Networks of Audience Overlap in the Consumption of Digital News

Subhayan Mukerjee*, Sílvia Majó-Vázquez**, and Sandra González-Bailón*

*Annenberg School for Communication, University of Pennsylvania
** Reuters Institute for the Study of Journalism, University of Oxford

Abstract: How do people consume news online? Here, we propose a novel way to answer this question using the browsing behavior of web users and the networks they form while navigating news content. In these networks, two news outlets are connected if they share a fraction of their audiences. We propose two crucial improvements to the methodology employed in previous research: a statistical test to filter out non-significant overlap between sites; and a thresholding approach to identify the core of the audience network. We explain why our approach is better than previous approaches using two data sets: one tracks digital news consumption during the 2016 Brexit referendum in the UK and the other during the 2016 Presidential Election in the US. We show that our filtering technique produces a completely different ranking of top sites, uncovering structural properties in the audience network that would go unnoticed otherwise.

Keywords: news; web; fragmentation; centralization; network analysis; legacy media; digital-born media.

Corresponding author: Sandra González-Bailón, Annenberg School for Communication, University of Pennsylvania, 3620 Walnut Street, PA 19104, Philadelphia, U.S.

Email: sgonzalezbailon@asc.upenn.edu

Abstract

How do people consume news online? Here, we propose a novel way to answer this question using the browsing behavior of web users and the networks they form while navigating news content. In these networks, two news outlets are connected if they share a fraction of their audiences. We propose two crucial improvements to the methodology employed in previous research: a statistical test to filter out non-significant overlap between sites; and a thresholding approach to identify the core of the audience network. We explain why our approach is better than previous approaches using two data sets: one tracks digital news consumption during the 2016 Brexit referendum in the UK and the other during the 2016 Presidential Election in the US. We show that our filtering technique produces a completely different ranking of top sites, uncovering structural properties in the audience network that would go unnoticed otherwise.

Keywords: digital news; web; fragmentation; centralization; network analysis; legacy media; digital-born media.

Networks of Audience Overlap in the Consumption of Digital News

The way in which the public consumes news reveals important information of how they engage with the political process. In information-rich environments (which today characterize most liberal democracies) the selection of media sources reveals individual preferences and propensities: some individuals will be more inclined to actively consume news and compare the information provided by different media outlets; others will consume a narrower set of sources, consistent with their predefined political attitudes; and yet others will be exposed to political news only as a byproduct of other activities (Prior, 2007). On the aggregate, these individual behaviors help us determine the levels of selective exposure and general political knowledge in any given society, which in turn help us understand the formation of public opinion and, more generally, the quality of those democracies as they operate in practice. Most normative theories of democracy rely, in the end, on informed citizens: "Political information", one analogy goes, "is to democratic politics what money is to economics: it is the currency of citizenship" (Delli Carpini & Keeter, 1996, p. 8). The way in which people consume news tells us, in other words, how many resources they have available to invest in the democratic process, to improve their understanding of the political world, or to take informed decisions.

The way in which technologies shape news consumption, on the other hand, gives us information about how the media environment constrains or diversifies those individual choices – which do not take place in a vacuum but respond, instead, to the institutions, regulations, and market forces that determine which news outlets exist, how they can be accessed, and whether or not they prosper. The media environment has been defined as "the different media sources routinely available to people at any point in time" (Prior, 2007, pp. 28-29). Most research on news consumption has focused on individual attributes like demographics or preferences not because they are more important than the features that define the media environment (i.e. features that determine the opportunity structure for news consumption) but because they are easier to measure, mostly through surveys. Yet prevalent measures of media exposure are known to suffer from the many issues of reliability and validity that derive from self-reported data (de Vreese & Neijens, 2016; Dilliplane, Goldman, & Mutz, 2013; Goldman, Mutz, & Dilliplane, 2013; Prior, 2009a, 2013). Imperfect recall of the media consumed is one part of the problem; the other is that survey instruments impose a limit on the number of media outlets that can be

analyzed – effectively cutting out the long tail of media options that is so characteristic of high-choice environments like the web and, more broadly, the internet. The conventional way in which the literature differentiates media environments boils down to the poor-choice/high-choice dichotomy – a binary classification that sums up the transition from mass media to digital forms of political communication (Delli Carpini & Keeter, 1996; Neuman, 2016; Prior, 2007; Williams & Delli Carpini, 2011). This dichotomy, however, hides the diversity that high-choice media environments still contain and limits our ability to understand news consumption in the digital domain.

Here, we propose a new way to characterize the diversity of media environments by mapping them through the choices that audiences make, given the available sources. Our goal is to devise a methodology that helps us map the media landscape in a way that can be compared across countries in a statistically robust way. This methodology captures not only the number of news outlets available for consumption, but how audiences navigate those outlets. There are two aspects of the media environment that interest us: fragmentation, or the extent to which different audiences self-select around specific news outlets given the higher number of available options; and centralization, or the extent to which audience attention concentrates around a few news sources, regardless of how long the tail is or how many information "niches" are contained in that tail (Anderson, 2006; Stroud, 2011). As with all empirical research, the way in which theoretical concepts like fragmentation or centralization are measured has a great impact on the interpretation of findings: a flawed measurement will lead to flawed conclusions. The analytical approach we take here aims to overcome the limitations of previous research that also looks at online audience behavior (e.g., Ksiazek, 2011; Taneja & Webster, 2016; Webster & Ksiazek, 2012). While our overall goal is similar (i.e. build networks of audience overlap to assess levels of fragmentation and centralization) we differ substantially in how we execute that goal – a departure that, as the rest of this article argues, has important implications for how we theorize about audience behavior in this digital age.

Prior literature suggests that digital technologies have increased fragmentation or the ability to self-select, and that they have reduced the media centralization so characteristic of the broadcast era – a consequential shift because centralization in the consumption of news guaranteed common exposure and the existence of a public domain where all citizens converged

(Katz, 1996; Prior, 2007; Stroud, 2011; Sunstein, 2001, 2007, 2017; Turow, 2012). We test this assumption of fragmentation with the use of observational data tracking online news consumption in two countries: the US and the UK. We focus on these two countries because they represent radically different media environments: the UK has a long history of public service media that is widely used and well-funded; the US, on the other hand, is dominated by private corporations, and characterized by a highly atomized supply and very weak government intervention. These differences in the media environment are very difficult to capture by studies that only pay attention to individual-level attributes and self-reported accounts of news exposure.

Our approach aims to capture those environmental characteristics through the analysis of aggregated networks of audience overlap among news sources, and the calculation of robust statistics that can be compared across countries. Changes in the structure of those networks can be quantified to pin down the way in which media environments differ – beyond the number of media outlets available. We propose, in other words, using network measures to compare media environments and summarize their features with statistics that translate their differences to the same scale. Before we describe the details of our approach, however, we give an overview of past research that has also made use of audience data to answer questions about media consumption patterns.

Audience Duplication Research

The analysis of audience duplication data has a long history in media research. In the 1960s, for instance, researchers analyzed the percentage of the audience of a TV program that also watched another program to develop metrics of cross-channel and within-channel duplication (Goodhardt & Ehrenberg, 1969). These metrics were mostly used to advance advertising goals, but the measurements also helped uncover "patterns of viewing" (Goodhardt, Ehrenberg, & Collins, 1975) that led to indices of brand loyalty and inheritance effects (Webster, 1985). The same measurements also informed theories of democratic engagement, for which media choices and exposure to news are important predictors of political attitudes and behavior (Delli Carpini & Keeter, 1996; Prior, 2007). Audience duplication is, in the end, a measure of cross-cutting exposure and of polarization when it refers to news outlets that have opposing ideological slants.

Since the early days of audience duplication research different proposals have been made to adapt the measurements to an increasingly complex media environment (i.e., Cooper, 1996). A recent approach suggests analyzing audience duplication data as a network, where media outlets are the nodes and the level of duplication between their audiences are represented by ties (Ksiazek, 2011). One of the advantages of this approach is that it allows a more sophisticated empirical measure of fragmentation. Traditionally, fragmentation has been operationalized as the number of media outlets that are available for consumption. This number has drastically increased in recent years given the sharp reduction in the costs of publishing news online and the consequent rise of niche content (Anderson, 2006; Turow, 2012). The analysis of audience overlap networks, on the other hand, pays attention to how audiences cluster around media outlets. Instead of measuring the number of different outlets available for consumption, the analysis of audience networks identifies groups of outlets that are consumed by the same people.

