What Counts as a Weak Tie?

A Comparison of Filtering Techniques to Analyze Co-Exposure Networks

Subhayan Mukerjee[1], Tian Yang[1], Georg Stadler[2], and Sandra González-Bailón[1,]*

[1] University of Pennsylvania, [2] New York University

Abstract: Co-exposure networks offer a useful tool for analyzing audience behavior. In these networks, nodes are sources of information and ties measure the strength of audience overlap. Past research has used this method to analyze exposure to content on social media and the web. However, we still lack a systematic assessment of how different choices in the construction of these networks impact the results. Here we evaluate three different filtering rules that have been used in the literature to eliminate noise in raw data and identify the strongest connections (i.e., those above a certain weight). Moreover, we also provide a mathematical heuristic to choose the optimal threshold. To illustrate our approach, we use two observed networks measuring co-exposure to news sources on the web. We then formulate the problem of filtering the networks as a trade-off between network sparsity (i.e., the need to remove the weakest ties) and connectedness (i.e., the need to preserve the observed connectivity). Our mathematical approach resolves this problem by finding the threshold that maximizes the number of edges removed while minimizing the number of nodes becoming isolates. This analytical technique is generalizable and can be applied to the analysis of any weighted structure that requires solving a similar trade-off between network measures.

Keywords: weighted graphs; co-exposure networks; thresholding; L-curve method.

* 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

Acknowledgements: work on this paper was funded by NSF grants 1729412 and 2017655.

What Counts as a Weak Tie?

A Comparison of Filtering Techniques to Analyze Co-Exposure Networks

Abstract: Co-exposure networks offer a useful tool for analyzing audience behavior. In these networks, nodes are sources of information and ties measure the strength of audience overlap. Past research has used this method to analyze exposure to content on social media and the web. However, we still lack a systematic assessment of how different choices in the construction of these networks impact the results. Here we evaluate three different filtering rules that have been used in the literature to eliminate noise in raw data and identify the strongest connections (i.e., those above a certain weight). Moreover, we also provide a mathematical heuristic to choose the optimal threshold. To illustrate our approach, we use two observed networks measuring co-exposure to news sources on the web. We then formulate the problem of filtering the networks as a trade-off between network sparsity (i.e., the need to remove the weakest ties) and connectedness (i.e., the need to preserve the observed connectivity). Our mathematical approach resolves this problem by finding the threshold that maximizes the number of edges removed while minimizing the number of nodes becoming isolates. This analytical technique is generalizable and can be applied to the analysis of any weighted structure that requires solving a similar trade-off between network measures.

Keywords: weighted graphs; co-exposure networks; thresholding; L-curve method.

Networks offer a useful tool to analyze co-exposure to content. Measures of audience overlap, for instance, allow us to draw a picture of how much co-exposure exists between information sources, which in turn allows us to shed empirical light on questions of selective exposure: if no overlap exists between two sources (i.e., if there are no ties connecting them) it means that those sources attract different types of audiences. This type of networks are weighted structures where the tie weights are proportional to the strength of the overlap – or the number of people co-exposed to the sources. Co-exposure networks usually result from the one-mode projection of bi-partite structures (or affiliation networks) where individuals are linked to sources. As is common in these projections, the networks tend to have many cliques with varying edge weight distributions (Newman 2018). The high density of these projections is one of the reasons why filtering techniques can be a useful step in the analysis: they help identify the parts of the network where the strongest connections exist.

The ubiquity of digital technologies, and the trails these technologies leave behind, have made it much easier to obtain reliable information on how people gain exposure to content and how much co-exposure they get from different sources. However, behavioral data drawn from digital traces can also be very noisy, reflecting measurement error or stochastic data gathering processes (Golder and Macy 2014). Filtering techniques designed to eliminate edges as a function of their weight can help identify the core of activity in the network but also reduce the noise introduced during data collection. However, choosing a specific threshold that identifies the ties to be removed can be a rather subjective exercise. This is the case even when statistical benchmarks are used to determine the significance of tie strength – much in the same way as *p* values used in standard statistical procedures reflect a convention rather than a ground truth. Here, we define the choice of a filtering threshold as a mathematical trade-off between sparsity (i.e., removing the weakest ties) and connectedness (i.e., preserving the observed connectivity of the network). Our goal is to present a solution to the problem of finding the optimal threshold based on an objective criterion, rather than on *ad hoc* choices.

In what follows, we first describe three different filtering techniques employed in prior research to make networks sparser, and we compare their effects on network topology using two different datasets that track web browsing behavior in two different countries (the US and the UK). We want to illustrate the impact that choosing a particular threshold has on the conclusions we can draw from observed networks – but our contribution is essentially methodological and intended to generalize to other data sets and substantive areas. We then describe the trade-off that exists between removing as many weak ties as possible (i.e., maximizing sparsity) and preserving the connectedness of the network (e.g., minimizing the number of nodes that become isolates). The formation of overlapping ties is the key building block of co-exposure networks – it is what determines other structural properties, offering a map of how co-exposure is distributed among nodes (in our case, news outlets). Our trade-off is motivated by the need to preserve as much information as possible about these co-exposure patterns, retaining only the most relevant ties; but this presents the technical problem of defining "relevant" to accomplish such a task. Our goal here is to propose a solution to this technical question.

To study the trade-off, we use an adaptation of the L-curve method employed in parameter estimation (Vogel, 2002). We propose a technique to identify the threshold that best solves the tension between sparsity and connectedness, and we compare the outputs for the two observed networks and for two randomized versions of the data. Our datasets map co-exposure to news sources on the web, but our analyses can be applied to any type of weighted structure – especially if the structure is very dense or clustered and its analysis would benefit from eliminating the statistically insignificant connections. Network ties signal affinity and closeness (i.e., in the context of co-exposure networks, closeness is measured as overlap in the audience base), and the overall connectedness of the network is indicative of how much affinity exists between the nodes (i.e., in our context, information sources). The trade-off exists because eliminating the weakest ties disrupts network connectivity – potentially distorting significant features of the empirical data.

