.82
Dev-from-random dup		*Dev-from-random dup*		*Dev-from-random dup*	
www.nytimes.com	206	news.yahoo.com	1154	www.washingtonpost.com	1
www.washingtonpost.com	206	www.commentarymagazine.com	1033	www.nytimes.com	0.98
thehill.com	203	www.intellectualconservative.com	908	www.cnn.com	0.87
www.usatoday.com	203	www.myrecordjournal.com	776	www.foxnews.com	0.83
www.bbc.com	201	www.civilbeat.org	590	www.usatoday.com	0.82
Disparity filter at 0.05 α		*Disparity filter at 0.05* α		*Disparity filter at 0.05* α	
www.nytimes.com	141	www.nytimes.com	2471	www.washingtonpost.com	1
www.washingtonpost.com	135	www.foxnews.com	1549	www.nytimes.com	0.98
www.cnn.com	110	www.cnn.com	1460	www.cnn.com	0.86
www.foxnews.com	110	www.washingtonpost.com	1186	www.usatoday.com	0.81
www.usatoday.com	92	www.spokesman.com	1027	www.foxnews.com	0.8
Disparity filter at 0.01 α		*Disparity filter at 0.01* α		*Disparity filter at 0.01* α	
www.nytimes.com	40	www.nytimes.com	761	www.washingtonpost.com	1
www.washingtonpost.com	40	www.washingtonpost.com	468	www.nytimes.com	0.98
www.foxnews.com	18	www.foxnews.com	451	www.foxnews.com	0.55
www.cnn.com	9	www.today.com	87	www.cnn.com	0.54
www.cbsnews.com	5	www.cnn.com	40	www.usatoday.com	0.41

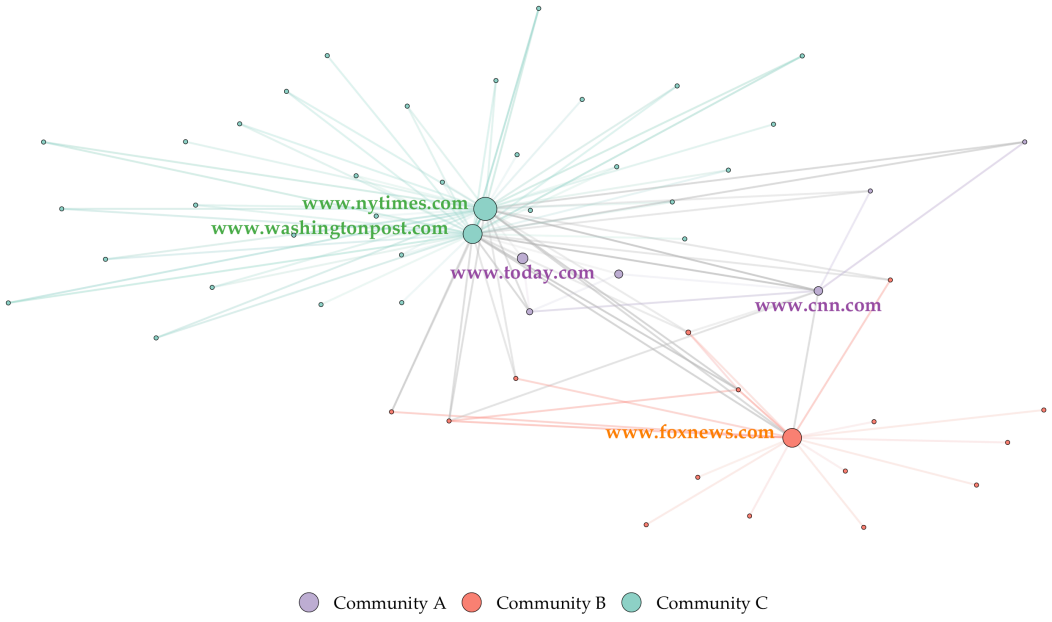


FIGURE 5: Graphical depiction of outlet-end backbone network at the 0.01 α significance level

Figure 5 presents the outlet-end network rendered after extracting the statistically significant edges at the 0.01 α level. In the graph, a node's size represents its betweenness and the tie thickness represents the weight of the edge — the number of users that consume both outlets. Additionally, the extracted networks were subjected to the Louvain method of community detection. In the case of the backbone network at 0.01 α, three communities were detected — represented by each color. The additional graphic renditions of the networks and the table containing the outlets categorized by communities can be found in Appendix B.

IDEOLOGICAL HOMOPHILY

The level of detail intrinsic to individual trace data allows me to construct the consumer-end of the audience overlap network. The network is conceived as an undirected valued network, just like its outlet-end analog. The adjacency matrix formulates each subject as a node, the duplication of consumption as edges, and the number of outlets corresponding to the overlap as the values. The resulting matrix underwent the assessment of edge statistical significance through the disparity filter method.

Table 5 showcases a set of statistics for the outlet-end network under each specification. As can be seen, statistical significance edge assessment reduces the number of ties substantially. For this analysis, I extract the edges significant at the 0.05 and 0.01 α levels of significance. Figure 6 presents a graphical depiction of 0.01 α the network. Nodes and edges are scaled to betweenness centrality and the weight of the overlap respectively. The color scheme represents the self-reported political ideology of the subjects.

TABLE 5: Network statistics for the outlet-end network under each specification

	Unfiltered	Disparity filter at 0.05	Disparity filter at 0.01
Number of Nodes	1117	638	173
Number of Edges	383651	14154	757
Centralization	0.37	0.77	0.64
Density	0.62	0.07	0.05
Transitivity	0.8	0.26	0.21
Max Degree	1096	534	119
Min Degree	10	1	1

As it was established in the previous section, Exponential Family Random Graph Models (ERGMs) are conceived to model the random behavior of the adjacency matrix Y conditional on the covariate matrix X. In this case, this means that the randomness of edge occurrence amongst individuals can be modeled conditional on a set of attributes specific to the respondents. As opposed to mainstream modeling approaches, ERGMs can deal with the existence of dependencies between network edges; however, they maintain a degree of similarity with commonly used generalized linear models that allows for parsimony when examining their results. To illustrate the promises of network statistical inference tools, I explore ideological homophily in the network.

Homophily refers to "the principle that a contact between similar people occurs at a higher rate than among dissimilar people ... [it] implies that distance in terms of social characteristics translates into network distance" (McPherson et al., 2001, p. 416). In the case, the consumer-end network allows us to assess whether individuals are more likely to be connected with others who share their self-reported ideological stance. A coefficient plot containing the estimated odds ratios with 95% confidence intervals are presented in Figure 7.