Research following the network approach to the analysis of audience overlap has already cast serious doubts on the fragmentation hypothesis. One of the indicators used, network centralization, measures how concentrated audiences are around a few media outlets (Ksiazek, 2011, pp. 245-246). The analysis of audience data for broadcast, TV channels, and Internet brands during March 2009 revealed a low centralization score "and, thus, a tendency for audiences to duplicate relatively evenly across the outlets in the media network" (p. 246). A later study using the same data concluded that "the people who use any given TV channel or Web site are disproportionately represented in the audience for most other outlets", adding that this result was "consistent with recent research that finds little evidence of ideological segmentation in media use" (Webster & Ksiazek, 2012, p. 50). These results were also echoed as evidence of "a massively overlapping culture", an expression used to describe the fact that "when we look at what people actually do, rather than what they say they do, we find a surprising, if rather messy, commonality to cultural consumption" (Webster, 2014, p. 20). Audience overlap, in other words, is prevalent across and within media channels but it goes undetected if we only pay attention to the number of media outlets available, not to how people navigate those outlets.

A more recent study, published in 2016, shifted attention from national to international media by looking at the top 1,000 global Web domains, ranked by monthly unique users during June 2012 (Taneja & Webster, 2016). The authors found that the centralization score in this

network was "moderately high", which responded to the fact that "there are a relatively small number of sites that do get links from most sites"; as the authors explained, "this is not surprising given that the top websites (…) get a disproportionately high amount of traffic and, consequently, have audience overlaps with large number of sites" (p. 171). Their findings also suggested that (unsurprisingly) language and geographic similarities are the most powerful predictors of audience overlap – which confirms the intuitive fact that people living in the same country tend to consume the same media. A related study published in the same year with the same data reports the same conclusions, namely that the audience overlap network "exhibits the tendency of audiences to gravitate toward culturally proximate content" (Taneja, 2016, p. 15).

These studies, summing up, show that the analysis of audience overlap data can offer relevant insights on exposure to information and media diets. They also show that the network analysis of audience behavior can offer a powerful methodological approach to uncover the characteristics of media diets and the venues where audiences concentrate more clearly. However, we argue that these prior studies are significantly weakened by the choices made in the construction of the audience networks. The goal of the analyses that follow is to describe these problems, propose a solution, and illustrate why our approach to the analysis of audience overlap offers valuable empirical insights that would go unnoticed otherwise.

Unlike the prior work just described, our data focuses exclusively on the consumption of news media content. We analyze novel data collected around two major political events: the 2016 Brexit referendum in the UK, and the 2016 Presidential Election in the US. In the analysis section we show that centralization scores (and, by extension, the overall network structure) change substantially before and after insignificant ties are removed – a test of statistical significance that was missing in prior work. We also show that the strength of the overlap, also disregarded in prior research, is crucial to understand the structure of news consumption patterns and, in particular, to identify a cluster of media sites that attract a disproportionate amount of attention from online audiences. The following section offers details of how audience networks were analyzed in the past – and why that approach should be improved.

The Analysis of Audience Networks

Audience overlap networks are the media-level projection of the bipartite network linking individuals with the media they consume, as illustrated in Figure 1, panel A. Bipartite networks are a conventional approach in network science to map affiliation ties (Newman, 2010, p. 123 and ff.; Wasserman & Faust, 1994, p. 298 and ff.). In this case, affiliation ties map the association between individuals and media outlets. In the most basic definition, ties in the single mode projections exist when two media outlets (top) or two individual consumers (bottom) share at least one consumer or media, respectively. The analysis of audience networks has, so far, focused on the media projection, where nodes are media outlets and ties measure the number of consumers they share – that is, the strength of their overlap. The analysis of this network allows measuring fragmentation in a more sophisticated way than just counting the number of media outlets available or the reach of those outlets. For instance, in the schematic representation of Figure 1, panel A, the media network has two components formed by outlets that share no audience: in this hypothetical scenario, fragmentation results from the decision of individual consumers to select into specific outlets. Whether such patterns of fragmentation emerge from observed news consumption, or whether they differ across national media landscapes are, of course, empirical questions that require measurement and analysis.

-- Figure 1 about here –

Prior Work

The first paper to use audience duplication data to build media networks was published in 2011 (Ksiazek, 2011). In this paper, the author makes a distinction between primary duplication and absolute duplication. He defines the first as "the percentage of the audience of one outlet that is also exposed to a second outlet" (p. 239). This measure translates into a directed tie, as illustrated in Figure 1, panel B. As Ksiazek specifies, "when using primary duplication, note that it makes a difference which outlet is treated first (i) and second (j)". The reason is that the base reach (i.e. the amount of monthly visitors) for each outlet differs so the percentage that is exposed to a second outlet will consequently change as well: an overlap of 100 readers will be 10% of site's reach with a total audience of 1,000 people, but it will be 50% for an outlet with 200 readers. Absolute duplication, on the other hand, is defined as "the percentage of the total audience that is exposed to both outlets in a given pair" (p. 240). This measure translates into an

undirected tie, as illustrated in Figure 1, panel C. What changes in this operationalization is the denominator of the overlap: in this case, it refers to total audience (for instance, the size of the online population), not to the audience of one specific outlet (e.g. the reach of the media sites in the dyad). The distinction between what Ksiazek calls primary and absolute duplication is important because the analytical techniques applied to the network will vary if the ties are directed or undirected.

In both instances, the author operates with percentages. This choice masks the total magnitude of the overlap – and makes the term "absolute overlap" misleading: in his metrics, overlap is always relative to a base that is hidden because of the way in which percentages transform frequencies. This causes a serious problem for his third operationalization of audience duplication, which Ksiazek calls "deviation from-random duplication". This measure is intended to indicate "the degree to which the observed duplication between two outlets differs from the expected duplication between those outlets" (p. 240). He calculates the difference between observed and expected overlap following the numerator of the formula used to calculate the chi-squared statistic but with one important caveat: he does not use frequencies but percentages.

In Ksiazek's words, "adapted to audience duplication, the row and column marginals are the *reach percentage* of the two outlets (i.e. the percentage of the audience that is exposed to each outlet separately), and N is equal to 100%, or the maximum reach percentage points for any given outlet" (p. 241). What the author does not explain is that those percentages refer to very different quantities because media outlets differ greatly in their overall total reach, i.e. some may have a reach of 1 million users, others of just a couple of thousand unique visitors; and this diversity in reach crucially affects the strength of the overlap between media outlets. The decision to use percentages instead of frequencies disregards, in other words, the substantive differences in reach. Hence, this approach treats as equivalent the least and the most accessed outlets, thus offering a misleading interpretation of the strength of the overlap – a weakness that affects the rest of the analyses applied to the network. Ksiazek's approach also seems oblivious to the fact that the chi-squared test requires working with frequencies, not percentages, because only then differences in size across groups can be appropriately taken into account.

In addition to this important limitation, this prior research is also limited in other respects. First, it makes the wrong assumption that the difference between expected and observed overlap is enough to assess statistical significance. To quote again the original paper, "a positive value indicates a non-random tendency to attend to a given outlet, a negative value would suggest the opposite (i.e., a tendency to avoid a given outlet), relative to the entire audience. Thus, the cell values in the resulting Primary Duplication-from-Random Duplication matrix indicate the degree to which the observed duplication deviates from random and, thus, also indicate varying levels of attendance or avoidance of particular outlets" (p. 242). However, the analyses reported offer no test to determine whether the departure from the null hypothesis of no difference is actually significant.

In a conventional test of independence, the chi-squared statistic summarizes how close the expected frequencies fall to the observed frequencies: the larger the statistic, the larger the difference is, and the greater the evidence against the null hypothesis. The sampling distribution of the chi-squared statistic indicates how large the statistic must be before we can conclude that the evidence is strong enough to reject the null hypothesis. The shape of this distribution depends on the degrees of freedom (df). For $df = 2$, for instance, the chi-squared statistic needs to be at least $X^2 = 5.99$ for a probability $p = 0.05$. The convention, then, is that if the summary of differences between observed and expected frequencies is less than 5.99, those differences are assumed to be statistically insignificant (for a $p = 0.05$).