Applying a filtering threshold might, for instance, isolate nodes that would be otherwise connected to the rest of the network, preventing us from analyzing various instances of (in our example) legitimate co-exposure. On the other hand, retaining all ties can obfuscate the signal of significant overlap, especially if the network is extremely dense. Our approach aims to maximize the number of ties eliminated for being "noisy" or not adding meaningful signal to the data, while minimizing the disruption of the overall connectivity levels. Resolving this trade-off enables us to choose a less subjective standard for eliminating weak ties. To that end, our study helps define the "weakness" of ties by formulating an optimization problem and providing a mathematical heuristic that identifies the optimal threshold.

1. Prior Work

Bi-partite networks are formed by two types of nodes (e.g., actors and organizations, users and websites) linked through at least one tie across the sets (e.g., donations, affiliations, exposure etc., Newman 2018; Wasserman and Faust 1994). The one-mode projections of the original incidence matrix map the connections between one type of nodes. For instance, in the analysis of co-exposure networks, nodes are usually information sources and the edges encode the number of people co-exposed to those sources. Researchers have used this type of weighted structure to analyze a wide variety of empirical phenomena, ranging from interlocking corporate directorates (Burt 1978), the formation of creative teams (Uzzi and Spiro 2005), job mobility across academic institutions (Fowler, Grofman and Masuoka 2007), and edits across linguistic editions of Wikipedia (Ronen et al. 2014). Co-exposure to news organizations in social media platforms offer another prominent example (e.g., Grinberg et al. 2019; Schmidt et al. 2017). The analysis of these co-exposure networks sheds light on clusters formed by news outlets that are tightly connected through shared audiences.

Across all these examples, the weight of the ties adds valuable information to the clustering in the network by allowing us to identify the nodes that are most strongly connected. For instance, in the Wikipedia network, strong ties exist across language editions that share most bilingual editors; in the social media co-exposure networks, strong ties exist across news sources that share most readers. The weight of a tie is not independent of the nodes it connects: the strongest ties in the Wikipedia network connect the most spoken languages, and the strongest ties in the co-exposure network connect the most popular news sources. However, the strength of the correlation between network topology and weight distribution changes across empirical settings (Yan et al. 2018). In some networks, the most central nodes (i.e., those with the highest number of connections) also maintain the strongest ties; but in some other networks, centrality and tie strength may be less correlated (e.g., the connections with the highest strength might be located at the periphery of the network). The empirical analysis of weighted networks allows us to differentiate these different scenarios and network configurations.

The meaning of tie strength and how we use this attribute in the analysis of an observed network depends on the context of the data. This is because the network representation of empirical phenomena is, in the end, a theoretical exercise (Butts 2009). In the context of co-exposure networks, tie strength is a measure of proximity or similarity between sources, revealed by the extent of the overlap between their audiences. This measurement has been used in recent years by a growing body of work looking at audience behavior. One focus of this work has been the study of audience fragmentation (Majó-Vázquez et al., 2019; Mukerjee et al., 2018; Yang et al., 2020) and selective exposure (Mukerjee, 2021). Another focus has been exposure to fake news and clustering in the consumption of false information (Grinberg et al. 2019), as well as polarization and partisanship in news consumption (Del Vicario et al. 2017; Schmidt et al. 2017; Weaver et al. 2019).

The increasing adoption of these network analytic techniques serves as testament to their effectiveness in uncovering relevant social phenomena. On the other hand, their use has also led to the implementation of various analytical choices that have important downstream consequences for the findings reported. One of those choices relates to how the network is pruned or filtered prior to the analysis. In the context of co-exposure research, there are several substantive reasons why filtering observed networks is meaningful and relevant: one is that it helps identify the news sources that are most similar in their audience base; another is that it eliminates the effects of random browsing behavior resulting from measurement error introduced during data collection.

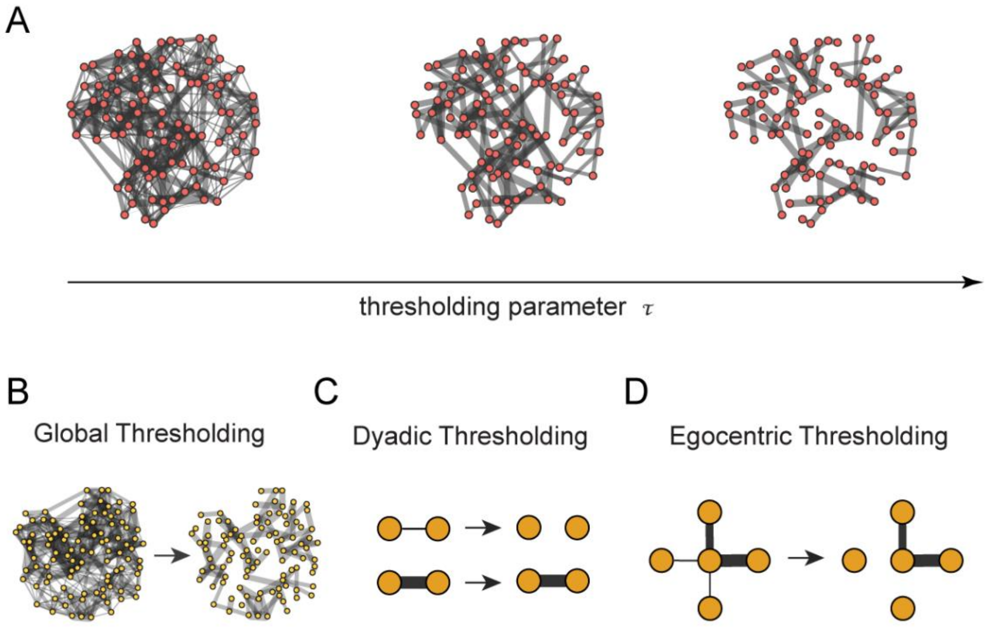
As we explain in the following section, a review of the literature reveals several ways in which researchers have dealt with this pre-processing step in data analysis. In fact, past work has applied different filtering rules to similar network structures, begging the question of how sensitive the reported results are to the choices made in data preparation. The next section discusses three of the approaches most commonly used to transform dense weighted networks into sparser structures. Given their prominence in recent research, we use networks of co-exposure to news as a working example to illustrate how the filtering rules operate, but our approach is intended to serve as a generalizable framework to analyze other weighted structures. Our goal is to assess the impact that these different filtering techniques have on the network topology, and to determine which method offers the most useful filtering mechanism given the trade-off that exists between network sparsity and connectedness.