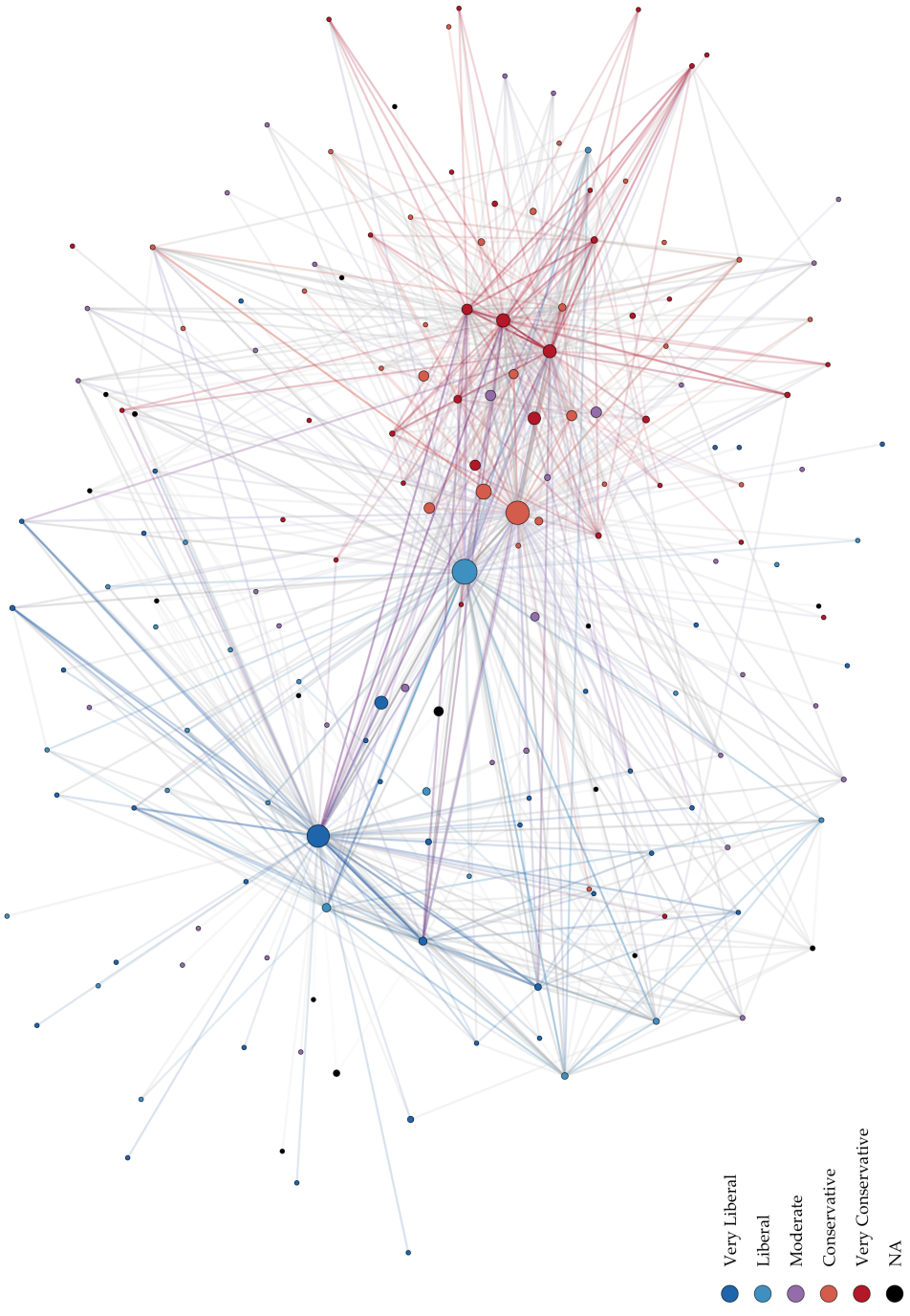
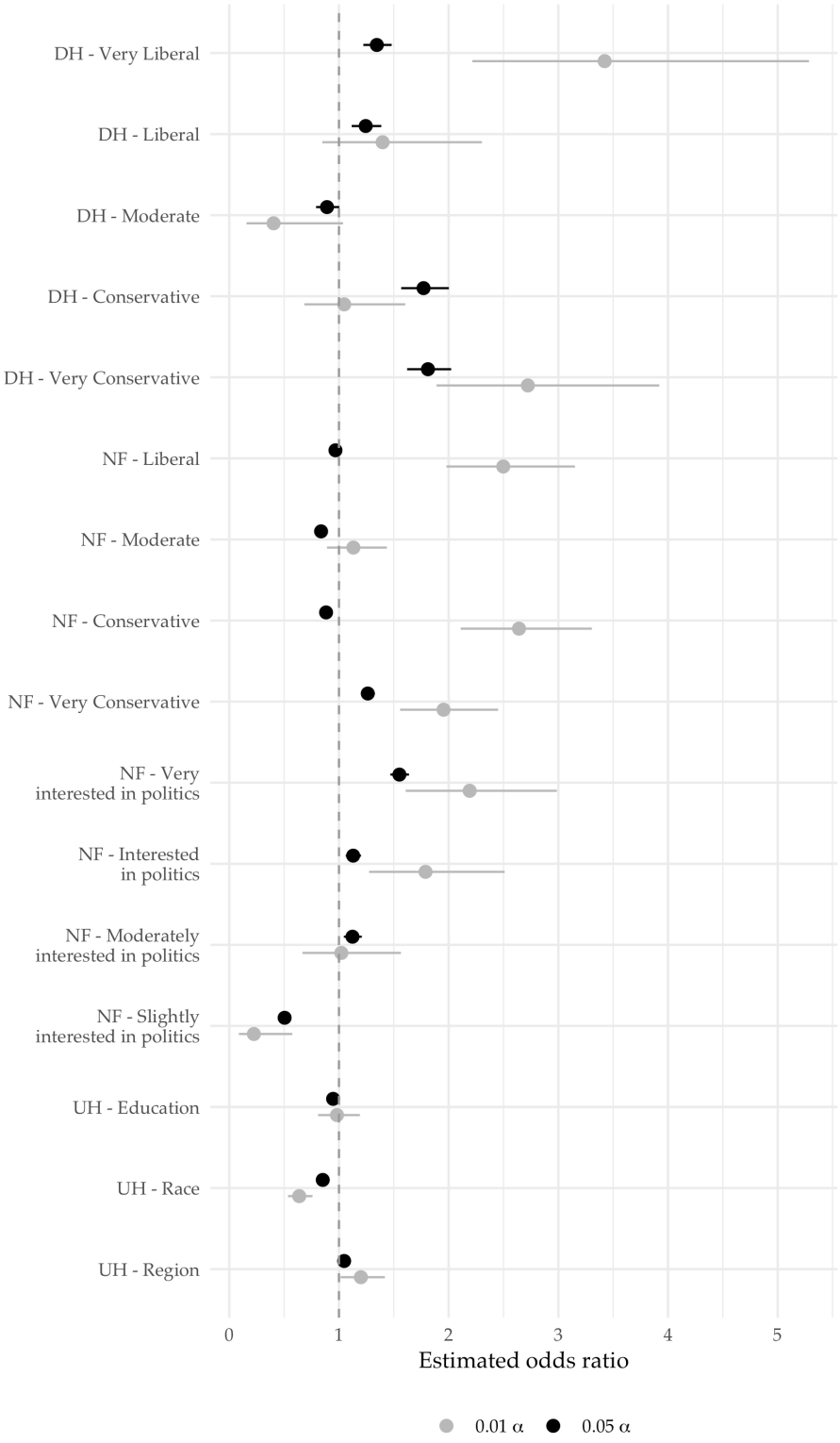


FIGURE 6: Consumer-end backbone network at the 0.01 α level

The model contains uniform homophily terms for educational level, race, and region. The estimate in the plot represents the change of conditional odd ratios when two individuals have the same level of the covariate. For political ideology, the measure is of differential homophily, where each level has its estimate. Finally, the model contains nodal factor terms for each of the factors of political ideology and political interest, following the recommendations from Lusher et al. (2013), to capture the change that comes as a consequence of nodes from a particular level being more active, hence forming more ties. A positive parameter indicates that nodes with the attribute tend to have higher network activity with reference to the baseline factor.

The results suggest that in the consumer-end backbone network at the 0.01 α level, individuals at the extreme ends of the ideological spectrum have a tendency to form edges with others who hold their political ideology. The latter is not the case for self-described leans and moderates. On the other hand, in the 0.05 α level network differential homophily is found at every ideological level except for the self-described moderates. There is a trade-off between being more conservative on the edge statistical significance side and sample size. The estimates rendered by the 0.01 α network have very large standard errors, while the less stringent probability value renders more precise estimates largely due to the sample size. The standard model output can be found in Appendix C.

The modeling performed in this section was a dyadic independence ERGM, which serves an illustrative introductory purpose to ERGMs to convey the possibilities that network statistical inference can bring to media consumption research. I note the existence of more sophisticated modeling strategies such as dyadic dependence ERGMs, and more recently, Frailty Exponential Random Graph Models (FERGMs) (Box-Steffensmeier et al., 2018), which can model the complicated dependence patterns and unobserved heterogeneity not captured by the more simple probability models. Dyadic dependence ERGMs employ Markov chain Monte Carlo (MCMC) algorithms, which makes them more computationally intensive and open to degeneracy (see Snijders et al., 2006; Hunter and Handcock, 2006; Hunter, Goodreau and Handcock, 2008; Lusher et al., 2013, for more information about dyadic dependence ERGMs).



Note. DF — Differential homophily; NF — Node factor; UH — Uniform homophily. The baseline for the nodal factors of ideology and interest in politics are *Very liberal* and *Not interested at all.*

FIGURE 7: Coefficient plot of the dyadic independence ERGMs for the two disparity filter specifications

PREDICTIVE CLUSTERS

For the last application, I explore the potential of network derived measures for statistical analysis. I employ the backbone consumer-end networks at the 0.01 and 0.05 α levels. I leverage the Louvain method for community detection by Blondel et al. (2008) to explore the predictive power of the audience network modularity on party affiliation and vote choice. The method detected three communities for the 0.01 and four for the 0.05 network. Table 6 presents an overview of the extracted communities. I regress four binary dependents on the extracted communities: a) Vote for Donald Trump, b) vote for Hillary Clinton, c) self-identified Democrat, and d) self-identified Republican. The Louvain method maximizes the density of links inside communities compared to links between communities. In this particular application, it entails that the communities represent clusters of more densely connected individuals through their media diets. Figures 8 and 9 present the predicted probabilities extracted from the four logistic regression for each filtering specification.

TABLE 6: Number and proportion of observations in each detected community

	0.01 α		0.05 α	
	n	$\hat{p}$	n	$\hat{p}$
Community A	28	0.16	23	0.04
Community B	87	0.50	83	0.13
Community C	58	0.34	185	0.29
Community D	—	—	347	0.54

In both networks, two communities capture more than eighty percent of the respondents. The models present a captivating picture, where these communities holding the gross of the sample, are exceptionally predictive of vote choice and to some extent partisan affiliation. The results indicate that the clustering based on optimizing group edge density captures underlying features relating media consumption patterns to political opinions and behavior. The outcomes are robust to further model specifications controlling for individual demographic attributes (see Appendix D). This application is not intended to offer evidence on the state of media consumption, rather the aim is to illustrate how network measures open new avenues for analysis. The results offer a different perspective about the conceptualization of diversity in news media use. The networks do not present an environment where Trump and Clinton's voters do not overlap, yet the way they are connected to each other appears to be different as it is captured through the cluster extraction method.