Ksiazek offers no criterion to determine whether the differences he calculates are significant (assuming they are meaningful, since, again, they are calculated on the basis of percentages, not frequencies). Considering that media measurement companies like Nielsen or comScore (the main sources of audience data) produce estimates based on samples, and that 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. Instead, Ksiazek does the following: "I dichotomize the Primary Deviation-from-Random Duplication matrix so that values > 0 (i.e., the observed duplication is greater than expected) are converted to a value of '1', and those < 0 are converted to '0'" (p. 246). He then proceeds to calculate network centralization as a measure of media fragmentation. His working assumption is that "a high score indicates a high level of inequality in the in-degree

scores […] of the nodes in a network, whereas a low score signifies greater equality" (p. 246). With a centralization score of 17.08% he concludes that his network shows "a tendency for audiences to duplicate relatively evenly across the outlets in the media network" (Ibid). However, the problem is that this network contains many ties that are unlikely to be significant from a statistical point of view.

These limitations have not precluded the application of the method in other published articles, which have uncritically adopted the approach. In Webster & Ksiazek (2012), for instance, the authors write: "As there will always be some level of audience duplication just 'by chance', we wanted a conservative standard. Our approach was to compare the observed duplication between two outlets to the 'expected duplication' due to chance alone. Expected duplication was determined by multiplying the reach of each outlet. So, for example, if outlet A had a reach of 30% and outlet B a reach of 20%, then 6% of the total audience would be expected to have used each just by chance. If the observed duplication exceeded the expected duplication, a link between two outlets was declared present (1); if not, it was absent (0) (see Ksiazek, 2011, for a detailed treatment of this operationalization)" (p. 46). The authors analyze the same data presented in the 2011 paper, but here they reach the conclusion that the network centralization score is 0.86%. On the basis of this score, the authors conclude: "this suggests a high level of equality in degree scores and thus evidence that the audience of any given outlet, popular or not, will overlap with other outlets at a similar level" (p. 49). However, having dichotomized the ties, it is impossible to determine how different the overlap between pairs of media is in terms of actual magnitude. It would have been far more informative to preserve the difference between observed and expected overlap as edge weights, that is, as an attribute of the ties capturing their actual strength. Plotting the distribution of the tie strengths would certainly have been more insightful than plotting the distribution of degree scores (figure 5 in their article). Since the method presented in the 2011 article and re-used in the 2012 article does not allow capturing the actual departure from random overlap, by construction most outlets will be equally connected to most other outlets in the network analyzed.

The more recent study analyzing the top 1,000 global Web domains (Taneja & Webster, 2016) applies, yet again, the same procedure to build the audience overlap network. The authors state: "popular domains are likely to have audiences in common due to random chance.

Therefore, in order to declare a tie, we have set the bar above the level of duplication expected by chance alone. Ksiasek (2011) describes this procedure" (pp. 168-169). As in the previous two papers, they dichotomize the ties (i.e. eliminate their weight) and then they calculate network centralization. The same method is applied in yet another related paper: "since any two websites are expected to have some amount of audience overlap, we considered a tie to exist only when the duplication was above what one would expect by random chance. For instance, if in a given time period a certain website A has a unique user reach of 10% of all Internet users and website B has a reach of 50%, then assuming the consumption of both are independent events, 5% would be the expected number of users to visit both A and B" (Taneja & Wu, 2014, p. 303). Once more, there is no way to determine how many users are comprised by that 5% or whether this difference is indeed a significant departure from 0.

So, summing up, we argue that the methodology applied in prior work has clear limitations to accomplish the stated goals of the research. In spite of these flaws, the methodology has started to percolate in the literature, as the references above show (see also Weeks, Ksiazek, & Holbert, 2016). More recent research, which partly builds on these previous studies, offers a test of statistical significance by applying a chi-squared test to every dyad in the network (Fletcher & Nielsen, 2017, p. 486); however, the audience overlap metrics used in this study derive from a survey and are therefore affected by the recall problems already discussed (as the authors themselves acknowledge, p. 484); in addition, the strength of the overlap is not considered in the analyses (i.e. ties are unweighted), and the networks consist only of the 14 most popular news outlets, as defined by the survey. Our approach preserves the test of statistical significance while overcoming the other limitations. The analysis of audience networks have so far disregarded important developments in network science that are not even recent enough to justify their omission, for instance, developments to analyze weighted networks. Communication research lags, in this respect, far behind other fields (e.g., Barabási, 2016; Borgatti, Everett, & Johnson, 2013; Easley & Kleinberg, 2010; Newman, 2010; Watts, 2003) even though serious attempts have been made to introduce advanced network techniques in the research agenda (Monge & Contractor, 2003). The following section describes in more detail how we improve prior work by capitalizing on some of those methodological developments.

Our Approach

To determine the strength of audience overlap between two media outlets, we apply a technique that relies on the conventional *phi* coefficient, a correlation measure that is related to the chi-squared statistic. Instead of applying it to a contingency table, we apply the coefficient to the matrix representing the number of common visitors to two media outlets. If D_{ij} is the absolute audience duplication (measured as counts, not percentages) between outlets i and j; A_i and A_j are the total audience reach of each source (again, measured as counts); and N is the total sample population (not 100% but the actual total audience size), then the coefficient is expressed as:

$$t\phi_{ij} = \frac{D_{ij}N - A_iA_j}{\sqrt{A_iA_j(N-A_i)(N-A_j)}}.$$

$$\phi_{ij} = \frac{D_{ij}N - A_iA_j}{\sqrt{A_iA_j(N-A_i)(N-A_j)}}.$$

This coefficient is positive when the overlap is larger than expected by chance. Media measurement companies like comScore – which provided the data for our analyses – or Nielsen – used in past research – do not actually give the number of panelists accessing a given outlet; instead, they offer estimates that relate to the entire online population. ComScore measures reach as the total number of unique users that access a given site on a monthly basis. Unlike most previous work, we use these counts, not the percentages, to take into account the fact that media outlets vary drastically in their absolute reach. The estimates provided by comScore or Nielsen are affected by sampling error; the weights applied to make the panels representative of the entire online population; and by random browsing behavior, so it is important not only to determine how much overlap there is but also whether this overlap is beyond a margin of error as determined by usual probability procedures.

To determine the statistical significance of the *phi* correlations, therefore, we make use of the *t* statistic, which is also a conventional tool to determine how likely it would be to observe the measured overlap if the null hypothesis of no overlap were true. The *t* value is calculated with the formula:

$$t_{ij} = \frac{\phi_{ij}\sqrt{N-2}}{\sqrt{1-\phi_{ij}^2}}.$$

For a more stringent test, we use max (A_i, A_j) instead of N, following the same criterion followed in other published work that also filters weighted networks (Ronen et al., 2014). For a significance level $p < 0.05$, t values need to be $t_{ij} > 1.96$, or $t_{ij} > 2.58$ for a probability $p < 0.01$. In our approach, overlapping ties with t values below the probability threshold $p < 0.01$ are eliminated as non-significant.

We are only aware of one other study applying this filtering technique to the analysis of audience overlap networks (Majó-Vázquez, Cardenal, & González-Bailón, 2017). It is also worth noting that there is an alternative technique to filter weighted ties on the basis of a null model, the so-called backbone extraction (Serrano, Boguñá, & Vespignani, 2009). Unlike the *phi* correlation approach, which operates with dyads, the backbone extraction technique defines the null model on the level of ego-networks. A full comparison of these two approaches is beyond the scope of this paper, and we leave that comparison to future research. Here, we focus on the *phi* correlation because it offers a more direct comparison with the findings reported in prior published research, which also defined statistical significance (albeit with the flaws already discussed) on the dyadic level. As the findings reported below show, the networks that result from this statistical test look very different from the networks that do not eliminate insignificant ties.

Data

We use two datasets to illustrate why a statistical test of significant overlap is necessary before audience networks can be analyzed reliably. The data was obtained from comScore, and it contains similar information to the data analyzed in the literature reviewed above: for every media outlet under analysis we have information of their total reach and total overlap with other outlets. We pay attention to news media sources (legacy and digital-born) in the UK during the three months surrounding the 2016 Brexit Referendum (May, June, July) and in the US during the 2016 Presidential Election (October, November, December). Consistent with previous literature, we define legacy media as news sources that predate the advent of the internet; we define digital-born media, on the other hand, as outlets that operate online only and were created after internet technologies became mainstream. This last category includes portals like Yahoo

News and hybrid news sources like Buzzfeed, which combine their own content with syndicated news from external sites. We look into audience consumption during two major political events because this is a critical period for news consumption. We averaged activity over the three months to smooth out random fluctuations (previous research only provides data for one-month snapshots).