2. Filtering Techniques

The goal of network filtering is to eliminate the least relevant ties so that the core of the network, or the subgraph containing the most relevant information, can be uncovered. The definition of what counts as "relevant" is, of course, to a large degree subjective. The approaches we discuss here define relevance (and thus "weakness") to varying degrees of statistical rigor. Our goal is to determine which of the operational definitions of "weak tie" is the most suitable for analytical purposes. We define an optimization problem and provide an answer that helps us identify an optimal threshold for "weakness". We should add that tie strength (or weakness) in the context of our approach refers exclusively to the weight of the edges, i.e., in our case, their bandwidth in terms of overlapping audience. This raises the question of whether this operationalization can also be connected to a more conceptual or theoretical understanding of the importance of weak ties (to suggest, for instance, that these ties act as bridges connecting otherwise disconnected clusters, following the classic hypothesis of Granovetter, 1973). This question, however, cannot be answered by looking at edge weight alone; it requires examining the full network structure, which can be greatly affected by how we define (and filter out) weak ties to begin with.

Figure 1 illustrates different approaches to filtering weighted networks. Panel A summarizes the general logic of applying a filtering mechanism to weighted structures: by increasing the value of a thresholding parameter τ, we progressively eliminate the weakest ties (i.e., those with lower weight), thereby increasing the sparseness of the network. This exercise helps remove the noise that is often intrinsic to empirical measurements, and it allows us to identify the connections that encode the strongest signal.

Figure 1. Approaches to Filtering Weighted Networks



Note: Schematic representation of different approaches to filtering weighted networks. Weighed edges are filtered according to a threshold parameter τ (panel A). Prior work has used three approaches to selecting the specific value of τ. The first is a global approach that selects a cutting point in the overall weight distribution (panel B). The second is a dyadic approach that measures the difference between observed edge weight and the randomly expected weight for a given pair of nodes (panel C). And the third is an egocentric approach that compares observed and expected strength given the weight distribution around a focal node using the disparity filter (panel D).

One of the simplest and most intuitive filtering techniques is known as *global thresholding* (Figure 1B). This technique requires examining the weight distribution of ties and then progressively removing the weakest ties that are below a given threshold value. The impact of the chosen threshold on the resulting network depends on the statistical properties of the overall weight distribution: if it is very skewed, most of the ties will be eliminated as soon as the parameter τ reaches a relatively small value. The choice of a specific threshold is very consequential because it has an impact on the density of the network and, through that, on other network properties at the global and local levels. For example, past work analyzing email communication networks showed that with a small change in the threshold (from τ = 1 to τ = 5, where edge weight measures number of emails sent) the number of edges was reduced by an order of magnitude, and also that connectivity and clustering changed substantially (De Choudhury et al., 2010)(De Choudhury et al., 2010). These topological changes, however, offer no clear guidance on which τ should be preferred – the analyses are purely descriptive and, on their own, they offer no theoretical or statistical benchmark to decide on a threshold choice. Networks filtered according to a more or less stringent threshold are likely to look very different, in ways that are unknown if no sensitivity analyses are conducted.

Another limitation of the global thresholding approach is that it makes comparative research more difficult since the choice of a threshold is highly dependent on the features of the data analyzed. One solution to this limitation involves the use of a statistical benchmark, or null model, to assess departure from randomness. Prior work, for example, has used the conventional φ correlation coefficient to determine if a tie between two nodes is stronger than what is expected by chance, eliminating the ties that do not meet the usual criterion of statistical significance as assessed through the *t* value (Ronen et al. 2014). This *dyadic thresholding* approach (Figure 1C) has been applied to the analysis of co-exposure networks to identify the news sources that share a higher fraction of their audience than expected by chance: a positive φ correlation for a given dyad signals that there are more overlapping audience members than randomly expected (given the audience reach of each outlet in the pair, and the size of the total audience population). The associated *t* value captures the strength of that departure, offering a standardized way to operationalize the filtering parameter τ: the larger the absolute *t* value, the stronger the departure from randomness (i.e., the smaller the probability *p* of observing that departure), and the more significant the tie strength is for a particular dyad.

Unlike global thresholding, dyadic thresholding considers the tie weight distribution at the local, pairwise level, which is particularly appropriate for networks where nodes vary drastically in their ability to create ties with a given strength. In the news co-exposure network, to follow with our example, weaker ties tend to connect less prominent news outlets because these outlets have a smaller market share to begin with.

Another technique that also uses a statistical benchmark to determine the significance of tie strength is known as *disparity filter* (Serrano, Boguñá and Vespignani 2009). This approach, however, changes the definition of the null model by making it operate at the level of ego-networks instead of dyads (Figure 1D). Like dyadic thresholding, this method accounts for local disparities in tie strength, but it allows each node to evaluate the significance of all their adjacent ties; in other words, it only filters edges that are deemed insignificant by the nodes at both ends of the tie, given all their other connections. The technique makes tie significance conditional on the weight distribution of all the ties adjacent to the focal node. The null model is defined so that the normalized weights that correspond to the connections of a given node are randomly assigned from a uniform distribution. It then compares the observed weight distribution with the randomized distribution and calculates the probability that each tie could have occurred under the null model. The threshold for preserving ties is dictated by the probability of observing the weighted tie if the null model were true: a low probability (as measured by a low *p* value) implies a small chance that the tie is random and, therefore, it can be retained as statistically significant.