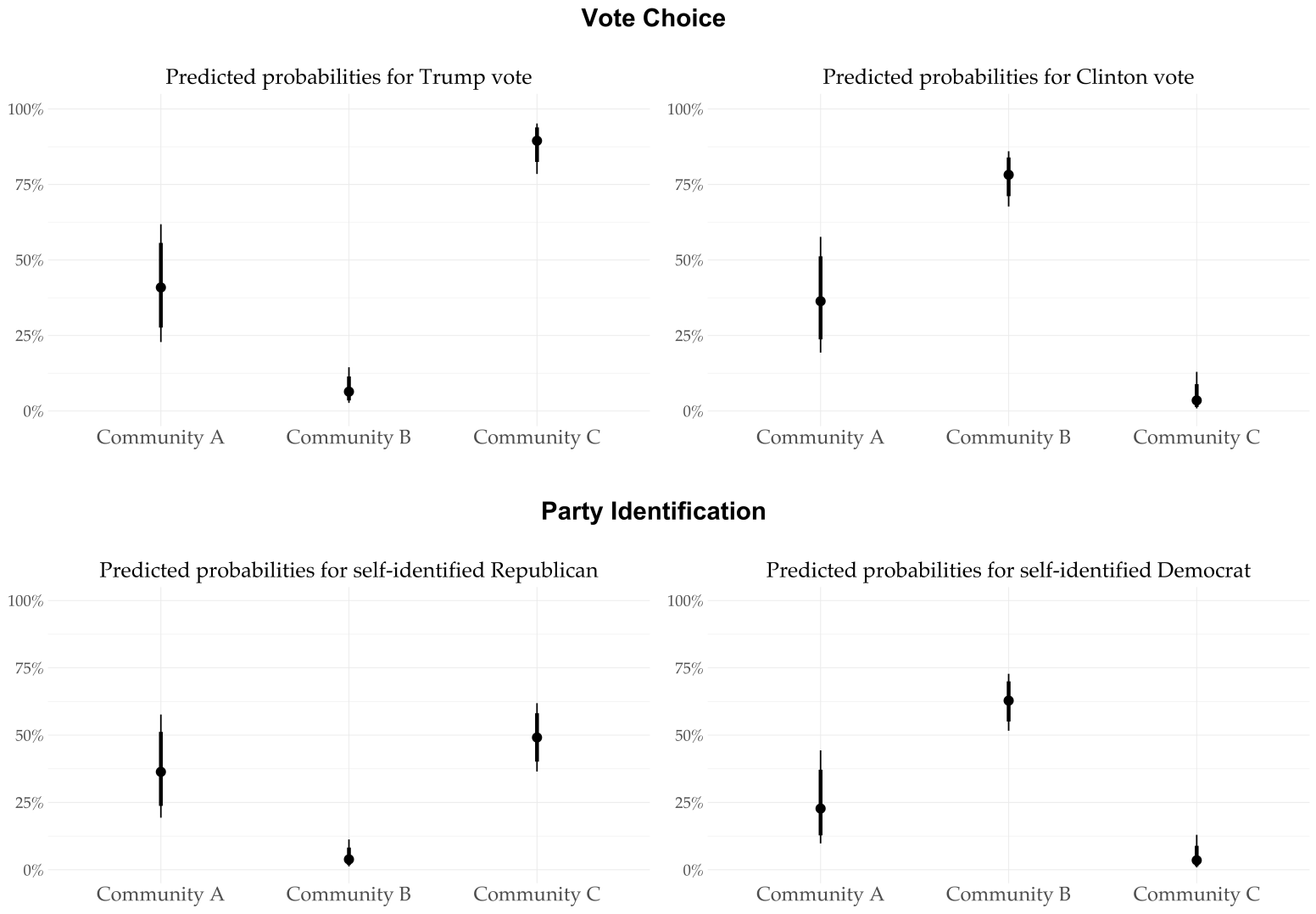


FIGURE 8: Predicted probabilities of communities for 0.01 α consumer-end network

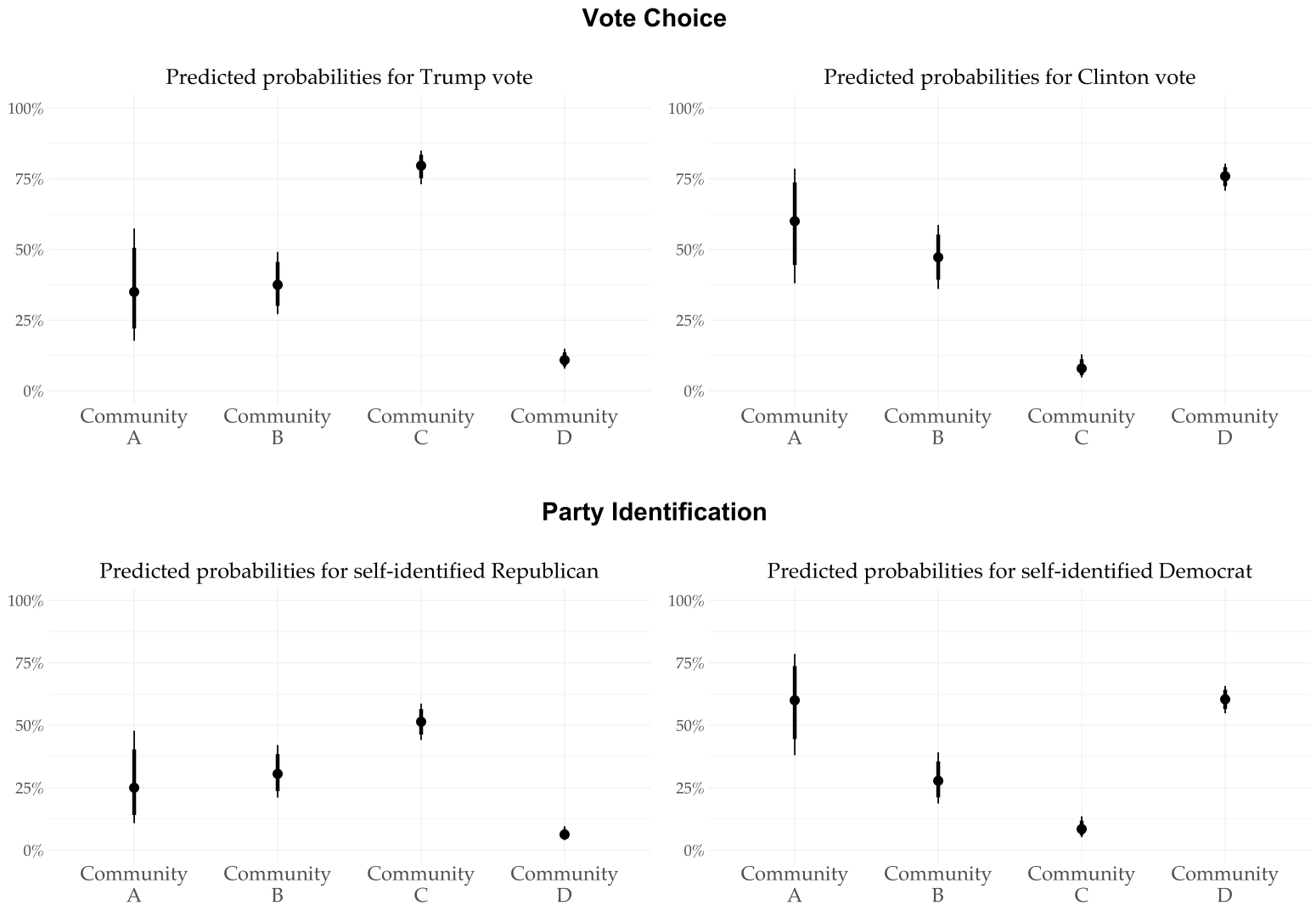


FIGURE 9: Predicted probabilities of communities for 0.05 α consumer-end network

DISCUSSION

The digital technologies, at the center of the substantive questions that concern this field of research, provide an unprecedented stream of data to explore social phenomena. Prior (2009) discusses the validity of self-reported news exposure measures at the center of media effects research. The author found 300% of overestimation of news consumption from self-reporting. This problem is no longer the main concern with regard to the validity of media consumption research thanks to novel data sources. Still, the validity of research output making use of these novel data, such as audience overlap network literature, has not gone without its setbacks. In this study, I have considered the importance of reliable methods to supply dependable networks for analysis. I have also explored the opportunities that the increasingly growing field of network science brings to the study of news media consumption.

As it had been previously pointed out by Mukerjee et al. (2018*a*), the methodological approach at the center of the audience overlap network research, deviation-from-random duplication, fails to assess the statistical significance of the ties in a network. The field lacked, however, a more detailed analysis of the method to put previous work in perspective. I made a comparison of deviation-from-random duplication with disparity filtering, a widespread method in network science to assess edge significance. As it was stated in the previous section, the observed results from the network at hand are not directly transferable to other networks. The outcome does, however, bring up questions about the extent to which the results of previous research are methodological artifacts. Although my analysis cannot offer conclusive evidence with regard to other studies, it calls to read previous findings with a degree of skepticism.

I explored the promises that the fusion of web-tracking and survey data brings to the study of media use. Thanks to the granularity of the individual level digital trace data, I was able to construct for the first time the consumer-end of the audience overlap network. The resulting adjacency matrix becomes meaningful as the survey data provide means to map the individual attributes of the consumers. The extracted backbone networks were employed to illustrate ways in which a network analytic approach can provide alternative perspectives to study consumption patterns.

First, I utilized Exponential Family Random Graph Models, a statistical analysis method specific to network data, to showcase new frames to assess substantive queries in news media usage patterns. The results hint that individuals at the ends of the political ideology spectrum are more likely to be

connected to ideological equivalents. To some extent in the observed network, birds of a feather flock together.

Second, I employed the communities detected through the Louvain method to examine how network measures can be incorporated into statistical inferential analysis. The communities extracted are significantly predictive of vote choice and to a lesser degree party identification. Modularity in the networks appears to capture underlying factors that relate to media consumption patterns and political opinions. The latter leads to a reflection about the theoretical frameworks and the construct validity of the measures to study phenomena in the high-choice digital news environment. A lot of the evidence points at moderately overlapping media diets and not an avoidance of cross-cutting media exposure, but digital trace data allow us to seek for more nuance beyond exposure metrics. It is perhaps not selective exposure to, rather "preferential engagement" with congenial content.

This study is not without limitations. The undirected valued conception of the networks treats media consumption as static. Future avenues for research open through the richness of individual digital trace data can explore information flows, as well as incorporating temporal dynamics (Borge-Holthoefer and González-Bailón, 2015). Since the data do not come with established mappings, a more comprehensive collection of news media outlets can lay out a more complete portrayal of news media consumption dynamics. Additionally, more intricate network modeling strategies that can better capture the complex dependence patterns and unobserved heterogeneity can provide further insight behind the micro-configurations that lead to the audience network.

CONCLUSION

The rise of the internet has brought about unprecedented dynamics in information creation, sharing, and consumption that have revolutionized the news media environment. The public debate is inundated with suggestions that these dynamics are in part responsible for the corrosion of the polity of many democratic states. The insinuation that the digital news environment is conducive to the creation of balkanized realities has concerned various scientific disciplines. So far the evidence suggests that the picture is one of overwhelming overlap between partisan and ideologically diverse audiences. This study puts in perspective a portion of the previous work that has employed problematic methodological decisions. It is crucial that in a time where sectors of the general public declare to be tired of

experts, that scientific rigor be the guiding force to assess what are the defining features of our time. New data sources allow for unparalleled analysis. These new data also call for new ways of exploring, assessing, and formulating scientific queries. This study presents an exploration of a course through which digital trace and survey data can be merged to gather insights on digital news consumption.

REFERENCES

All Sides. 2018. "Media bias ratings." *Allsides.com* .

Bakshy, Eytan, Solomon Messing and Lada A Adamic. 2015. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348(6239):1130–1132.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker and Richard Bonneau. 2015. "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological science* 26(10):1531–1542.

