What Counts as a Weak Tie? A Comparison of Filtering Techniques for Weighted Networks

Subhayan Mukerjee, Tian Yang, and Sandra González-Bailón*

University of Pennsylvania

Abstract: Social networks capture interdependence by mapping relationships between actors. The ties often encode information of the relationships strength, and several filtering techniques exist to eliminate the weakest connections. However, it is not always clear which specific threshold is the most appropriate to filter ties. Here we describe three different techniques and we compare their performance with two observed networks. One of these techniques relies on the global weight distribution; the other two operate at the local (dyadic and egocentric) level. The techniques differ in whether they use a null model to define strength significance and the assumptions made by that random benchmark. We argue that the choice of an appropriate null model is essential to eliminate the least relevant ties while preserving the structural properties of the network. We also argue that probabilistic thresholds offer a more standardized approach to weighted structures than a more subjective global threshold selection.

Keywords: bipartite networks, weighted graphs, thresholding, computational social science.

* 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 grant #1729412.

Abstract

Abstract: Social networks capture interdependence by mapping relationships between actors. The ties often encode information of the relationships strength, and several filtering techniques exist to eliminate the weakest connections. However, it is not always clear which specific threshold is the most appropriate to filter ties. Here we describe three different techniques and we compare their performance with two observed networks. One of these techniques relies on the global weight distribution; the other two operate at the local (dyadic and egocentric) level. The techniques differ in whether they use a null model to define strength significance and the assumptions made by that random benchmark. We argue that the choice of an appropriate null model is essential to eliminate the least relevant ties while preserving the structural properties of the network. We also argue that probabilistic thresholds offer a more standardized approach to weighted structures than a more subjective global threshold selection.

Keywords: bipartite networks, weighted graphs, thresholding, computational social science.

What Counts as a Weak Tie? A Comparison of Filtering Techniques for Weighted Networks

Networks are the hidden architecture of social life. Their analysis illuminates the structure and effects of interdependence, with implications for how we understand phenomena like social influence, diffusion, and mobilization (Granovetter, 1973, 1978; Katz & Lazarsfeld, 1966; Rogers, 1962; Watts, 2004). With the explosion of digital technologies, the number of sources that provide high-resolution network data have multiplied. This allows us to reconstruct networks on a much larger scale (Lazer et al., 2014), as they evolve over time (Butts, 2009; Holme & Saramäki, 2012), or as ties are activated or kept dormant (Perra, Gonçalves, Pastor-Satorras, & Vespignani, 2012). One of the implications of this profusion of observational data is that the strength of ties is now easier to measure (Barthélemy, Barrat, Pastor-Satorras, & Vespignani, 2005). We can track not only connections between actors but also the bandwidth of those connections – for instance, the amount of information that flows between people (Aral & Alstyne, 2011) or the time spent in communication (Onnela et al., 2007). In its richness, however, digital data have also made the need to reduce complexity and noise ever more important. Filtering techniques that eliminate nodes or connections according to some definition of relevance allow us to fulfill that task.

In this paper, we describe three different filtering mechanisms employed in prior research to make networks sparser, and we compare their effects on network topology using two different datasets. Our goal is to determine the impact that choosing a particular threshold has on the conclusions we can draw from observed networks. In particular, we aim to resolve the problem of finding a trade-off between reducing complexity and the distortions caused by noisy

measurement, on the one hand, and preserving the substantive features that characterize the observed network, on the other. In what follows, we first discuss prior work analyzing weighted networks and the theoretical relevance of taking the strength of ties into account. We then discuss the three filtering techniques that we employ in our analyses and we present our data, which derives from web logs tracking news browsing behavior in two different countries. We look at the impact that the three filtering techniques have on the global structural properties of the networks (i.e. size, density, centralization, transitivity, and modularity) as well as the local properties (i.e. centrality rankings of individual nodes). We conclude by offering a discussion of these findings and we offer some recommendations on how to best analyze similar weighted structures.