The disparity filter has seen widespread application in diverse contexts. Olson and Neal, for instance, used this technique to extract the backbone of a topic network on Reddit (2015); Conover and colleagues (2013) used it to analyze a Twitter retweet network during the Occupy protests; Bajardi et al. (2011) applied it to a network of cattle movements in Italy; and Zhang and colleagues (2013) applied it to analyze article citation patterns. In the analysis of co-exposure networks, this technique has been used to analyze exposure to news sources on social media platforms (Schmidt et al. 2017). This study uncovers clustering patterns that reveal selective exposure and polarization through the application of community detection methods, shedding important light on how social media platforms mediate access to news and political information. However, in line with other research, the study offers no clear justification of the threshold employed to filter the weakest connections prior to the analyses reported. As the authors state, they use the disparity filter "to determine the connections that form the backbones of [the] networks" (Schmidt et al. 2017), but the explicit *p* value used is not reported, and there is no justification of why that threshold was selected.

This discussion on sensitivity is also absent in another recent study analyzing co-exposure networks in social media (Grinberg et al. 2019). In this case, rather than focus on a particular *p* value, the authors filter the network so that "only the top 2% most meaningful relationships between sites are retained" (Ibid, SM, p. 45). The methodological discussion in the article offers no information of why this particular threshold was selected, what the associated *p* value was, or how sensitive the results are to different filtering rules. Focusing on the top 2% of edges (ordered by weight) is an appropriate choice for the purposes of the analyses presented: it directs our attention to the news outlets with the highest levels of user exposure and, therefore, to the most important sources of information. However, it is difficult to assess how the results are shaped by the choice of a specific threshold – or whether a different percentage of top edges would be most appropriate for a different network. More generally, it is not obvious if we can directly compare results across this and other similar studies because different filtering rules are applied with no sensitivity analyses provided.

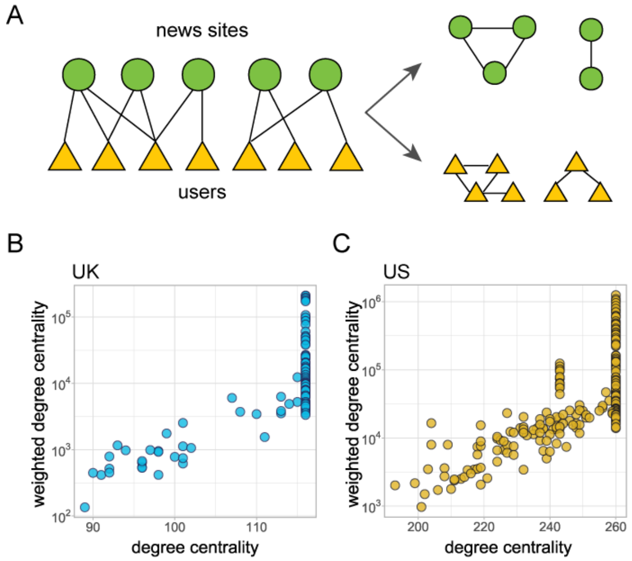
Here, we offer a systematic assessment of how network topology changes under different filtering rules and provide some evidence to guide the discussion of which filtering approach is the most appropriate if the goal is to maximize the number of ties removed while preserving the connectedness of the network. We further propose a mathematical heuristic to decide which threshold to use (e.g., which p-value), offering a more objective approach to what, otherwise, look like *ad hoc* choices. Instead of selecting a threshold by way of convention (such as $p = 0.05$) or arbitrary percentiles (e.g., top 2%), our approach is rooted in formal mathematical considerations borne out of formulating the trade-off between network sparsity and connectedness as a geometric curve-fitting exercise. In other words, our approach sees threshold selection as an optimization problem that can be solved as long as there is a trade-off between two network statistics (in our case, density and connectedness). The following section describes in more detail the data we use to illustrate this approach.

3. Data

Our data track the browsing behavior of web users accessing the sites of news outlets through mobile, tablet, and desktop devices. The data were provided by Comscore, an online audience measurement company that maintains representative panels of the online population in several countries (including the US and the UK). We use their MMX Multi-Platform audience duplication and cross-visiting reports. These data have been used in the past to cast light on selective exposure, segregation, and ideological bias in online news consumption (e.g., Gentzkow and Shapiro 2011; Mukerjee et al., 2018; Majo-Vaquez et al., 2019; Yang et al., 2020). The data are collected by means of software that is installed on the devices of consenting panelists and that passively tracks their web-browsing behavior. The device-level data is integrated with server-side data of page-views that are collected by means of tags placed in the source code of the websites and triggered every time the webpage is accessed. These data generate two kinds of monthly estimates that are particularly relevant for the analysis of co-exposure networks. The first is audience reach, or the number of unique visitors each media outlet receives every month. The second is audience overlap, which is the number of unique visitors that get co-exposed to a pair of outlets in each month.

Figure 2, panel A shows a schematic representation of the data structure. In its raw format, the data associates the panelists (triangles) to the web domains they visit (circles). From this affiliation structure, the total reach of web sites as well as their audience overlap are estimated. The ties in the one mode projection of news sources (green nodes) measure the number of unique users that visited any two news domains during a particular month. We analyze these data aggregated for June 2016 in the UK case (i.e., the month of the Brexit referendum) and November 2016 in the US case (i.e., the month of the Presidential Election). It is worthwhile noting here that the temporal granularity of the data determines the distribution of edge weights and even the connectivity of the network. Instead of monthly aggregations, we could have chosen another temporal window – a choice that would have affected the structure of the networks. Our goal is to use the monthly networks as examples to illustrate the impact that different thresholding rules have on network topology and to demonstrate our approach to threshold selection.