Bennett, W Lance and Shanto Iyengar. 2008. "A new era of minimal effects? The changing foundations of political communication." *Journal of communication* 58(4):707–731.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. 2008. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment* 2008(10):P10008.

Borgatti, Stephen P, Martin G Everett and Jeffrey C Johnson. 2018. *Analyzing social networks*. Sage.

Borge-Holthoefer, Javier and Sandra González-Bailón. 2015. "Scale, time, and activity patterns: Advanced methods for the analysis of online networks." *The SAGE handbook of online research methods* pp. 259–276.

Box-Steffensmeier, Janet M, Dino P Christenson and Jason W Morgan. 2018. "Modeling unobserved heterogeneity in social networks with the frailty exponential random graph model." *Political Analysis* 26(1):3–19.

Butts, Carter T. 2015. *network: Classes for Relational Data*. The Statnet Project (http://www.statnet.org). R package version 1.13.0.1.
URL: *https://CRAN.R-project.org/package=network*

Carpini, Michael X Delli and Scott Keeter. 1996. *What Americans know about politics and why it matters*. Yale University Press.

Csardi, Gabor and Tamas Nepusz. 2006. "The igraph software package for complex network research." *InterJournal* Complex Systems:1695.
URL: *http://igraph.org*

Festinger, Leon. 1957. *A theory of cognitive dissonance.* Stanford University Press.

Flaxman, Seth, Sharad Goel and Justin M Rao. 2016. "Filter bubbles, echo chambers, and online news consumption." *Public opinion quarterly* 80(S1):298–320.

Garrett, R Kelly. 2009*a*. "Echo chambers online?: Politically motivated selective exposure among Internet news users." *Journal of Computer-Mediated Communication* 14(2):265–285.

Garrett, R Kelly. 2009*b*. "Politically motivated reinforcement seeking: Reframing the selective exposure debate." *Journal of communication* 59(4):676–699.

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

Goodhardt, Gerald J. 1966. "Constant in duplicated television viewing." *Nature* 212(5070):1616–1616.

Goodhardt, Gerald J and Andrew SC Ehrenberg. 1969. "Duplication of television viewing between and within channels." *Journal of Marketing Research* 6(2):169–178.

Guess, Andrew M. 2019. "(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets." *Unpublished manuscript* .

Habermas, Jurgen. 1991. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky and Martina Morris. 2019. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (https://statnet.org). R package version 3.10.4.
URL: *https://CRAN.R-project.org/package=ergm*

Hindman, Matthew. 2008. *The myth of digital democracy*. Princeton University Press.

Hunter, David R and Mark S Handcock. 2006. "Inference in curved exponential family models for networks." *Journal of Computational and Graphical Statistics* 15(3):565–583.

Hunter, David R, Mark S Handcock, Carter T Butts, Steven M Goodreau and Martina Morris. 2008. "ergm: A package to fit, simulate and diagnose exponential-family models for networks." *Journal of statistical software* 24(3):nihpa54860.

Hunter, David R, Steven M Goodreau and Mark S Handcock. 2008. "Goodness of fit of social network models." *Journal of the American Statistical Association* 103(481):248–258.

Knobloch-Westerwick, Silvia, Cornelia Mothes and Nick Polavin. 2020. "Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information." *Communication Research* 47(1):104–124.

Ksiazek, Thomas B. 2011. "A network analytic approach to understanding cross-platform audience behavior." *Journal of Media Economics* 24(4):237–251.

Lusher, Dean, Johan Koskinen and Garry Robins. 2013. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Majó-Vázquez, Sílvia, Rasmus K Nielsen and Sandra González-Bailón. 2019. "The backbone structure of audience networks: A new approach to comparing online news consumption across countries." *Political Communication* 36(2):227–240.

McPherson, Miller, Lynn Smith-Lovin and James M Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual review of sociology* 27(1):415–444.

Mukerjee, Subhayan, Sílvia Majó-Vázquez and Sandra González-Bailón. 2018*a*. "Networks of audience overlap in the consumption of digital news." *Journal of Communication* 68(1):26–50.

Mukerjee, Subhayan, Sílvia Majó-Vázquez and Sandra González-Bailón. 2018*b*. "Response to Webster and Taneja's response to "Networks of audience overlap in the consumption of digital news"." *Journal of Communication* 68(3):E15–E18.

Mutz, Diana C. 2001. "Facilitating communication across lines of political difference: The role of mass media." *American Political Science Review* 95(1):97–114.

Nelson, Jacob L and James G Webster. 2017. "The myth of partisan selective exposure: A portrait of the online political news audience." *Social Media+ Society* 3(3).

Neuman, W Russell, Russell W Neuman, Marion R Just and Ann N Crigler. 1992. *Common knowledge: News and the construction of political meaning*. University of Chicago Press.

Pattison, Philippa and Stanley Wasserman. 1999. "Logit models and logistic regressions for social networks: II. Multivariate relations." *British Journal of Mathematical and Statistical Psychology* 52(2):169–193.

Prior, Markus. 2007. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Prior, Markus. 2009. "The immensely inflated news audience: Assessing bias in self-reported news exposure." *Public Opinion Quarterly* 73(1):130–143.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org*

Sears, David O and Jonathan L Freedman. 1967. "Selective exposure to information: A critical review." *Public Opinion Quarterly* 31(2):194–213.

Serrano, M Ángeles, Marián Boguná and Alessandro Vespignani. 2009. "Extracting the multiscale backbone of complex weighted networks." *Proceedings of the national academy of sciences* 106(16):6483–6488.

Snijders, Tom AB, Philippa E Pattison, Garry L Robins and Mark S Handcock. 2006. "New specifications for exponential random graph models." *Sociological methodology* 36(1):99–153.

Stier, Sebastian, Johannes Breuer, Pascal Siegers and Kjerstin Thorson. 2019. "Integrating survey data and digital trace data: key issues in developing an emerging field." *Social Science Computer Review* .

Stier, Sebastian, Nora Kirkizh, Caterina Froio and Ralph Schroeder. 2020. "Populist Attitudes and Selective Exposure to Online News: A Cross-Country Analysis Combining Web Tracking and Surveys." *The International Journal of Press/Politics* .

Stroud, Natalie Jomini. 2008. "Media use and political predispositions: Revisiting the concept of selective exposure." *Political Behavior* 30(3):341–366.

Sunstein, Cass R. 2001. *Republic.com*. Princeton university press.

Sunstein, Cass R. 2018. *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Taneja, Harsh. 2016. "Using commercial audience measurement data in academic research." *Communication Methods and Measures* 10(2-3):176–178.

Taneja, Harsh and James G Webster. 2016. "How do global audiences take shape? The role of institutions and culture in patterns of web use." *Journal of Communication* 66(1):161–182.

Teixeira, Filipe. 2018. *skynet: Generates Networks from BTS Data*. R package version 1.3.0. **URL:** *https://CRAN.R-project.org/package=skynet*

Wasserman, Stanley and Philippa Pattison. 1996. "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp." *Psychometrika* 61(3):401–425.

Webster, James G. 1985. "Program audience duplication: A study of television inheritance effects." *Journal of Broadcasting & Electronic Media* 29(2):121–133.

Webster, James G. 2014. *The marketplace of attention: How audiences take shape in a digital age*. MIT Press.

Webster, James G and Harsh Taneja. 2018. "Building and interpreting audience networks: A response to mukerjee, majo-vazquez & gonzalez-bailon." *Journal of Communication* 68(3):E11–E14.

Webster, James G and Thomas B Ksiazek. 2012. "The dynamics of audience fragmentation: Public attention in an age of digital media." *Journal of communication* 62(1):39–56.

Weeks, Brian E, Thomas B Ksiazek and R Lance Holbert. 2016. "Partisan enclaves or shared media experiences? A network approach to understanding citizens' political news environments." *Journal of Broadcasting & Electronic Media* 60(2):248–268.

STATEMENT OF AUTHORSHIP

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on May 25, 2020 is identical to the printed version to be submitted to the Examination Office.

Sebastian Ramirez Ruiz
Berlin
May 25, 2020

Mapping the Online News Environment: Leveraging survey and web-tracking data for audience networks

Appendix

APPENDIX A LIST OF OUTLETS

TABLE A1: List of included outlets with number of users and total reach

Outlet	Users	Reach	Outlet	Users	Reach
www.washingtonpost.com	501	44.85%	www.investors.com	39	3.49%
www.nytimes.com	489	43.78%	hotair.com	37	3.31%
www.foxnews.com	442	39.57%	jezebel.com	37	3.31%
www.cnn.com	402	35.99%	www.ocregister.com	37	3.31%
www.usatoday.com	359	32.14%	www.upworthy.com	36	3.22%
www.huffingtonpost.com	318	28.47%	www.rollcall.com	35	3.13%
www.nbcnews.com	308	27.57%	www.redstate.com	34	3.04%
www.cbsnews.com	306	27.39%	ijr.com	33	2.95%
www.businessinsider.com	297	26.59%	www.teenvogue.com	33	2.95%
www.cnet.com	266	23.81%	www.bostonherald.com	32	2.86%
thehill.com	263	23.55%	www.economist.com	32	2.86%
nypost.com	261	23.37%	www.scientificamerican.com	32	2.86%
www.dailymail.co.uk	261	23.37%	www.mediamatters.org	28	2.51%
www.theguardian.com	255	22.83%	www.theepochtimes.com	28	2.51%
news.google.com	253	22.65%	www.ft.com	27	2.42%
www.npr.org	237	21.22%	www.ibtimes.com	27	2.42%
www.forbes.com	233	20.86%	www.judicialwatch.org	27	2.42%
abcnews.go.com	231	20.68%	www.mcclatchydc.com	27	2.42%
www.cnbc.com	228	20.41%	www1.cbn.com	26	2.33%
www.politico.com	217	19.43%	www.propublica.org	24	2.15%
www.latimes.com	215	19.25%	www.rasmussenreports.com	24	2.15%
www.bbc.com	200	17.91%	www.sfchronicle.com	24	2.15%
time.com	199	17.82%	www.politicususa.com	23	2.06%
www.theatlantic.com	193	17.28%	newrepublic.com	22	1.97%
www.thedailybeast.com	191	17.1%	observer.com	22	1.97%
www.vox.com	186	16.65%	www.frontpagemag.com	22	1.97%
www.bloomberg.com	178	15.94%	www.pressherald.com	22	1.97%
www.newsweek.com	171	15.31%	boingboing.net	21	1.88%
www.wsj.com	168	15.04%	www.csmonitor.com	21	1.88%
www.chicagotribune.com	161	14.41%	www.democracynow.org	21	1.88%
www.newyorker.com	156	13.97%	www.jpost.com	21	1.88%
www.reuters.com	153	13.7%	www.theamericanconservative.com	21	1.88%
www.usnews.com	142	12.71%	www.cookpolitical.com	20	1.79%
www.rollingstone.com	136	12.18%	www.eurekalert.org	20	1.79%
www.marketwatch.com	132	11.82%	www.mtv.com	20	1.79%
www.pbs.org	132	11.82%	www.post-gazette.com	20	1.79%
www.telegraph.co.uk	128	11.46%	spectator.org	19	1.7%
www.independent.co.uk	127	11.37%	amgreatness.com	17	1.52%
nymag.com	126	11.28%	wgntv.com	17	1.52%
www.dailykos.com	124	11.1%	www.projectveritas.com	17	1.52%
www.dailywire.com	123	11.01%	www.ksl.com	16	1.43%
www.nydailynews.com	122	10.92%	www.truthorfiction.com	15	1.34%
www.buzzfeednews.com	114	10.21%	truthout.org	14	1.25%
www.miamiherald.com	113	10.12%	www.kqed.org	13	1.16%
dailycaller.com	112	10.03%	www.pri.org	13	1.16%