Our list of media sites includes those classified by comScore under the category of "information/news" with at least a percentage reach of 0.01% (previous published work report using a 3% reach threshold). The reason why we only consider media sites that have this minimum reach is that domains that receive visits from less than 0.01% of the online population do not produce very reliable estimates, especially for audience overlap metrics (the comScore panels we use have sizes N ~ 67,000 for the UK and N ~ 210,000 for the US). Finally, we manually checked the lists of news sites to categorize them as legacy or digital-born news outlets. In total, the UK dataset contains 133 sites (103 legacy media, 30 digital-born); the US dataset contains 332 sites (253 legacy media, 79 digital-born).

As an additional test of data quality, we compared the comScore reach rankings with those provided by Alexa, another online measurement company. Alexa estimates are based on a global panel of internet users, whereas comScore yields estimates based on panels that are representative of the online population of each country. In spite of their different methodologies, the rank position of news outlets according to reach is very similar in the two panels: the correlation coefficients are 0.71 for the UK data and 0.78 for the US data. Figure 2 shows the scatterplots of this association using the logarithmic transformation of reach (which is highly skewed).

-- Figure 2 about here –

Findings

Network Structure Before and After the Significance Test

The methodology we propose here eliminates overlapping ties that do not reach the significance level for a probability value $p < 0.01$. Figure 3 shows the total number of ties (edges) in the two

audience networks before and after this filtering is applied. In total, 796 ties are eliminated in the UK audience network (12% of all original ties), and 7,456 ties are eliminated in the US network (14% of all the ties in the original network). The elimination of these ties has a clear impact on centralization scores, which is the statistic that prior research used to assess the levels of fragmentation in media consumption. Centralization is a network-level measure that summarizes the centrality scores of the nodes in the network, that is, it indexes "the degree to which the centrality of the most central point exceeds the centrality of all other points" (Freeman, 1979, p. 227). In the networks we analyze here, the points are media outlets and their centrality measures the number of other outlets they share audience with (e.g. their degree centrality). More formally, the centralization of audience networks can be defined as:

$$C_A = \frac{\sum_{i=1}^{g}[C_A(n^*) - C_A(n_i)]}{max \sum_{i=1}^{g}[C_A(n^*) - C_A(n_i)]}$$

The numerator is the sum of the differences between the largest centrality score in the network ($C_A(n^*)$) and all other observed values ($C_A(n_i)$); the denominator considers all possible networks with the same observed size $N = g$ and determines how large the sum of the differences can be. The resulting index is always between 0 and 1: when it equals 0, all actors have exactly the same centrality score; when it equals 1, one actor dominates the other actors (Wasserman & Faust, 1994, pp. 176-177). As Figure 3 shows, in both networks the centralization score increases by more than 10 percentage points when insignificant ties are eliminated.

-- Figure 3 about here –

Prior research has also used the degree distribution to characterize the structure of audience networks and as evidence of "massively overlapping" media consumption patterns (Webster, 2014). However, as discussed, the methodology followed in that prior work disregards the actual strength of the overlap and assumes that all ties are equivalent -- which proves to be an overly simplifying (and misleading) assumption. Figure 4 shows the degree distribution of the US and UK media networks (filtered according to the test of statistical significance). Degree has been normalized so that all scores fall in the range [0, 1]. Again, this measure of centrality is simply a count of other outlets a given media site shares audience with, disregarding the strength

of the overlap. As the histograms show, most sites have a very high degree centrality, which means they share audience with most other outlets. This is consistent with one of the main findings of prior research and the "massive overlap" argument. However, the distribution of centrality scores changes drastically if we take the strength of the overlap (i.e. the tie weights) into account.

The histograms in the second and third columns of Figure 4 show how different the results are if we take into account tie weights. Here we calculate centrality using the eigenvector scores, which correspond to the values of the first eigenvector of the graph's adjacency matrix (Bonacich, 1987). The intuitive interpretation of this measure is that the centrality of a node is not only defined by how many other nodes it is connected to but also, and crucially, by how central adjacent nodes are themselves. In other words, the centrality of a media outlet is proportional to the sum of the centralities of the other outlets it connects to (via audience overlap). If the cells in the adjacency matrix are binary values indicating the mere existence or absence of a tie (as prior research assumes), the eigenvector centrality distribution looks very similar to its degree counterpart: most sites are connected to most other sites. However, if the cells of the adjacency matrix encode the strength of the overlap, as we propose here, then the resulting centrality distribution looks drastically different: now, a small minority of sites have a significantly higher centrality score than the vast majority. The reason is that by using the strength of the overlap as a tie weight, eigenvector centrality gives high scores to outlets that either share audience with many other outlets or share large fractions of their audience with a few others (Newman, 2004). Since news outlets vary drastically in their audience reach, the strength of their overlapping ties vary drastically as well.

-- Figure 4 about here –

Taking the strength of the overlap into account reveals, in other words, that the "massively overlapping culture" is in fact not so massive but rather narrowly concentrated. The top five sites in the UK network as revealed by weighted eigenvector centrality are *bbc.co.uk*, *dailymail.co.uk*, *telegraph.co.uk*, *theguardian.com*, and *mirror.co.uk* (which also happen to be the top 5 sites in terms of total reach). By contrast, the top 5 sites according to the unweighted version of the network (which, in line with prior research, assumes that all overlapping ties are

equivalent) are *news.sky.com*, *itv.com/news*, *thesun.co.uk*, *standard.co.uk*, and *express.co.uk* – a completely different ranking. The rankings for the US network reveal a similar discrepancy. According to the weighted measure of centrality, the top 5 sites are *yahoo.com/news*, *cnn.com*, *aol.com*, *nytimes.com*, and *washingtonpost.com* (which are, again, the top news providers in terms of total reach). The top 5 sites according to the unweighted version of centrality includes *freebeacon.com*, *hngn.com*, *mcclatchydc.com*, *bostonherald.com*, and *9news.com* (there are several other sites that share the same unweighted centrality score). What these results suggest, then, is that taking the strength of the overlap into account is necessary to uncover a hierarchy in the network of news consumption that goes unnoticed otherwise – and that, indeed, went unnoticed in prior work.

The Core Structure of Audience Networks

Another advantage of preserving the strength of the overlap is that we can use it as a filtering mechanism to progressively "peel away" the weakest ties – and uncover the core of the media network (an approach that is known in the literature as "thresholding", see Borge-Holthoefer & González-Bailón, 2017). Figure 5 shows the number of sites that remain connected in the largest component of the network (vertical axis) as the weakest ties are progressively removed (horizontal axis; tie weights are normalized to fall in the interval [0, 1]). For comparison, we did the same thresholding exercise with simulated networks ($N = 1,000$) that preserved the observed structure and tie weight distribution but permuted the tie weights by randomly sampling from the empirical weight distribution without replacement (Ripley, 1987). Figure 5 plots the mean values and standard deviations (± 2) of those random permutations. What the figure shows is that removing the weakest overlapping ties (in the normalized interval [0, 0.1]) reduces the observed networks by more than 80%: only 21 sites remain connected in the UK case, and 48 in the US case. By contrast, in the networks that permute the allocation of tie weights, removing the weakest ties does not have such a drastic impact on the connectivity of the network: the size of the largest component is, on average, still 88 for the UK and 292 for the US. Once again, this finding highlights that not all ties are equally relevant to understand media consumption patterns (i.e., their strength is correlated with the total audience reach of news outlets), and that the decision of prior research to binarize the ties eliminates crucial information from the analyses.