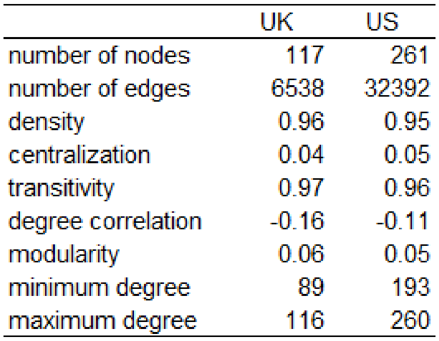
Figure 2. Raw Data Structure and Correlation of Centrality Distributions in Observed Networks



Note: We analyze the news sites projection of a bipartite structure in which unique web users are associated to the domains they visit: nodes are news outlets and the strength of edges measures the number of users co-exposed to any two outlets (panel A). The degree centrality and weighted degree centrality distributions are highly skewed and they are correlated in the two networks we analyze, which are also characterized by high density (e.g., most news outlets are connected to most other outlets through shared audiences, panels B and C). The clusters of data points on the right edge of the horizontal axes indicate that many outlets are connected to the same number of other outlets (up to the maximum of N – 1), but these connections vary greatly in terms of their strength, hence the vertical dispersion (note that the *y*-axis, which measures weighted degree centrality, is log transformed).

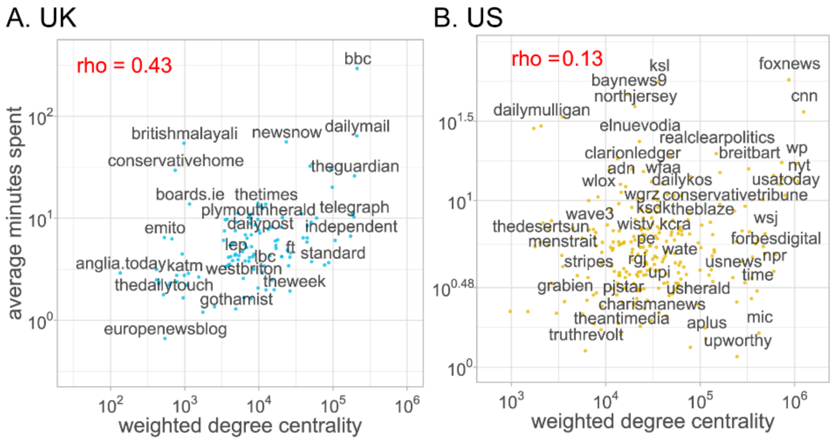
Table 1 offers descriptive statistics for the two networks. In total, the UK network includes $N = 117$ news domains and the US network includes $N = 261$ domains. Figure 3 offers some more details on the correlation between weighted network centrality and the average time (measured in minutes) unique users spend in those domains. In general, both networks are extremely dense, which means that most news sites are connected through shared audiences; but they are also very heterogeneous in terms of their weighted degree centrality, which means that most of the sites are connected to each other by weak ties. Figure 3 also shows that our two example networks differ in the correlation of the two measures (user engagement and weighted centrality). The low correlation suggests that, in general, the sites where users spend the most time are not necessarily the most central in the network – especially so for the US data.

Table 1. Descriptive Statistics for the Two Observed Networks



	UK	US
number of nodes	117	261
number of edges	6538	32392
density	0.96	0.95
centralization	0.04	0.05
transitivity	0.97	0.96
degree correlation	-0.16	-0.11
modularity	0.06	0.05
minimum degree	89	193
maximum degree	116	260

Figure 3. Correlation of Network Centrality and Time Spent



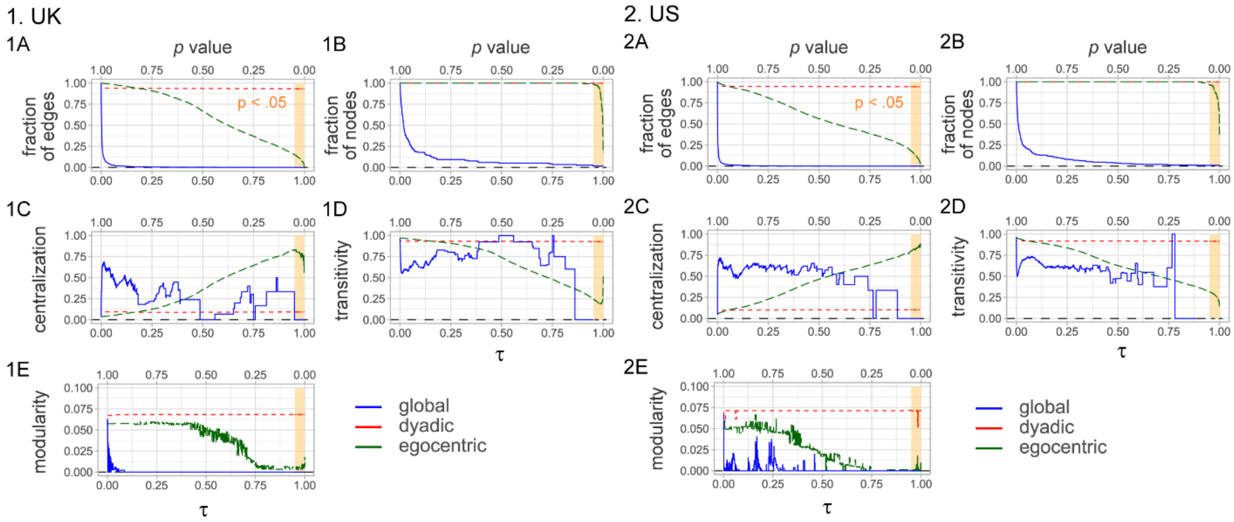
Note: These scatterplots summarize the association between the weighted network centrality (also known as graph strength) of news domains and the average time that users spend on those domains, measured in minutes. The correlations, though significant, are moderate in the UK network ($p <$ 0.001) and weak in the US network ($p < 0.05$).

Given the high density of these networks, and the skewness of the weight distribution, the question we try to answer is how to best filter these structures so that the most relevant overlap (i.e., audience similarity) becomes apparent. The three filtering techniques we evaluate here define "relevant overlap" in different ways – our first goal is to compare these three approaches and assess the impact they have on network structure. Our second and more central goal is to offer systematic guidance on how to reduce density (especially in extremely dense networks like these) while maintaining the observed connectedness (i.e., by restricting the number of nodes that become isolates). In other words, the goal is to solve the trade-off between two network statistics to identify the optimal threshold, i.e., the threshold that maximizes sparsity while minimizing the number of isolated nodes. The following section starts by assessing the topological impact of the three different filtering rules.