1. Weighted Networks and the Strength of Ties

Weighted networks capture more information of social structure than their binary, unweighted versions. In unweighted networks, all ties are equivalent: they reduce connections to binary values, each representing the presence or absence of an interaction or relationship. Weighted networks, on the other hand, are more flexible at encoding differences across ties. These differences might relate, for example, to emotional closeness or the effort invested in a particular relationship. A classic approach to the strength of ties sees in that attribute a "combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (Granovetter, 1973, p. 1361). Traditionally, the measurement instruments that allowed eliciting this type of information relied on questions about the frequency of interaction or how often respondents saw a particular contact (Ibid., p. 1371). Another classic approach to the analysis of weighted structures relies on

ethnographic observations of persistent interactions in small groups or the number of social contexts in which those interactions take place (e.g., Zachary, 1977). In this case, the definition of tie strength does not rely on self-reported data but on a more objective measurement of interaction opportunities. Both approaches, however, emphasize the correlation of tie strength with embeddedness in the network: they show that weaker ties tend to be less redundant than stronger ties and that they therefore create bridges in the network, connecting nodes that would otherwise be farther apart or disconnected.

The use of digital technologies to engage in social interactions, and the trails these technologies leave behind, have made it easier to obtain reliable information on tie strength without having to rely on memory or subjective observations. These new observational data have encouraged theoretical advances around the question of why tie strength matters. For example, we are now able to study the diversity-bandwidth trade-off that was masked in past research by measurement constraints (Aral & Van Alstyne, 2011). What this trade-off highlights is that weak ties are not always the main conduit of novel information: even though weak ties increase diversity by connecting remote parts of a network (where information is less likely to be redundant), they are also activated less frequently, so they allow information to flow at a slower rate. The temporal aspect of these interactions, and the actual novelty contained in the information exchanged, were difficult to approximate with old measurement instruments. Likewise, it is now possible to determine with robust empirical data why weak ties are topologically strong. The analysis of communication networks, for instance, shows the extent to which connectivity depends on weak connections: networks are robust to the removal of the strong ties but they quickly break into disconnected components if the weak ties are removed (Onnela et al., 2007). The analysis of population-level network data also shows that the strength

of ties does not decline monotonically with network distance, as suggested by previous research: ties that span a longer distance are actually nearly as strong as the ties that connect small circles of friends (Park, Blumenstock, & Macy, 2018). These empirical patterns, which are consistent across a range of networks and communication technologies, contradict theoretical intuitions derived from the analysis of smaller weighted structures.

One type of weighted network that is common in observational research derives from projections of bipartite or two-mode structures, that is, networks formed by two types of nodes (e.g., actors and organizations) linked through at least one tie across the sets (e.g., donations, memberships or affiliations, Wasserman & Faust, 1994, p. 39). The one-mode projections of the original incidence matrix map the connections between one of the two types of nodes. For instance, if we are mapping individual affiliations to clubs, one projection is the network formed by individual actors: the ties indicate co-membership, and the weights measure the actual number of clubs to which any two actors belong. Alternatively, we can also project the network formed by the clubs, and here the ties indicate how many people belong to any given pair of organizations. One-mode projections make clusters more apparent because each group in the bipartite network results in a cluster of nodes, or a clique (Newman, 2010, p. 124). This type of weighted structures has been used to analyze a wide range of empirical phenomena, including interlocking corporate directorates (Burt, 1978); the formation of creative teams (Uzzi & Spiro, 2005); job mobility across academic institutions (Fowler, Grofman, & Masuoka, 2007); edits across linguistic editions of Wikipedia (Ronen et al., 2014); or co-exposure to news sources in social media platforms (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019; Schmidt et al., 2017). Across all these examples, tie weight adds valuable information to the clustering in the network by highlighting the ties (and the nodes) that are most strongly connected. For