www.bustle.com	112	10.03%	www.richmond.com	13	1.16%
apnews.com	111	9.94%	www.tallahassee.com	12	1.07%
slate.com	110	9.85%	katu.com	11	0.98%
www.msnbc.com	110	9.85%	www.oann.com	11	0.98%
www.theverge.com	108	9.67%	www.spokesman.com	11	0.98%
www.breitbart.com	106	9.49%	www.thecollegefix.com	11	0.98%
mashable.com	104	9.31%	www.truthdig.com	11	0.98%
www.sfgate.com	104	9.31%	quillette.com	10	0.9%
www.today.com	104	9.31%	www.courier-journal.com	10	0.9%
www.washingtontimes.com	100	8.95%	www.statesman.com	10	0.9%
www.vanityfair.com	99	8.86%	washingtonmonthly.com	9	0.81%
www.washingtonexaminer.com	94	8.42%	www.countable.us	9	0.81%
www.westernjournal.com	94	8.42%	psmag.com	8	0.72%
www.realclearpolitics.com	91	8.15%	www.city-journal.org	8	0.72%
fivethirtyeight.com	90	8.06%	www.cjr.org	8	0.72%
www.bostonglobe.com	90	8.06%	www.commercialappeal.com	8	0.72%
www.rawstory.com	88	7.88%	www.defenseone.com	8	0.72%
townhall.com	87	7.79%	grist.org	7	0.63%
www.mercurynews.com	86	7.7%	www.currentaffairs.org	7	0.63%
lifehacker.com	85	7.61%	www.saturdayeveningpost.com	7	0.63%
thinkprogress.org	81	7.25%	www.yesmagazine.org	7	0.63%
www.thegatewaypundit.com	79	7.07%	ivn.us	6	0.54%
www.nationalreview.com	78	6.98%	www.dailyprogress.com	6	0.54%
www.esquire.com	76	6.8%	www.liveaction.org	6	0.54%
www.ajc.com	75	6.71%	www.thefiscaltimes.com	6	0.54%
www.politifact.com	75	6.71%	www.univision.com	6	0.54%
www.newsmax.com	73	6.54%	fair.org	5	0.45%
www.salon.com	73	6.54%	jacobinmag.com	5	0.45%
www.axios.com	70	6.27%	www.arkansasonline.com	5	0.45%
techcrunch.com	67	6%	www.conservativehq.com	5	0.45%
theweek.com	66	5.91%	www.texasobserver.org	5	0.45%
www.mediaite.com	65	5.82%	lasvegassun.com	4	0.36%
www.dallasnews.com	64	5.73%	vtdigger.org	4	0.36%
www.sacbee.com	63	5.64%	www.commentarymagazine.com	4	0.36%
thefederalist.com	62	5.55%	www.foreignaffairs.com	4	0.36%
qz.com	61	5.46%	www.whatfinger.com	4	0.36%
www.motherjones.com	60	5.37%	www.dailypress.com	3	0.27%
www.vice.com	60	5.37%	www.delcotimes.com	3	0.27%
www.azcentral.com	58	5.19%	www.mrc.org	3	0.27%
theintercept.com	57	5.1%	www.wgbh.org	3	0.27%
www.theblaze.com	57	5.1%	thelibertarianrepublic.com	2	0.18%
pjmedia.com	50	4.48%	www.civilbeat.org	2	0.18%
www.factcheck.org	50	4.48%	www.idsnews.com	2	0.18%
www.theroot.com	50	4.48%	www.nationaljournal.com	2	0.18%
www.alternet.org	49	4.39%	www.redandblack.com	2	0.18%
www.cnsnews.com	46	4.12%	www.timescall.com	2	0.18%
www.wnd.com	46	4.12%	www.wandtv.com	2	0.18%
www.c-span.org	45	4.03%	www.wisconsingazette.com	2	0.18%
freebeacon.com	44	3.94%	dailynorthwestern.com	1	0.09%
reason.com	44	3.94%	michellemalkin.com	1	0.09%
www.dailysignal.com	44	3.94%	news.mit.edu	1	0.09%
www.infowars.com	43	3.85%	news.yahoo.com	1	0.09%
splinternews.com	42	3.76%	rightwingnews.com	1	0.09%
www.drudgereport.com	42	3.76%	rsbnetwork.com	1	0.09%
chicago.suntimes.com	41	3.67%	www.ajplus.net	1	0.09%
www.weeklystandard.com	41	3.67%	www.better-angels.org	1	0.09%
www.sciencedaily.com	40	3.58%	www.bgdailynews.com	1	0.09%
www.thenation.com	40	3.58%	www.insidephilanthropy.com	1	0.09%
www.aljazeera.com	39	3.49%	www.intellectualconservative.com	1	0.09%
www.americanthinker.com	39	3.49%	www.myrecordjournal.com	1	0.09%