-- Figure 5 about here –

The news sites that remain connected after the weakest ties are filtered out form the core of the network – a core that becomes increasingly smaller as the weight threshold becomes more stringent. This core is the elite of news sources that most online users consume. However, there is also a hierarchy within the hierarchy. Figure 6 shows the network of sites that remain connected once the weakest ties (in the normalized interval [0, 0.1]) are removed. In both cases, there is a clear core-periphery structure (Borgatti & Everett, 1999): there is a core is formed by sites that have stronger overlapping ties with each other (labels in italics), and the periphery is formed by sites that share audience with core sites but not with other peripheral sites (regular font). In the US case, the network was so dense (inset) that we only visualize the strongest ties, i.e., ties that connect media sites at the core of the core. One interesting finding is that digital-born outlets like BuzzFeed, which publishes its own content as well as aggregated news, and The Huffington Post and portals like Yahoo News and msn.com, are part of the media core: as Figure 2 showed, some digital-born media outlets perform substantially better in terms of reach than most legacy media, despite their significantly shorter history. This fact provides evidence that digital-born outlets can challenge the monopoly of legacy media brands as the most prominent sources of news online. For the most part, however, legacy media are still the preferred sources of news for online users. The strongest audience overlap takes place across these legacy media sites, which arise as the main points of convergence of audiences searching for news online.

-- Figure 6 about here –

All in all, these analyses reveal that audiences searching for news online are not fragmented in a way consistent with the hypothesis of self-selection or selective exposure: their behavior reveals a structure of overlap that is very dense, where most news outlets are connected with many other news outlets and they are all part of the same component. These overlapping ties, however, are very heterogeneous in their strength, which means that audiences effectively concentrate around the few news outlets that have disproportionate reach. This pattern characterizes both the US and the UK data, but the analyses also reveal a meaningful difference between these two media environments: the UK network is substantially more centralized than the US network. This is consistent with the prominence of British public broadcasting: the BBC has double the reach of the second most accessed outlet for the period we consider (e.g. the

tabloid newspaper Daily Mail). The US network is, by comparison, more homogenous in the distribution of reach and centrality scores, consistent also with the more atomized supply of the media sector in this country.

Discussion

The analysis of audience duplication networks can shed important light on the dynamics of news consumption and how varied news diets are in terms of the sources consumed. Sparse or fragmented networks are an indication of individual-level selective exposure: the absence of ties between media sites results, in the end, from lack of overlap between audiences. On the other hand, dense and highly cohesive networks (which is what we find with our data, in line with prior work) suggest that audiences consume news sources widely. However, merely analyzing the structure of audience networks does not capture the whole story of how media environments vary: the analysis of tie weights (or the magnitude of the overlap, which prior research failed to take into account) is also important to identify clusters within clusters where audience overlap is strongest. Our results suggest that this core is formed predominantly by high-profile legacy media brands, and that consumption patterns at this core show no evidence of fragmentation. This lack of support for the fragmentation hypothesis is also consistent with prior research failing to find strong evidence of segregation online (Flaxman, Goel, & Rao, 2016; Gentzkow & Shapiro, 2011). When consuming digital news, users form highly cohesive networks with their browsing behavior, which means that large segments of the online population obtain their news from a diverse range of media – even if most of the attention is concentrated in a minority of sources at the core.

More generally, our analyses suggest that instead of a "massively overlapping culture" (Webster, 2014), audience networks can be best characterized by a small core of outlets that act as the primary source and then a much larger periphery of secondary sources, distributed across several layers of decreasing reach. Legacy and digital-born media are represented in all parts of the network, but the core is mostly formed by long-standing legacy media outlets, which speaks to the importance of offline brand awareness. This is consistent with well-known asymmetries in online visibility, and the rich-get-richer dynamics that are manifested in the long-tails that are typical of online activity (Hindman, 2009). What our analyses reveal is that news sites in the tail

of the distribution, in terms of reach, share also a substantive fraction of their audiences – which sets those sites clearly apart from more peripheral, secondary sources.

These patterns suggest that selective exposure is not the main driving force underlying choices in digital news consumption. However, our data show limits on selective exposure only at the outlet level, that is, across news sites; our findings are much less informative about selective exposure within platforms of news distribution, for example, social media and, in particular, Facebook or Twitter, where filter bubbles might be more prevalent (although there is research that suggests otherwise, see Barberá, Jost, Nagler, Tucker, & Bonneau, 2015).

To put in context this distinction between self-selection across news sites and within platforms of news distribution, figure 7 shows data on the main entry points to the outlets in our audience networks. According to Alexa statistics, search engines and direct navigation are the main sources of traffic to the news sites we analyze: referrals from social media or from links on other websites account for only about a quarter of all referrals to the sites in our networks. Similar findings suggesting that users go directly to their preferred news sources have been reported in prior work (Allcott & Gentzkow, 2017; Flaxman et al., 2016). This means that there is high awareness of news media brands and that users going online make a resolute effort to get news from different sources – this is why we find weak evidence of self-selection across news outlets. Of course, these users, who are highly engaged in the consumption of political information, are not necessarily representative of the overall online population. Less attentive or less politically engaged users might get most of their news indirectly through social media platforms and their feeds, where the imprint of self-selection might be stronger. There is, as of today, little comparative data that allow us to assess how prevalent filter bubbles are across news sources and within social media platforms in different media environments. Our analyses offer one step in that direction.

To add some more contextual information to this discussion, the news source with maximum reach in the UK is the BBC site, with about 43% of the online population accessing its content; the most popular site in the US is CNN, with a reach of about 57% of the online population. If we compare these numbers with the reach of social media (Facebook has a reach of 60% in the UK, and 88% in the US; Twitter has a reach of 22% and 46%, respectively; all

percentages according to comScore data), it is clear that there is a fraction of the online population (around 20-30%) that are likely to obtain their online news solely from social media, in particular Facebook. And considering that social media referrals to the websites of news outlets amounts to just about 20% of their total traffic, it also probably means that social media users consume most of their news directly from their feeds (which have been shown to be biased by self-selection and filter-bubble dynamics, see Bakshy, Messing, & Adamic, 2015). With the data we analyze here, however, we can only speculate about these dynamics: social media activity goes undetected unless users decide to click on links that will take them to a news website – i.e., the activity tracked by comScore panels. Individual-level data on engagement with news content within social media platforms would be necessary to properly substantiate claims on the impact that passive exposure has on individuals who are, otherwise, not greatly interested in searching for political news. Our analyses, however, help contextualize the role that social media play as entry points to news within the larger picture of digital news consumption – which, again, is characterized by broad browsing patterns (at least, amongst those who are engaged with news sources).

There are some important limitations with audience data as collected by measurement companies like comScore, Nielsen or Alexa. One limitation is that the data provided are aggregated monthly, thus making it impossible to differentiate individuals who access several sources regularly (e.g. daily) as opposed to just occasionally (e.g. once a month). Another important limitation is that individual-level data browsing behavior is not available to reconstruct the consumer projection in Figure 1, panel A. One of the reasons is that the panels are weighted to make audience estimates representative of the overall online population. The analysis of individual-level data would allow a more fine-grained understanding of news consumption patterns, as recent research using the browsing behavior of individuals illustrates (Athey, Mobius, & Pal, 2017; Flaxman et al., 2016). However, these studies are affected by sampling limitations; for instance, they rely on specific web technologies (e.g. Microsoft products) that are not necessarily representative of the online population at large. Companies like comScore or Nielsen provide panel data that can be generalized – although their methodology is proprietary and, therefore, not entirely transparent. One more limitation is that most studies on audience behavior do not provide data that captures browsing activity below the domain level, so it is difficult to parse out the actual content to which users are exposed when they access a news site.

Finally, the time that different individuals spend in each news site can vary drastically, from seconds to several minutes. These differences are of course also meaningful to understand online audience behavior and should be considered in future studies – for now, they are also masked under the aggregated monthly measures, which is the lowest temporal resolution that is available for research.

In spite of these limitations, the analysis of digital traces based on observed browsing behavior as opposed to self-reported information is providing new ways to measure the elusive construct of 'news exposure' (de Vreese & Neijens, 2016; Prior, 2009a, 2009b, 2013). The consumption of political information has been traditionally measured through surveys asking respondents to recall their exposure to news – a data collection strategy that, as already mentioned in the introduction, is known to be weakened by problems with reliability and validity (Dilliplane et al., 2013; Goldman et al., 2013; Prior, 2013). The analysis of digital traces, in the form of browsing behavior or engagement with social media content, promises to offer a more accurate representation of how people consume news and how those patterns change over time and across media environments.