instance, in the Wikipedia network, the strongest ties exist across language editions that share most bilingual editors; in the news co-exposure networks, the strongest ties exist across the news sources that share most audiences. Of course, the weight of a tie is not independent of the number of people that speak the different languages or consume the news sources at the two ends of the connection: the strongest ties will tend to connect the most spoken languages and the most popular news outlets. However, network topology and weight distribution do not always correlate in the same way. In some networks, the most central nodes (i.e. those with a higher number of connections) might also maintain the high-bandwidth ties; but in some other networks, centrality and tie strength might not necessarily be correlated: the highest bandwidth connections might, in fact, 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 this attribute is used in the analysis of an observed structure 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). Tie strength can represent bandwidth and potential for novel information to flow, as in most communication networks; or it can represent proximity or similarity, as in interlocking corporate directorates and other affiliation structures. A given research question might require dichotomizing a weighted network or pruning it so that the weakest connections are deleted. In both cases, the researcher needs to choose a threshold that defines the minimum weight requirement for a tie to be preserved. There are several substantive reasons why we might want to prune an observed network: perhaps we only care about the existence of relationships that allow a given communication bandwidth; or perhaps we want to eliminate the most fleeting interactions or

those signaling lower similarity and weaker overlap. Regardless of the reasons, the choice of a threshold is still a rather subjective exercise, with important consequences for the resulting network and the analyses reported. In the next section, we discuss three thresholding approaches previously used in the literature to transform dense weighted networks into sparser structures. To the best of our knowledge, these three approaches have never been compared in a systematic manner. Our goal is to assess the impact that these different thresholding rules have on the same network topology – and to determine which method offers the most useful filtering mechanism given the substantive nature of the data.

2. Filtering Techniques

The goal of network thresholding 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. Figure 1 illustrates this network reduction approach. 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 or bandwidth), which means we are increasing the sparseness of the network – and, therefore, making it easier to identify core structural features. This exercise helps remove the noise that is always intrinsic to empirical measurements but, most importantly, it allows us to identify the backbone of the network – the connections, for example, through which actors exchange most of the information.

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, an analysis of email communication networks showed that with a small change in the threshold (from $\tau = 1$ to $\tau = 5$, with edge weight measuring number of emails sent) the number of edges was reduced by an order of magnitude and connectivity and clustering changed substantially (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 settle on a threshold choice. As a result, published research offers findings based on networks that are likely to look very different had the authors decided to filter the ties according to a more or less stringent threshold.

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 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) is particularly useful with network data that results from the one-mode projection of a bipartite structure. For example, in the network mapping edits across linguistic editions of Wikipedia a positive φ correlation for a given dyad signals that there are more editors connecting two languages than expected by random chance (given the number of speakers in each language).

The associated *t* value captures the strength of that departure, offering a standardized way to operationalize the thresholding parameter τ: the larger the *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 for a particular dyad.

Unlike the global approach, dyadic thresholding considers the tie weight distribution at the local, pairwise level, which is particularly appropriate for multiscale networks where nodes vary drastically in their ability to create ties with a given bandwidth. In the Wikipedia network, to follow with the same example, minority languages are connected by weaker ties just because there are less people speaking those languages. Another technique that also uses a statistical benchmark to determine the significance of tie strength is known as *disparity filter* (Serrano, Boguñá, & 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 weigh 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. The technique 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 again dictated by the probability of observing the weighted tie if the null model were true: a low probability (a low *p* value) implies a small chance that the tie is random and, therefore, it can be retained as statistically significant.

This method has seen widespread application in diverse contexts. Olson and Neal, for instance, used the disparity filter 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 the disparity filter to a network of cattle movements in Italy; and Zhang and colleagues (2013), applied it to analyze article citation patterns. This technique has also been used to analyze networks of co-exposure to news sources in social media platforms (Schmidt et al., 2017). In this example, which is closer to the data analyzed here, nodes are news outlets and ties measure how many users are co-exposed to those outlets. The study uncovers clustering patterns that reveal selective exposure and polarization through the application of community detection methods to the networks analyzed. The results shed novel and interesting light on how social media mediates access to news and political information. However, the study offers no clear justification of the threshold employed to filter the weakest connections prior to the analyses. The authors write that they use the disparity filter algorithm "to determine the connections that form the backbones of our networks" (Schmidt et al., 2017, p. 3039), but there is no indication of the *p* value employed or why that threshold was selected.