APPENDIX B STATISTICALLY SIGNIFICANT EDGE EXTRACTION

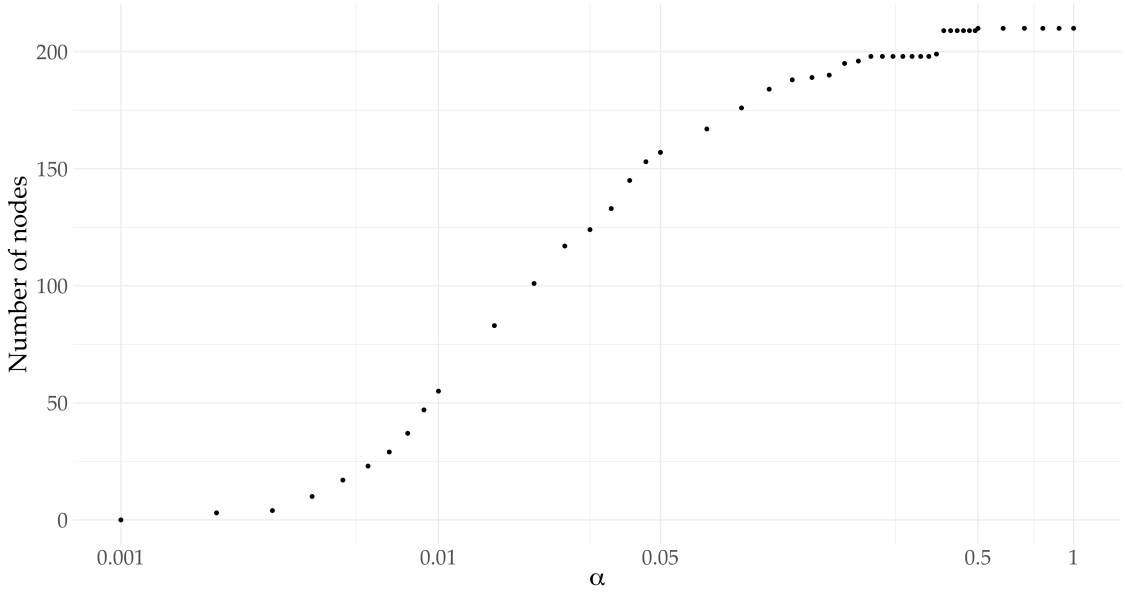


FIGURE B1: Node counts under different α specification for outlet-end network

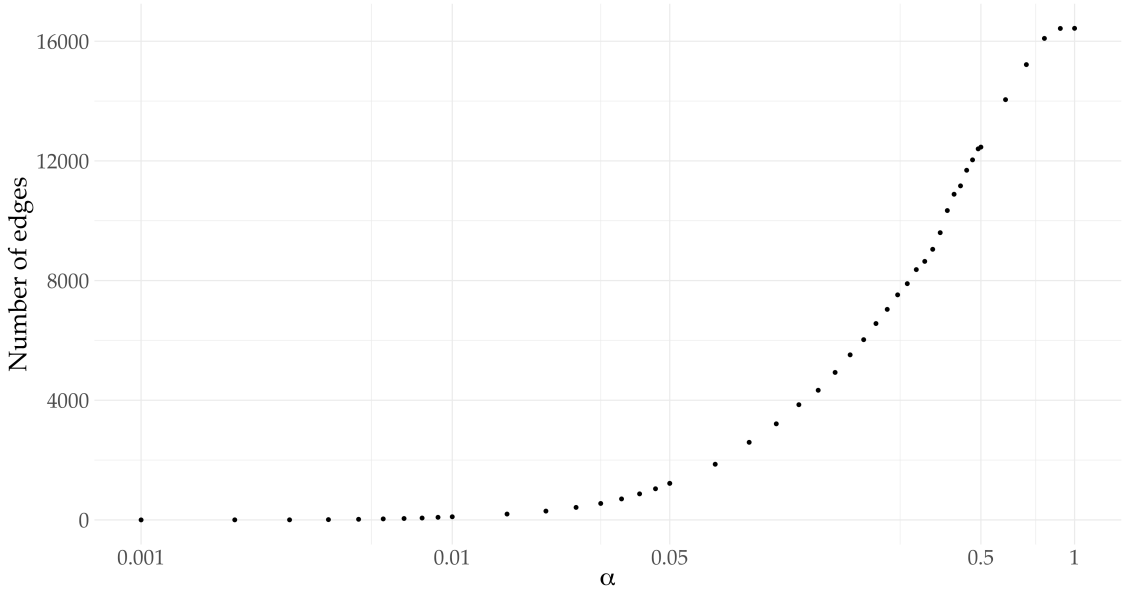


FIGURE B2: Edge counts under different α specification for outlet-end network

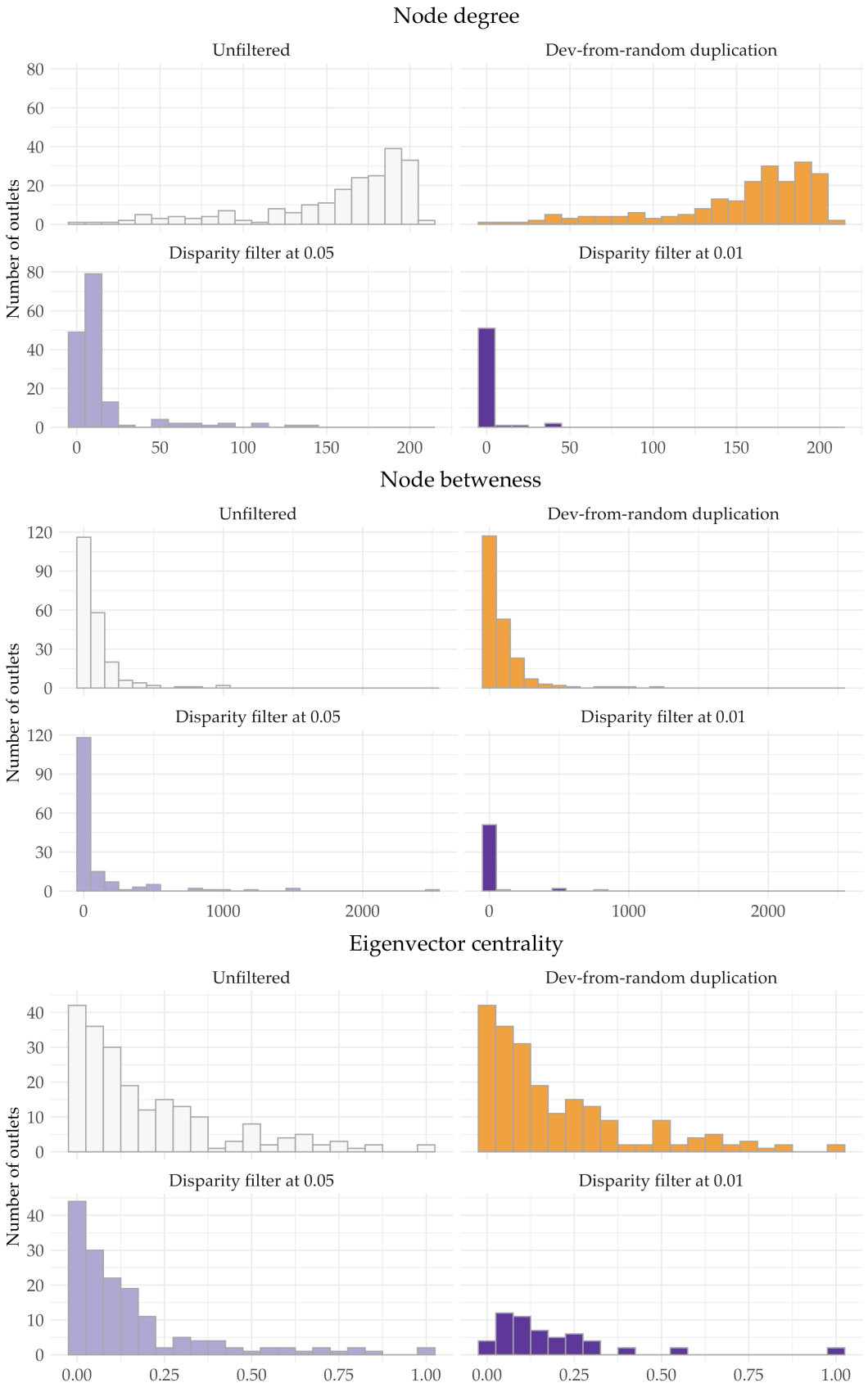


Figure B3: Distribution of node-level degree, betweenness and eigenvector centrality