4. The Impact of Threshold Choice on Global Structure

Figure 4 summarizes the effects that filtering has on a set of global network properties. The first row tracks the fraction of edges and the fraction of nodes that remain connected in the network as the threshold becomes increasingly stringent (the bottom horizontal axes index the global threshold τ, the top horizontal axes index the p value associated with the dyadic and egocentric approaches). What the curves show is that global thresholding is the most aggressive at reducing the density and the size of the networks. This is because the distribution of edge weights is heavily skewed, so the vast majority of ties are eliminated at relatively low values of τ. Likewise, the egocentric thresholding approach is stricter at determining the significance of ties than the dyadic approach: at $p < 0.05$, it retains only about a fourth of the ties that remain significant according to the dyadic null model. Both approaches, however, maintain the same network size (measured as number of non-isolated nodes) for most of their parametric space.

Figure 4. Effects of Filtering on Global Network Properties



Note: These plots summarize changes in network topology as the filtering threshold becomes more stringent. The bottom axes track the threshold defined globally, the top axes track the threshold as defined by the *p*-values in the dyadic and egocentric approaches. The shaded vertical bars highlight the parametric area conventionally defined as statistically significant. The egocentric approach strikes a balance between the global thresholding (which quickly eliminates many ties but also drastically reduces the fraction of nodes that remain connected) and the dyadic thresholding (which preserves network size for the full parametric space but eliminates very few ties).

The second row tracks the centralization and transitivity scores. Centralization is a measure of network inequality: it is higher when there is more heterogeneity in degree centrality, signaling a hierarchical structure in which a few nodes accumulate most of the connections (Freeman 1979). As panels 1C and 2C show, the disparity filter used in the egocentric approach yields networks that are substantially more centralized (i.e. more unequal) than the other two filtering techniques. Transitivity, on the other hand, is a measure of triadic closure or clustering aggregated for the overall graph (Wasserman and Faust 1994). Prior to filtering, the two networks are highly clustered but, as panels 1D and 2D show, the levels of clustering systematically go down as thresholds become more severe – except for the global approach applied to the UK network, where clustering levels actually increase for τ values in the range [0.25, 0.45]. Because of the smaller impact of the dyadic filter on network density, transitivity and centralization remain mostly unchanged under this approach. The disparity filter is clearly eliminating more ties that contribute to triadic closure.

Finally, the last row tracks the modularity scores associated to a community partition derived from a random walk algorithm (Pons and Latapy 2006). The modularity of a network with respect to a node partition (in this case, community membership) measures how good the partition is accounting for connectivity patterns – or how modular we can claim the network to be (Clauset, Newman and Moore 2004). As the plots show, the modularity of the observed networks degrades as the global and egocentric thresholds become stricter, and it barely changes under the dyadic approach (unsurprisingly, since this is the approach that removes the least number of ties).

In general, global thresholding is the most volatile across its parametric space compared to the other two methods. It is also the most subjective in its selection of a specific threshold value: this approach does not anchor the filtering rule to any assessment of statistical significance. However, using a null model to assess the strength of ties is no guarantee of better results: the dyadic approach is underwhelming in its elimination of overlapping ties, with the density of the networks barely changing across the parametric space. This means that the dyadic approach does not really help identify the most important connections. From the point of view of the trade-off that motivates us, the disparity filter seems the most successful in the task of increasing sparsity while reducing the number of nodes that become isolates. However, the exploration summarized in Figure 4 still does not help identify which specific *p* value to use as the threshold. In the following section, we discuss our approach to finding an answer to that question.

5. Choosing the Optimal Threshold

We are interested in the range of p values used in the specification of the disparity filter that result in a small fraction of edges while preserving the global connectedness of the network. To visualize the trade-off, Figure 5 plots the fraction of edges that remain for all values of p as it decreases from 1 to 0 (vertical axis) versus the fraction of connected nodes (horizontal axis, mirrored). The resulting graphs for the two observed networks (panels 1A and 2A) are approximately L-shaped, i.e., they include a region of large curvature close to the origin of the axes. This is the region where removing additional edges starts to significantly affect the connectedness of the network (by increasing the number of isolated nodes). To find the p-value that balances sparsity and connectedness, we identify the corresponding point in the L-shaped curve that has the largest curvature. This approach is adapted from parameter estimation problems, where it is known as L-curve method to optimally estimate regularization parameters (Vogel 2002). While the point of largest curvature can often be visually estimated, here we propose an automated (and thus more objective and scalable) approach.

To find the point of largest curvature, we fit circles to parts of the edge-node ratio data. For each p, we consider the corresponding data point $d^k = (x^k, y^k)$, where $k = k(p)$, and the *n* neighboring points on both sides that correspond to larger and smaller p values, i.e., $d^{k-n}, \dots, d^{k-1}$ and $d^{k+1}, \dots, d^{k+n}$. To estimate the curvature for the point d^k, we fit a circle to these points and we make use of the fact that the circle's curvature is the inverse of it radius. This means that finding the point of largest curvature is equivalent to finding the point for which the fitting circle has the smallest radius.