A challenge that future research needs to address is how to integrate mobile and social media data, especially given existing constraints on access because of the proprietary nature of the data and related privacy concerns. The promise is that the analysis of these behavioral traces can translate into novel metrics that will facilitate comparative research across different media environments, regardless of the specificities of their political contexts. As the analyses discussed above show, the properties of online browsing behavior can be quantified with indices and scores that can help rank specific outlets but also the full media environment for a given country. For instance, our analyses show that the UK media network is substantially more centralized than the US network, mostly because of the prominence of the British public service media. In future research we plan on expanding the comparison of audience networks to other countries and to analyze changes over time. In this article we focus on an observation period during which very salient political events took place (i.e. the Brexit referendum, a Presidential Election). It is a pending empirical question whether the networks of audience overlap change substantially at different points of the political cycle.

The line of research we advocate for here can also help contextualize the social impact of fake news and misinformation, especially to the extent that it can identify those individuals who are less actively engaged with news sources and are, therefore, more likely to be affected by passive exposure through social media. This would require the analyses of data tracking user engagement with the news in platforms like Facebook and Twitter, but the same methodology could be applied as long as users can be affiliated with specific news outlets using measures of engagement with their content. Another way of complementing the data we analyze is to look at the content of the news outlets to assess the actual diversity of the information they provide – we assume that diversity derives from exposure to different news outlets but that, of course, is an important assumption that would require further empirical evaluation.

Conclusion

Audience duplication data offer important information to characterize the media landscape and analyze how people navigate the news environment. In the context of digital news, audience networks help identify the media outlets that act as gravity centers of public attention – thus offering a map of how the media environment shapes news consumption choices. Here, we proposed a series of methodological improvements to how audience overlap networks were analyzed in the past. First, we showed that the structure of audience networks changes substantially when insignificant ties are removed (prior research did not apply a test of significance). Second, we showed that the strength of the overlap (also disregarded in previous research) is crucial to uncover the core-periphery structure of audience networks. The two networks we analyze exhibit a very cohesive core, with no evidence of fragmentation. This core contains a few digital-born news sources but it is fundamentally formed by legacy brands, which still stand (by far) as the main sources of news online. These results have implications for our theoretical understanding of how digital technologies mediate engagement with political information and, by extension, with the democratic process. We show that audiences that are interested enough in political information to access news sites display no evidence of selective exposure or self-selection.

References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Working Paper, Stanford University*.

Anderson, C. (2006). *The Long Tail. How Endless Choice is Creating Unlimited Demand*. London: Random House.

Athey, S., Mobius, M., & Pal, J. (2017). The Impact of Aggregators on Internet News Consumption. *Stanford University Graduate School of Business Research Paper No. 17-8*.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130-1132. doi:10.1126/science.aaa1160

Barabási, A.-L. (2016). *Network Science*. Cambridge: Cambridge University Press.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science, 26*(10), 1531-1542. doi:10.1177/0956797615594620

Bonacich, P. (1987). Power and Centrality: A family of Measures. *American Journal of Sociology, 92*(5), 1170-1182.

Borgatti, S. P., & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks, 21*(4), 375-395.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing Social Networks*. London: Sage.

Borge-Holthoefer, J., & González-Bailón, S. (2017). Scale, Time and Activity Patterns: Advanced Methods for the Analysis of Online Networks. In N. Fielding, R. Lee, & G. Blank (Eds.), *Handbook of Online Research Methods* (2nd ed.). London: Sage.

Cooper, R. (1996). The status and future of audience duplication research: An assessment of ratings-based theories of audience behavior. *Journal of Broadcasting & Electronic Media, 40*(1), 96-111. doi:10.1080/08838159609364335

de Vreese, C. H., & Neijens, P. (2016). Measuring Media Exposure in a Changing Communications Environment. *Communication Methods and Measures, 10*(2-3), 69-80. doi:10.1080/19312458.2016.1150441

Delli Carpini, M. X., & Keeter, S. (1996). *What Americans Know about Politics and Why it Matters*. New Haven: Yale University Press.

Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). Televised Exposure to Politics: New Measures for a Fragmented Media Environment. *American Journal of Political Science, 57*(1), 236-248. doi:10.1111/j.1540-5907.2012.00600.x

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York, NY: Cambridge University Press.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly, 80*(S1), 298-320. doi:10.1093/poq/nfw006

Fletcher, R., & Nielsen, R. K. (2017). Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication. *Journal of Communication, 67*(4), 476-498. doi:10.1111/jcom.12315

Freeman, L. C. (1979). Centrality in Social Networks: Conceptual clarification. *Social Networks, 2*(3), 215-239.

Gentzkow, M., & Shapiro, J. M. (2011). Ideological Segregation Online and Offline. *The Quarterly Journal of Economics, 126*, 1799-1839.

Goldman, S. K., Mutz, D. C., & Dilliplane, S. (2013). All Virtue Is Relative: A Response to Prior. *Political Communication, 30*(4), 635-653. doi:10.1080/10584609.2013.819540

Goodhardt, G. J., & Ehrenberg, A. S. C. (1969). Duplication of Television Viewing between and within Channels. *Journal of Marketing Research, 6*(2), 169-178. doi:10.2307/3149668

Goodhardt, G. J., Ehrenberg, A. S. C., & Collins, M. A. (1975). *The Television Audience: Patterns of Viewing*: Saxon House.

Hindman, M. S. (2009). *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.

Katz, E. (1996). And Deliver Us from Segmentation. *The ANNALS of the American Academy of Political and Social Science, 546*, 22-33.

Ksiazek, T. B. (2011). A Network Analytic Approach to Understanding Cross-Platform Audience Behavior. *Journal of Media Economics, 24*(4), 237-251. doi:10.1080/08997764.2011.626985

Majó-Vázquez, S., Cardenal, A. S., & González-Bailón, S. (2017). Digital News Consumption and Copyright Intervention: Evidence from Spain before and after the 2015 "Link Tax". *Journal of Computer-Mediated Communication, 22*(5), 284–301. doi:10.1111/jcc4.12196

Monge, P. R., & Contractor, N. S. (2003). *Theories of Communication Networks*. Oxford: Oxford University Press.

Neuman, W. R. (2016). *The Digital Difference. Media Technology and the Theory of Communication Effects*. Cambridge, MA: Harvard University Press.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E, 70*(5), 056131.

Newman, M. E. J. (2010). *Networks. An Introduction*. Oxford: Oxford University Press.

Prior, M. (2007). *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. Cambridge: Cambridge University Press.

Prior, M. (2009a). The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure. *Public Opinion Quarterly, 73*(1), 130-143. doi:10.1093/poq/nfp002

Prior, M. (2009b). Improving Media Effects Research through Better Measurement of News Exposure. *The Journal of Politics, 71*(03), 893-908. doi:doi:10.1017/S0022381609090781

Prior, M. (2013). The Challenge of Measuring Media Exposure: Reply to Dilliplane, Goldman, and Mutz. *Political Communication, 30*(4), 620-634. doi:10.1080/10584609.2013.819539

Ripley, B. D. (1987). *Stochastic simulation*. New York: Wiley.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences, 111*(52), E5616-E5622. doi:10.1073/pnas.1410931111

Serrano, M. Á., Boguñá, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences, 106*(16), 6483-6488. doi:10.1073/pnas.0808904106

Stroud, N. J. (2011). *Niche News: The Politics of News Choice*. New York: Oxford University Press.

Sunstein, C. R. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.

Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton, NJ: Princeton University Press.

Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ: Princeton University Press.

Taneja, H. (2016). Mapping an audience-centric World Wide Web: A departure from hyperlink analysis. *New Media & Society*. doi:10.1177/1461444816642172

Taneja, H., & Webster, J. G. (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. doi:10.1111/jcom.12200

Taneja, H., & Wu, A. X. (2014). Does the Great Firewall Really Isolate the Chinese? Integrating Access Blockage With Cultural Factors to Explain Web User Behavior. *The Information Society, 30*(5), 297-309. doi:10.1080/01972243.2014.944728

Turow, J. (2012). *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*: Yale University Press.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Watts, D. J. (2003). *Six Degrees. The Science of a Connected Age*. London: William Heinemann.

Webster, J. G. (1985). Program audience duplication: A study of television inheritance effects. *Journal of Broadcasting & Electronic Media, 29*(2), 121-133. doi:10.1080/08838158509386571

Webster, J. G. (2014). *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. Cambridge, MA: MIT Press.