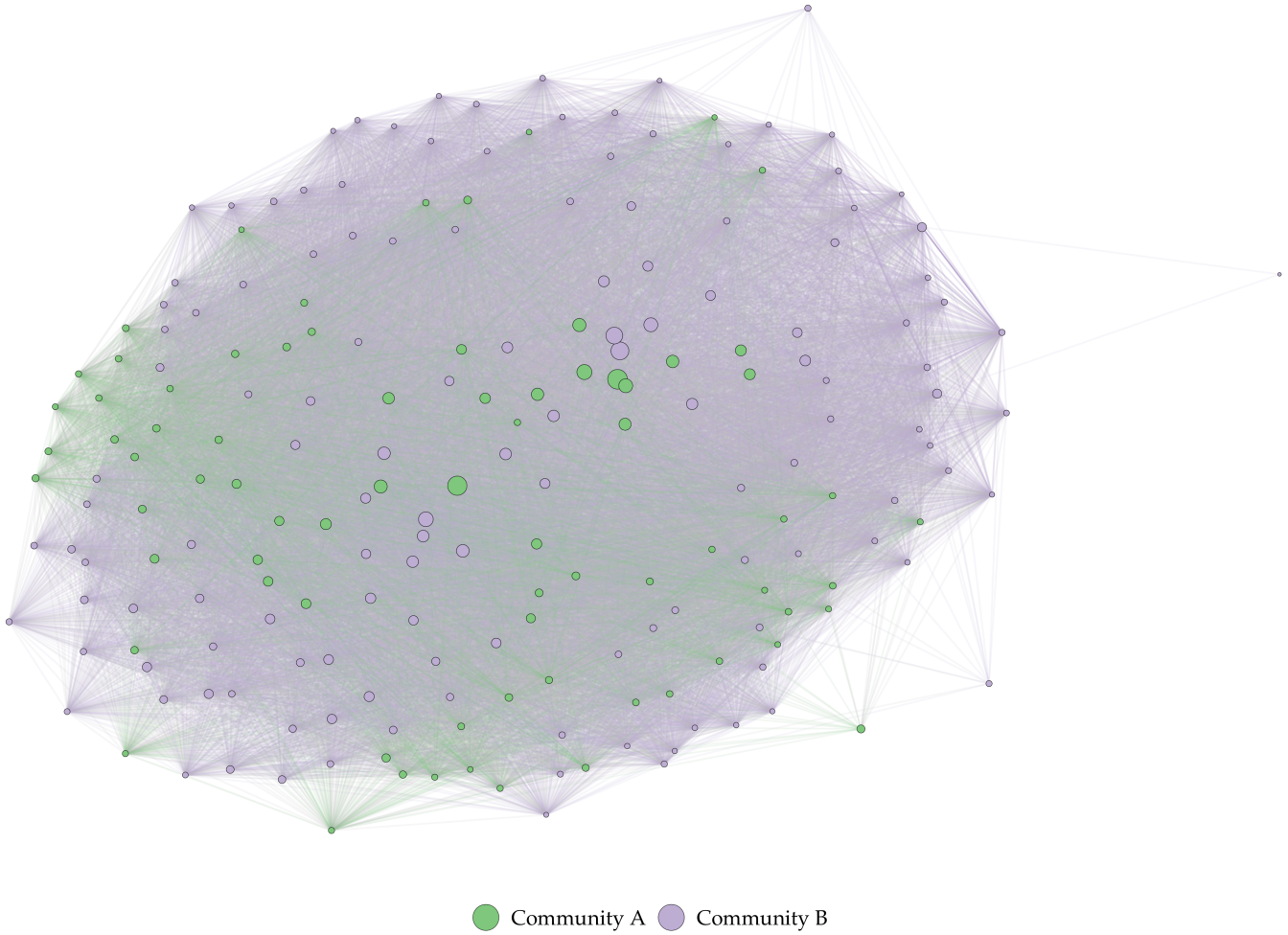


Figure B4: Graphical depiction of the outlet-end unfiltered network

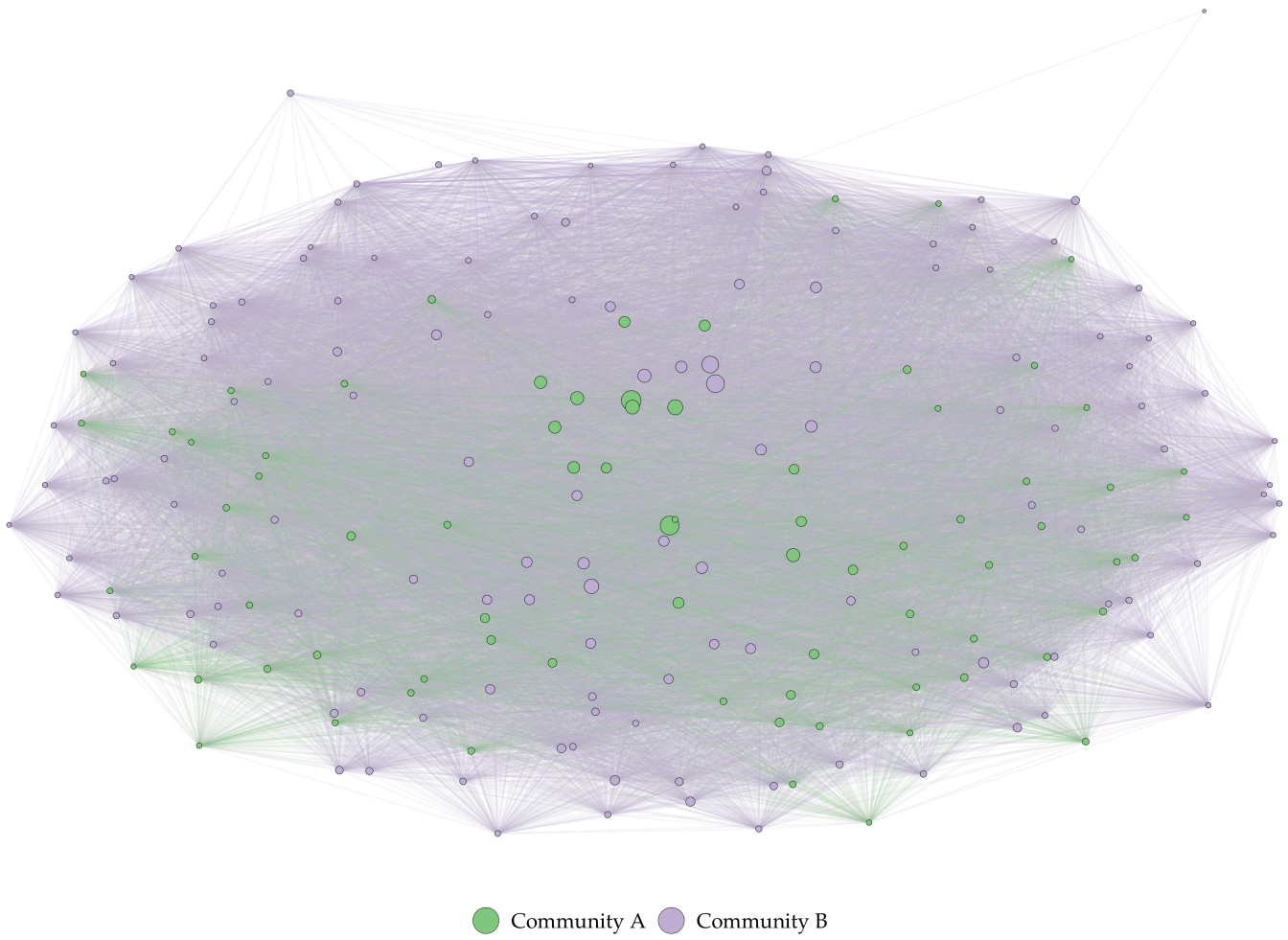


Figure B5: Graphical depiction of the outlet-end deviation from random duplication network

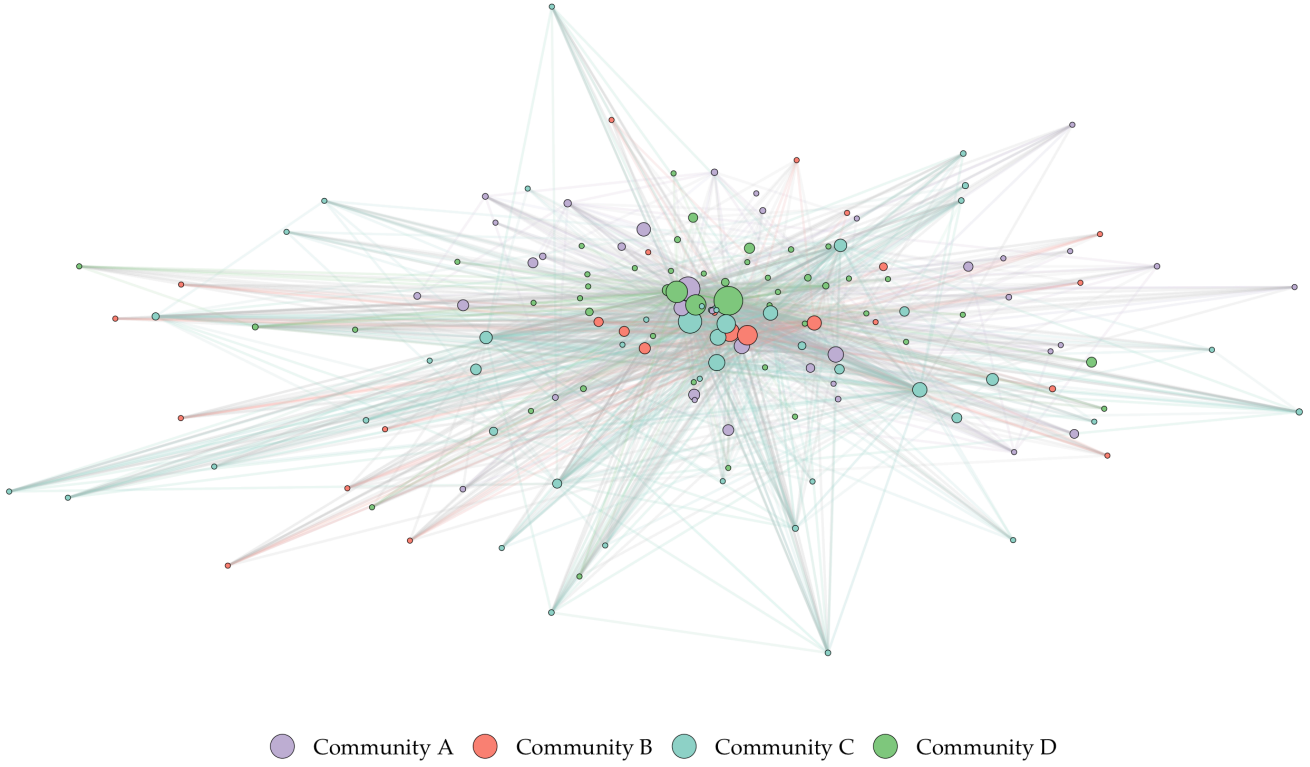


Figure B6: Graphical depiction of outlet-end backbone network at the 0.05 α significance level

Table B2: Communities obtained from backbone network after disparity filter at 0.01 α

Community 1			
cnn.com	huffingtonpost.com	nbcnews.com	today.com
forbes.com	msnbc.com		
Community 2			
dailycaller.com	cbsnews.com	foxnews.com	sfgate.com
nypost.com	cnet.com	mercurynews.com	usatoday.com
breitbart.com	dailymail.co.uk	miamiherald.com	usnews.com
businessinsider.com	dailywire.com	newsmax.com	westernjournal.com
Community 3			
abcnews.go.com	bloomberg.com	npr.org	thedailybeast.com
fivethirtyeight.com	bostonglobe.com	nytimes.com	theguardian.com
mashable.com	bustle.com	ocregister.com	theverge.com
news.google.com	chicagotribune.com	pbs.org	vox.com
nymag.com	cnbc.com	politico.com	washingtonpost.com
slate.com	dailykos.com	reuters.com	wsj.com
thehill.com	latimes.com	rollingstone.com	
time.com	newsweek.com	telegraph.co.uk	
bbc.com	newyorker.com	theatlantic.com	

Table B3: Communities obtained from backbone network after disparity filter at 0.05 α

Community 1			
amgreatness.com	townhall.com	www.frontpagemag.com	www.sfgate.com
chicago.suntimes.com	www.americanthinker.com	www.infowars.com	www.theblaze.com
dailycaller.com	www.azcentral.com	www.investors.com	www.thedailybeast.com
freebeacon.com	www.breitbart.com	www.mcclatchydc.com	www.thegatewaypundit.com
hotair.com	www.cnsnews.com	www.mercurynews.com	www.washingtontimes.com
ijr.com	www.dailymail.co.uk	www.miamiherald.com	www.westernjournal.com
nypost.com	www.dailysignal.com	www.mtv.com	www.wnd.com
pjmedia.com	www.dailywire.com	www.newsmax.com	
techcrunch.com	www.drudgereport.com	www.rasmussenreports.com	
thefederalist.com	www.foxnews.com	www.redstate.com	
Community 2			
apnews.com	www.bostonglobe.com	www.ft.com	www.newyorker.com
nymag.com	www.bostonherald.com	www.judicialwatch.org	www.nydailynews.com
spectator.org	www.c-span.org	www.mediaite.com	www.politico.com
thehill.com	www.cookpolitical.com	www.motherjones.com	www.rawstory.com
theintercept.com	www.dallasnews.com	www.msnbc.com	www.realclearpolitics.com
www.axios.com	www.eurekalert.org	www.nbcnews.com	www.washingtonexaminer.com
Community 3			
abcnews.go.com	www.businessinsider.com	www.huffingtonpost.com	www.tallahassee.com
fivethirtyeight.com	www.bustle.com	www.independent.co.uk	www.theatlantic.com
lifehacker.com	www.buzzfeednews.com	www.latimes.com	www.theguardian.com
mashable.com	www.cbsnews.com	www.newsweek.com	www.theroot.com
news.google.com	www.chicagotribune.com	www.npr.org	www.theverge.com
slate.com	www.cnet.com	www.politifact.com	www.today.com
thinkprogress.org	www.cnn.com	www.reuters.com	www.upworthy.com
time.com	www.countable.us	www.richmond.com	www.usatoday.com
www.ajc.com	www.dailykos.com	www.rollingstone.com	www.usnews.com
www.alternet.org	www.esquire.com	www.salon.com	www.vanityfair.com

www.bbc.com	www.factcheck.org	www.saturdayeveningpost.com	www.vice.com
www.bloomberg.com	www.forbes.com	www.scientificamerican.com	www.vox.com