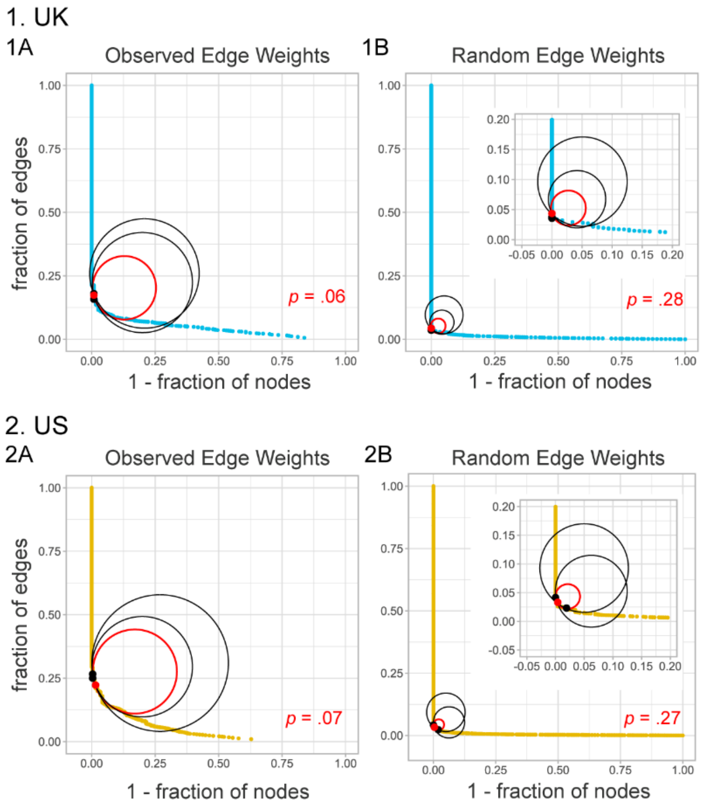
Fitting a circle to the points $d^{k-n}, \dots, d^{k+n}$ is a nonlinear regression problem with the circle's radius *r* and center (c, d) as parameters. To compute *r*, *c* and *d,* which characterize the best fitting circle, we solve the optimization problem:

$$\min_{r,c,d} \sum_{i=k-n}^{k+n} \left(\sqrt{(c - x^i)^2 + (d - y^i)^2} - r \right)^2 ,$$

which is computed numerically using an iterative Gauss-Newton algorithm (Nocedal and Wright 2006). After computing the best fitting radius r^k for each point d^k, the circle with the smallest radius r^k indicates the point of largest curvature and, thus, the point corresponding to the optimal trade-off between sparsity and connectedness. Since the edge-node ratio graph has discontinuities, we chose an *n* of moderate size for the circle regression problem.

As figure 5 shows, our approach fits circles to different parts of the curve to find the circle with the smallest radius (panels 1A and 2A). The data point on which this circle is anchored allows us to identify the *p* value that was used to filter the network: in the UK, the threshold that generates the best solution to the trade-off is *p* = 0.06; in the US, the threshold is *p* = 0.07. Both values are close to the threshold *p* < 0.05 conventionally used in social research as signaling statistical significance, but we want to emphasize the conventional nature of that choice. To find out how these solutions would change with a less skewed tie weight distribution, we generated random versions of the observed networks in which the edge weights were randomly reallocated and then averaged out over *N* = 1,000 iterations. This reshuffling process results in a network where the edge weight distribution is less centralized than the observed distribution. Panels 1B and 2B plot the corresponding L-curves and fitted circles (with insets maximizing the area of maximum curvature). For these randomized networks, the *p* values that best solve the trade-off are substantially higher (*p* = 0.28 and *p* = 0.27).

Figure 5. Solutions to the Sparsity-Connectedness Trade-Off



Note: Panels 1A and 2A plot the fraction of edges that remain for a sampling of 10,000 values of *p* in the disparity filter as it decreases from 1 to 0 (vertical axis) versus the fraction of nodes that remain connected (horizontal axis; we use 1 – fraction so that the curve visualizes as the letter L). The point that minimizes the radius of the best-fitting circle (in red) corresponds to *p* = 0.06 (UK data) and *p* = 0.07 (US data). Black circles are examples of two other circles fitting a different range of data points for different *k* with larger radius. Panels 1B and 2B plot the same curve for random versions of the networks where the edge weight distributions are arbitrarily reshuffled (*N* = 1,000 samples), with changes in the corresponding *p* value.

6. Discussion

Weighted networks are useful representations of observed relational data because they encode not just the structure of connections but also the strength of those ties. The analysis of empirical data often requires determining a threshold to define edge relevance, both to identify the most important parts of the network and to eliminate measurement error. However, it is not always obvious what the best filtering threshold should be or how that choice affects the resulting structure. Here, we have compared three techniques used in past work to determine output differences under different filtering rules. This comparison was motivated by the need to eliminate weak ties while preserving the connectedness of the observed network. We discuss a mathematical approach to this trade-off and identify the filtering technique that provides the optimal solution.

Our results suggest that defining a null model at the egocentric level offers the most compelling way to solve the tension between sparsity and connectedness. In the analysis of co-exposure networks, this allows us to highlight the parts of the network where audience overlap is stronger while preserving the information contained in global connectedness. The disparity filter provides a probabilistic rationale to select a particular threshold, offering a benchmark that can be used to compare networks drawn from different datasets. This approach is also stricter in the elimination of the least important ties while considering heterogeneity in weight distribution. And, crucially, the approach also minimizes the number of nodes that become isolates while maximizing the number of ties removed. Further, the technique we use to identifying the optimal threshold can also be used to compare networks with different weight distributions.

The question of what counts as a weak tie is contingent on the properties of the network analyzed as well as on the goals driving the research. We set up our work to solve the trade-off between network sparsity and connectedness in the analysis of co-exposure networks. We chose this trade-off for two reasons. The first is the extremely high density usually exhibited by co-exposure networks, which requires removing many ties to be able to focus attention on the most meaningful connections from a statistical point of view. The second reason is that we still want to be able to analyze co-exposure between all the sources available, which requires minimizing the number of nodes that become isolates because of filtering. Other empirical contexts might require studying different trade-offs involving other network statistics. The mathematical approach we take here, however, should also be applicable in those alternative contexts. Our approach makes the selection of a *p* value less *ad hoc* or subjective and paves the ground for more comparative work in the study of co-exposure networks. Identifying the optimal *p*-value to use as a threshold allows us to contextualize the network analyzed within the larger universe of possible, filtered networks.