Webster, J. G., & Ksiazek, T. B. (2012). The Dynamics of Audience Fragmentation: Public Attention in an Age of Digital Media. *Journal of Communication, 62*(1), 39-56. doi:10.1111/j.1460-2466.2011.01616.x

Weeks, B. E., Ksiazek, T. B., & Holbert, R. L. (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. doi:10.1080/08838151.2016.1164170

Williams, B. A., & Delli Carpini, M. X. (2011). *After broadcast news: Media regimes, democracy, and the new information environment*: Cambridge University Press.

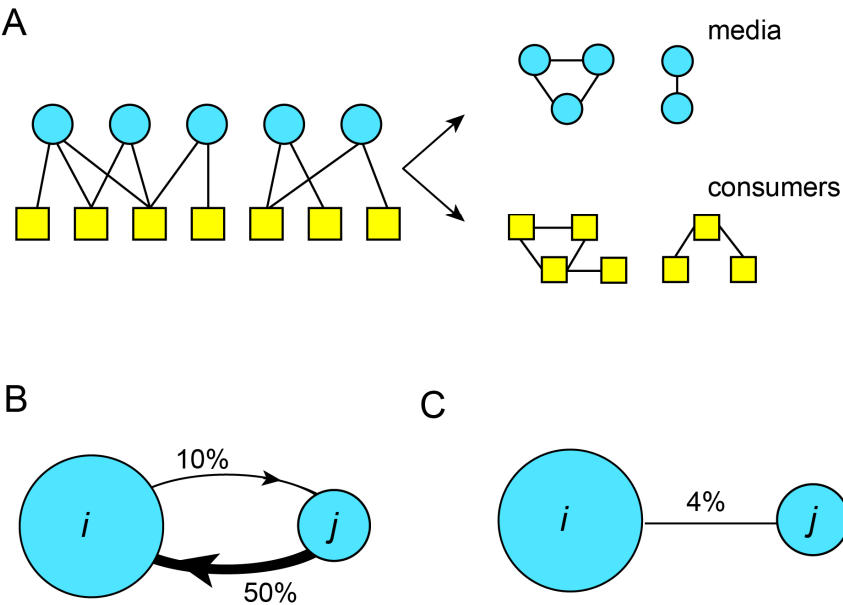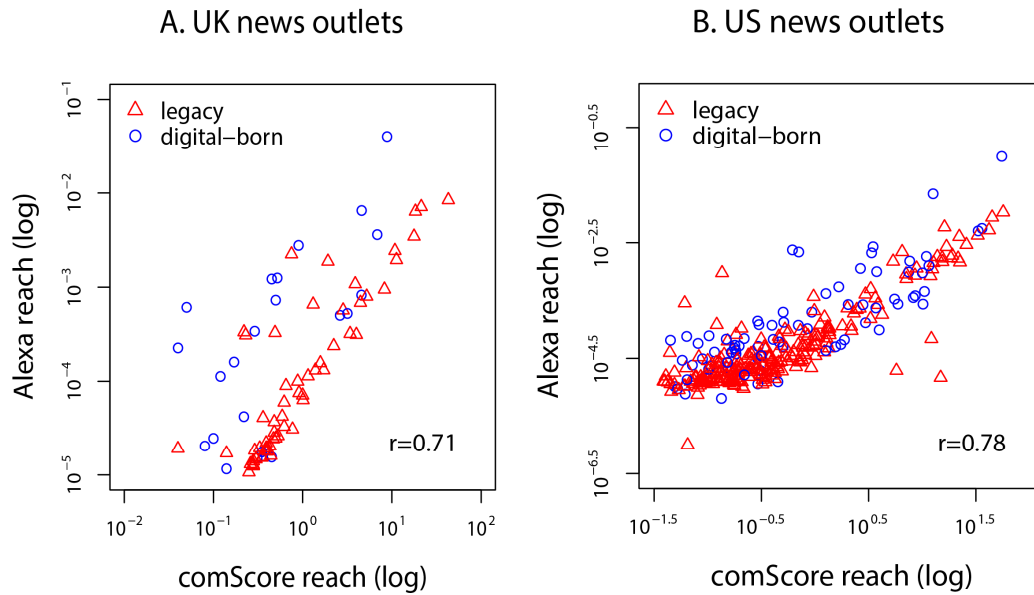Figure 1. The Construction of Audience Overlap Networks

Figure 2. Audience Reach of News Sites according to comScore and Alexa Panels

Note: the UK data are averaged over the months May, June, and July of 2016; the US

data are averaged over the months October, November, and December of 2016.

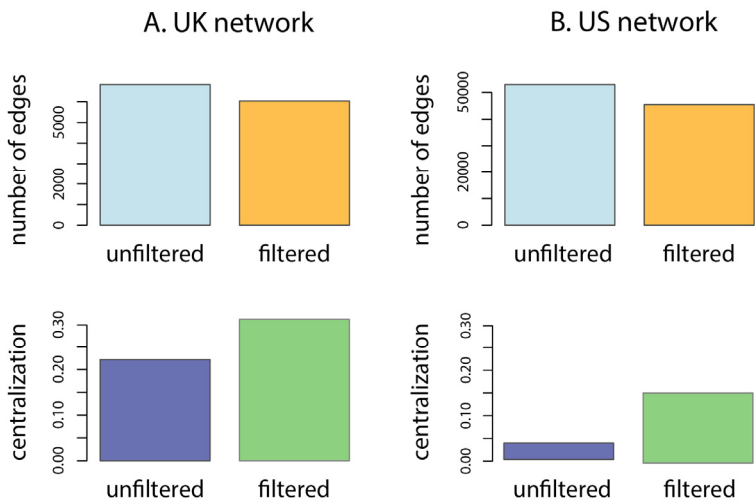Figure 3. Audience Networks Before and After Significance Test

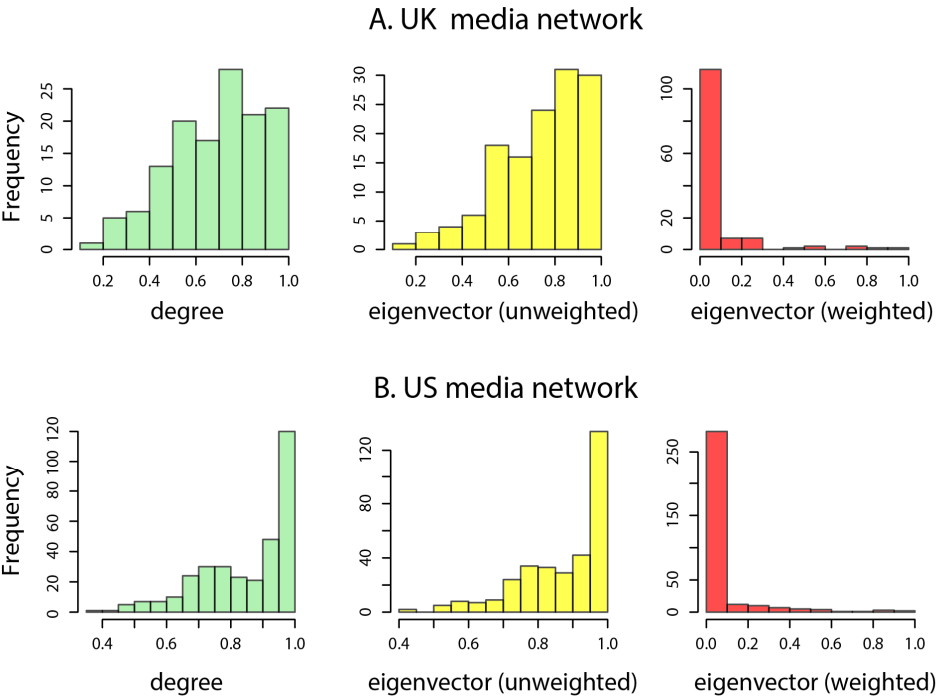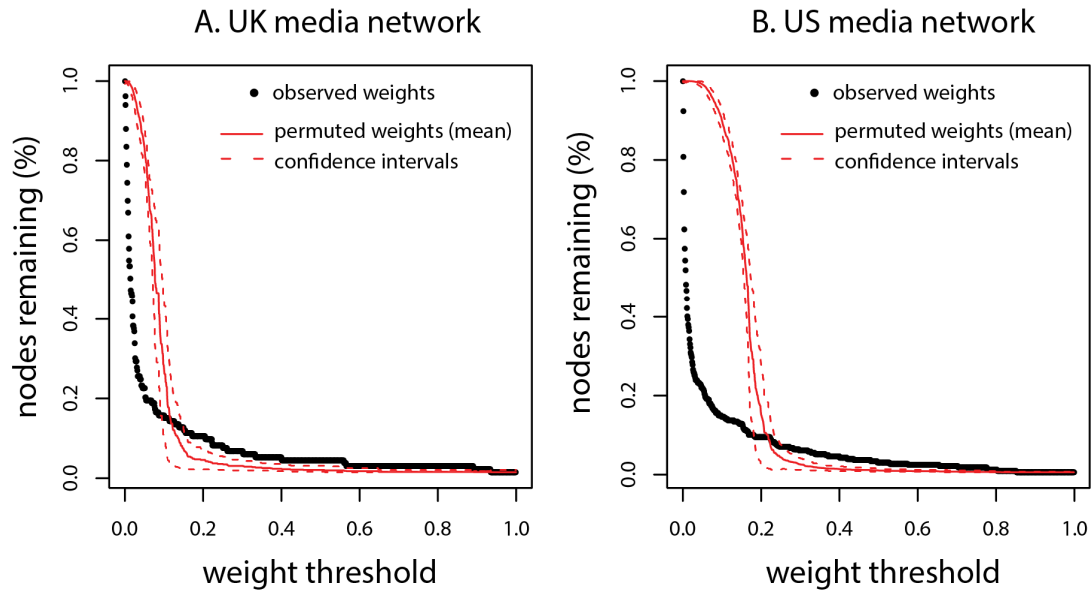Figure 4. Centrality Distributions in Networks with Unweighted and Weighted Ties