Community 4			
jezebel.com	www.cnbc.com	www.nationalreview.com	www.sfchronicle.com
katu.com	www.commercialappeal.com	www.nytimes.com	www.spokesman.com
newrepublic.com	www.courier-journal.com	www.ocregister.com	www.teenvogue.com
observer.com	www.csmonitor.com	www.pbs.org	www.telegraph.co.uk
quillette.com	www.currentaffairs.org	www.politicususa.com	www.theepochtimes.com
qz.com	www.democracynow.org	www.post-gazette.com	www.thenation.com
reason.com	www.economist.com	www.pressherald.com	www.truthdig.com
splinternews.com	www.ibtimes.com	www.pri.org	www.truthorfiction.com
theweek.com	www.kqed.org	www.propublica.org	www.washingtonpost.com
truthout.org	www.ksl.com	www.rollcall.com	www.weeklystandard.com
wgntv.com	www.marketwatch.com	www.sacbee.com	www.wsj.com
www.aljazeera.com	www.mediamatters.org	www.sciencedaily.com	www1.cbn.com

APPENDIX C IDEOLOGICAL HOMOPHILY

	0.01 α	0.05 α
Edges	-5.11^{***}	-3.06^{***}
	(0.26)	(0.04)
DH - Very Liberal	1.23^{***}	0.30^{***}
	(0.22)	(0.05)
DH - Liberal	0.33	0.22^{***}
	(0.25)	(0.06)
DH - Moderate	−0.91	−0.12
	(0.48)	(0.06)
DH - Conservative	0.05	0.57^{***}
	(0.22)	(0.06)
DH - Very Conservative	1.00^{***}	0.59^{***}
	(0.19)	(0.06)
NF - Liberal	0.92^{***}	−0.03
	(0.12)	(0.03)
NF - Moderate	0.12	-0.18^{***}
	(0.12)	(0.03)
NF - Conservative	0.97^{***}	-0.12^{***}
	(0.11)	(0.03)
NF - Very Conservative	0.67^{***}	0.23^{***}
	(0.12)	(0.03)
NF - Very interested in politics	0.78^{***}	0.44^{***}
	(0.16)	(0.03)
NF - Interested in politics	0.58^{***}	0.12^{***}
	(0.17)	(0.03)
NF - Moderately interested in politics	0.02	0.12^{**}
	(0.22)	(0.04)
NF - Slightly interested in politics	-1.49^{**}	-0.68^{***}
	(0.48)	(0.05)
UH - Education	−0.02	-0.05^{*}
	(0.10)	(0.02)
UH - Race	-0.45^{***}	-0.16^{***}
	(0.09)	(0.02)
UH - Region	0.18^{*}	0.05^{*}
	(0.09)	(0.02)
AIC	5557.85	100318.68
BIC	5687.18	100492.45
Log Likelihood	-2761.92	-50142.34

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE C4: Model output of Exponential Random Graph Models for backbone consumer-end networks at the 0.01 and 0.05 significance levels.

Note. DF — Differential homophily; NF — Node factor; UH — Uniform homophily. The baseline for the nodal factors of ideology and interest in politics are *Very liberal* and *Not interested at all.*

APPENDIX D PREDICTIVE CLUSTERS

D.1 CONSUMER-END 0.01 BACKBONE NETWORK

TABLE D5: Vote for Donald Trump — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.41	[0.23,0.62]	Community A	0.41	[0.23,0.62]
Community B	0.06	[0.03,0.14]	Community B	0.06	[0.03,0.14]
Community C	0.89	[0.78,0.95]	Community C	0.90	[0.78,0.95]

TABLE D6: Vote for Hillary Clinton — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.36	[0.19,0.58]	Community A	0.36	[0.19,0.58]
Community B	0.78	[0.68,0.86]	Community B	0.78	[0.67,0.86]
Community C	0.04	[0.01,0.13]	Community C	0.03	[0.01,0.13]

TABLE D7: Self-identified Republican — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.36	[0.19,0.58]	Community A	0.37	[0.2,0.59]
Community B	0.04	[0.01,0.11]	Community B	0.04	[0.01,0.1]
Community C	0.49	[0.36,0.62]	Community C	0.50	[0.37,0.63]

TABLE D8: Self-identified Democrat — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.23	[0.1,0.44]	Community A	0.23	[0.1,0.45]
Community B	0.63	[0.52,0.73]	Community B	0.62	[0.5,0.72]
Community C	0.04	[0.01,0.13]	Community C	0.04	[0.01,0.13]

The demographic controls consist of race, sex, and educational level. The conditional predicted probabilities were extracted with the *ggeffects* package (Lüdecke, 2018).

Figure D7: Vote choice — Proportion of individuals per attribute held by each community

Figure D8: Party Identification — Proportion of individuals per attribute held by each community

D.2 CONSUMER-END 0.05 BACKBONE NETWORK

Table D9: Vote for Donald Trump — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.35	[0.18,0.57]	Community A	0.33	[0.16,0.55]
Community B	0.38	[0.27,0.49]	Community B	0.37	[0.27,0.49]
Community C	0.80	[0.73,0.85]	Community C	0.80	[0.74,0.86]
Community D	0.11	[0.08,0.15]	Community D	0.11	[0.08,0.14]

Table D10: Vote for Hillary Clinton — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.60	[0.38,0.79]	Community A	0.62	[0.39,0.81]
Community B	0.47	[0.36,0.59]	Community B	0.46	[0.35,0.58]
Community C	0.08	[0.05,0.13]	Community C	0.07	[0.04,0.12]
Community D	0.76	[0.71,0.8]	Community D	0.77	[0.72,0.81]

Table D11: Self-identified Republican — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.25	[0.11,0.48]	Community A	0.22	[0.09,0.45]
Community B	0.31	[0.21,0.42]	Community B	0.30	[0.2,0.42]
Community C	0.51	[0.44,0.59]	Community C	0.53	[0.45,0.6]
Community D	0.06	[0.04,0.1]	Community D	0.06	[0.04,0.09]

Table D12: Self-identified Democrat — Predicted probabilities of community models

Communities			Communities + Demographic controls		
Community	Predicted	CI	Community	Predicted	CI
Community A	0.60	[0.38,0.79]	Community A	0.59	[0.37,0.78]
Community B	0.28	[0.19,0.39]	Community B	0.27	[0.18,0.39]
Community C	0.08	[0.05,0.14]	Community C	0.09	[0.05,0.14]
Community D	0.60	[0.55,0.66]	Community D	0.60	[0.54,0.65]

Figure D9: Vote choice — Proportion of individuals per attribute held by each community

Figure D10: Party Identification — Proportion of individuals per attribute held by each community

APPENDIX E SOFTWARE STATEMENT

The entire analysis was run under OS X 10.12.6 using R version 3.5.1 R Core Team (2018). In the empirical analysis, I employed of the following R software packages:

`broom` (Robinson and Hayes, 2018),
`cowplot` (Wilke, 2018),
`data.table` (Dowle and Srinivasan, 2018),
`dotwhisker` (Solt and Hu, 2018),
`dplyr` (Wickham et al., 2019),
`ergm` (Handcock et al., 2019),
`ggeffects` (Lüdecke, 2018),
`ggnewscale` (Campitelli, 2020),
`ggplot2` (Wickham, 2016),
`ggpubr` (Kassambara, 2018),
`ggthemes` (Arnold, 2018),
`igraph` (Csardi and Nepusz, 2006),
`intergraph` (Bojanowski, 2015),
`skynet` (Teixeira, 2018),
`statnet` (Butts, 2015),
`texreg` (Leifeld, 2013),
`tibble` (Müller and Wickham, 2019),
`xtable` (Dahl et al., 2018)

REFERENCES

Arnold, Jeffrey B. 2018. *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 4.0.1.
URL: *https://CRAN.R-project.org/package=ggthemes*

Bojanowski, Michal. 2015. *intergraph: Coercion Routines for Network Data Objects*. R package version 2.0-2.
URL: *http://mbojan.github.io/intergraph*

Butts, Carter T. 2015. *network: Classes for Relational Data*. The Statnet Project (http://www.statnet.org). R package version 1.13.0.1.
URL: *https://CRAN.R-project.org/package=network*

Campitelli, Elio. 2020. *ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'*. R package version 0.4.1.
URL: *https://CRAN.R-project.org/package=ggnewscale*

Csardi, Gabor and Tamas Nepusz. 2006. "The igraph software package for complex network research." *InterJournal* Complex Systems:1695.
URL: *http://igraph.org*

Dahl, David B., David Scott, Charles Roosen, Arni Magnusson and Jonathan Swinton. 2018. *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-3.
URL: *https://CRAN.R-project.org/package=xtable*

Dowle, Matt and Arun Srinivasan. 2018. *data.table: Extension of 'data.frame'*. R package version 1.11.8.
URL: *https://CRAN.R-project.org/package=data.table*

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky and Martina Morris. 2019. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (https://statnet.org). R package version 3.10.4.
URL: *https://CRAN.R-project.org/package=ergm*

Kassambara, Alboukadel. 2018. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.2.
URL: *https://CRAN.R-project.org/package=ggpubr*

Leifeld, Philip. 2013. "texreg: Conversion of Statistical Model Output in R to LaTeX and HTML Tables." *Journal of Statistical Software* 55(8):1–24.
URL: *http://www.jstatsoft.org/v55/i08/*

Lüdecke, Daniel. 2018. "ggeffects: Tidy Data Frames of Marginal Effects from Regression Models." *Journal of Open Source Software* 3(26):772.

Müller, Kirill and Hadley Wickham. 2019. *tibble: Simple Data Frames*. R package version 2.1.3.
URL: *https://CRAN.R-project.org/package=tibble*

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org*

Robinson, David and Alex Hayes. 2018. *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R package version 0.5.1.
URL: *https://CRAN.R-project.org/package=broom*

Solt, Frederick and Yue Hu. 2018. *dotwhisker: Dot-and-Whisker Plots of Regression Results*. R package version 0.5.0.
URL: *https://CRAN.R-project.org/package=dotwhisker*

Teixeira, Filipe. 2018. *skynet: Generates Networks from BTS Data*. R package version 1.3.0.
URL: *https://CRAN.R-project.org/package=skynet*

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Wickham, Hadley, Romain François, Lionel Henry and Kirill Müller. 2019. *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3.
URL: *https://CRAN.R-project.org/package=dplyr*

Wilke, Claus O. 2018. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 0.9.3.
URL: *https://CRAN.R-project.org/package=cowplot*