The choice of an appropriate null model is essential to eliminate ties that might result from measurement problems or other sources of noise. Our results show that probabilistic thresholds offer a more standardized approach to weighted structures than a more subjective global threshold selection, but also that the selection of the null model matters greatly. The two null models we consider here operate on the one-mode projections of bipartite structures, and projections are another step of data analysis that shape the results (Coscia and Rossi 2019). One alternative we do not consider, for instance, is to define a null model on the original bipartite network, randomizing, for example, the connections across the two sets of nodes. This exercise would involve making a set of assumptions about the generative mechanisms that underlie the emergence of the network – in our case, what drives people to be exposed to certain news sources. The filtering approaches we consider here fix the total reach of news outlets and the number of other outlets with which they share an audience; in other words, they just randomize how audiences are allocated (i.e., tie weight). Future research should consider how using other null models defined on the original bipartite structure might influence the outcomes – and the implications of choosing a specific baseline for how we think about the mechanisms that drive the emergence of observed networks.

The abundance of data associated with the use of digital technologies (like the web logs or social media trails used to reconstruct co-exposure networks) offers new research possibilities that have already started to materialize in theoretical developments (e.g., Aral & Van Alstyne, 2011; Park, Blumenstock, & Macy, 2018; Grinberg et al., 2019). We can now use richer data to illuminate aspects of social structure that were difficult to capture with measurement instruments designed for smaller networks. However, digital trails also present some important methodological challenges (Golder & Macy, 2014). One of the most important difficulties relates to data cleaning and preparation, and to how to disregard irrelevant and noisy information. The filtering techniques discussed in this paper aim to eliminate the most irrelevant information in weighted networks, using different rules to define what counts as "irrelevant". The three techniques we considered here have all been used in past work, but their performance has never been compared in a systematic fashion in the analysis of co-exposure networks. This is an area of developing research, but the accumulation of evidence is hampered by the absence of sensitivity analyses or comparable benchmarks across studies

By running sensitivity analyses on two empirically observed networks, we identify the changes in network topology that result from different filtering rules, and we then identify the technique that best solves the problem of balancing sparsity and connectedness (which allows us to reduce data while preserving information about the global structure). Our examples and results have direct implications for the study of audience behavior with co-exposure networks, but the comparative method we outline is context agnostic, and can be used to resolve similar analytical trade-offs in the analysis of other weighted structures. Based on our findings, we encourage future research to conduct similar sensitivity analyses when analyzing weighted networks to allow a better integration of the outputs into cumulative and comparative work.

References

Bajardi, Paolo, Alain Barrat, Fabrizio Natale, Lara Savini, and Vittoria Colizza. 2011. "Dynamical Patterns of Cattle Trade Movements." *PloS ONE* 6(5):e19869.

Burt, Ronald S. 1978. "A Structural Theory of Interlocking Corporate Directorates." *Social Networks* 1(1):415-35.

Butts, Carter T. 2009. "Revisiting the Foundations of Network Analysis." *Science* 325(5939):414-16.

Clauset, Aaron, M.E.J. Newman, and Cristopher Moore. 2004. "Finding community structure in very large networks." *Physical Review E* 70(6):066111.

Conover, Michael D., Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2013. "The Digital Evolution of Occupy Wall Street." *PloS ONE* 8(5):e64679.

Coscia, Michele, and Luca Rossi. 2019. "The impact of projection and backboning on network topologies." Pp. 286–93 in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Vancouver, British Columbia, Canada: Association for Computing Machinery.

Del Vicario, Michela, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. 2017. "Mapping social dynamics on Facebook: The Brexit debate." *Social Networks* 50:6-16.

Fowler, J. H., B. Grofman, and N. Masuoka. 2007. "Social networks in political science: Hiring and placement of Ph.D.s, 1960-2002." *Ps-Political Science & Politics* 40(4):729-39.

Freeman, Linton C. 1979. "Centrality in Social Networks: Conceptual clarification." *Social Networks* 2(3):215-39.

Golder, Scott A., and Michael W. Macy. 2014. "Digital Footprints: Opportunities and Challenges for Online Social Research." *Annual Review of Sociology* 40(1):129-52.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake news on Twitter during the 2016 U.S. presidential election." *Science* 363(6425):374-78.

Mukerjee, S. (2021). A systematic comparison of community detection algorithms for measuring selective exposure in co-exposure networks. *Scientific Reports*, *11*(1), 1-11.

Newman, M.E.J. 2018. *Networks*. Oxford: Oxford University Press.

Nocedal, Jorge, and Stephen Wright. 2006. *Numerical Optimization*. New York, NY: Springer.

Olson, Randal S., and Zachary P. Neal. 2015. "Navigating the massive world of reddit: using backbone networks to map user interests in social media." *PeerJ Computer Science* 1:e4.

Pons, Pascal, and Matthieu Latapy. 2006. "Computing Communities in Large Networks Using Random Walks." *Journal of Graph Algorithms and Applications* 10(2):191-218.

Ronen, Shahar, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, and César A. Hidalgo. 2014. "Links that speak: The global language network and its association with global fame." *Proceedings of the National Academy of Sciences* 111(52):E5616-E22.

Schmidt, Ana Lucía, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2017. "Anatomy of news consumption on Facebook." *Proceedings of the National Academy of Sciences* 114(12):3035-39.

Serrano, M. Ángeles, Marián Boguñá, and Alessandro Vespignani. 2009. "Extracting the multiscale backbone of complex weighted networks." *Proceedings of the National Academy of Sciences* 106(16):6483-88.

Uzzi, Brian, and Jarrett Spiro. 2005. "Collaboration and Creativity: the Small World Problem." *American Journal of Sociology* 111(2):447-504.

Vogel, Curtis R. 2002. *Computational Methods for Inverse Problems*. Philadelphia, PA: SIAM.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Weaver, Iain S., Hywel Williams, Iulia Cioroianu, Lorien Jasney, Travis Coan, and Susan Banducci. 2019. "Communities of online news exposure during the UK General Election 2015." *Online Social Networks and Media* 10-11:18-30.

Yan, Xiaoran, Lucas G. S. Jeub, Alessandro Flammini, Filippo Radicchi, and Santo Fortunato. 2018. "Weight thresholding on complex networks." *Physical Review E* 98(4):042304.

Zhang, Qian, Nicola Perra, Bruno Gonçalves, Fabio Ciulla, and Alessandro Vespignani. 2013. "Characterizing scientific production and consumption in Physics." *Scientific Reports* 3(1):1640.