Figure 5. The Elite Structure of Audience Networks

Note: the plots track the percentage of media sites that remain connected to the largest component (vertical axis) as the weakest ties are progressively removed from the network (horizontal axis). The range of weight values has been normalized to fall in the interval [0, 1]. The red lines offer a benchmark for comparison based on 1,000 random permutations of tie weights (confidence intervals based on ± 2 standard deviations).

Figure 6. Core of the Two Media Networks where Audience Overlap is Stronger

A. UK core media network
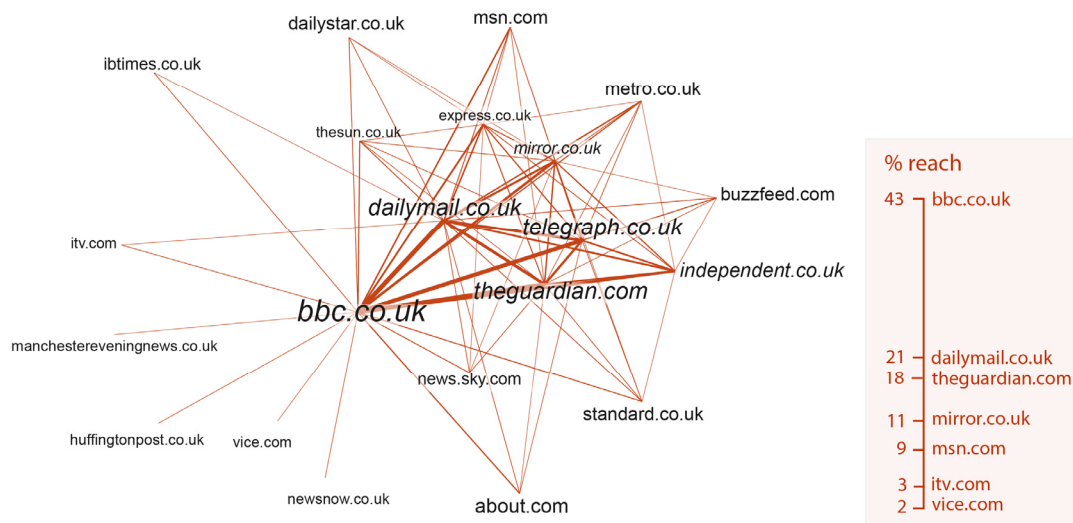
B. US core media network
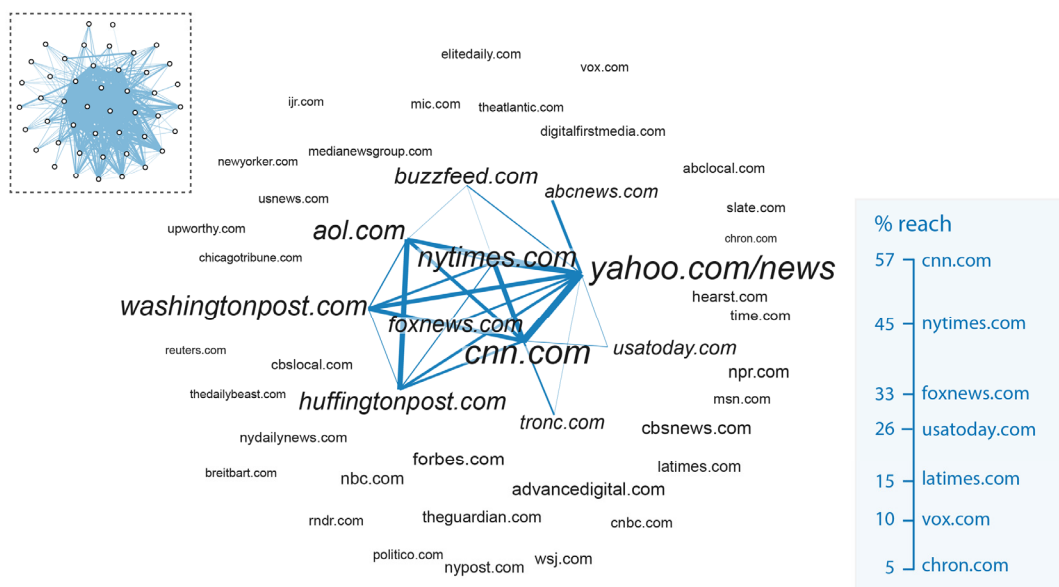
Note: nodes are news media sites and ties map audience overlap. The size of the labels is proportional to the total audience reach of each media outlet (scale and range are detailed in the legends; the scaling has been selected to enhance label legibility). The thickness of the ties is proportional to the audience news sites share.

Figure 7. Traffic Sources for the Media Sites in the Audience Overlap Networks
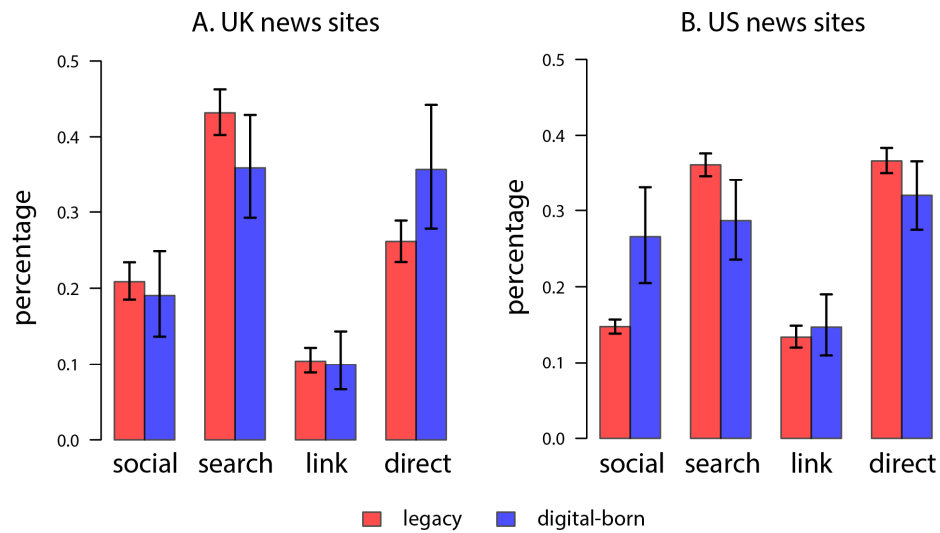
Note: percentages are based on Alexa data; the 95% confidence intervals are bootstrapped.

Search and direct traffic are the main entryways to news media sites in both countries. Referrals

from social media are significantly higher for digital-born outlets in the US, vis-à-vis digital

media, but social media still ranks third, after search and direct navigation.