In another recent article analyzing similar data but from a different social media platform, the authors apply a thresholding technique designed also to identify statistically significant connections between news outlets (Grinberg et al., 2019). In this case, they chose to filter the network so that "only the top 2% most meaningful relationships between sites are retained" (Ibid, SM, p. 45), but again there is no discussion of why this particular threshold was selected, what the associated *p* value was, or how sensitive the results are to different thresholding rules. This does not mean that focusing on the top 2% of edges (ordered by weight) is an inappropriate choice: it directs our attention to the news outlets with the highest levels of exposure and,

therefore, to the most important sources of information. However, it does mean that we cannot be sure about how much that specific threshold is driving the results or what percentage of connections would be necessary to eliminate if we were analyzing a different network and wanted to obtain comparable results. In other words, it is not obvious if the backbone network in this study is comparable to the backbone network in the study by Schmidt et al – even though they map very similar dynamics (i.e. engagement with the news in social media). This article offers some evidence to guide the discussion of which filtering approach is the most appropriate to build standardized, comparable filtering rules. The following section describes the data we use in more detail.

3. Data

The data we use to compare the three filtering techniques track the browsing behavior of web users accessing the sites of news outlets. The data were provided by an online measurement company that maintains representative panels of the online population in several countries (including the US and the UK, the two countries considered here). Figure 2, panel A shows a schematic representation of the data structure in its raw format. The data associate the panelists (white circles) to the web domains they visit (grey triangles) and estimate the reach of the web sites as well as their audience overlap. With this measure of audience overlap we can analyze the news sites projection of the original bipartite structure. In this network, the nodes are the web domains of news providers, and the ties measure the number of unique users that visited a given pair of news sources in a particular time window. We analyze these data aggregated for the month of June 2016 for the UK (the month during which the Brexit referendum took place) and November 2016 for the US (the month of the last Presidential Election). The same source of data

has been used in the past to explore questions of selective exposure, segregation and ideological bias in online news consumption (e.g., Gentzkow & Shapiro, 2011).

Table 1 offers descriptive statistics for the two networks. To be included in the data, web sites have to be accessed by a minimum percentage of the online population (otherwise it is difficult to draw estimates from the panels). In total, 134 news domains are included in the UK network (with the BBC, the Daily Mail, and the Mirror at the top of the reach distribution) and 322 domains are included in the US network (with CNN, the New York Times, and the Huffington Post leading the reach ranking). In general, both networks are very dense, which means that most news sites are connected through shared audiences; but they are also very heterogeneous in terms of degree centrality. Figure 2, panel B illustrates this distribution as well as the strong correlation between degree centrality and weighted degree centrality. This correlation derives from the fact that the most central sites in these networks are also the sites that have a higher reach – and, therefore, have a higher audience base to share with other sites. The correlation, however, is weaker at the periphery, which is likely to result from noisy measurement during data collection: by virtue of being less popular, these peripheral news sites do not provide as robust data for their visitors as the most popular sites at the core of the networks. This is precisely one of the reasons why we might want to remove the weakest connections: to eliminate noise.

The analysis of these networks sheds light on the similarity of news providers in terms of their audience base: the stronger the ties, the more similar two outlets are in the audiences they attract. In line with prior similar research, these co-exposure networks can be interpreted as a footprint of selective exposure: if similar people select similar news sources and avoid others, the networks will emerge fragmented and contain groups of web sites that are more strongly

connected internally. One network statistic that can be used to assess this level of fragmentation is modularity, which is designed to measure the strength of the division of a network into communities (Newman, 2006). This statistic ranges from -1 to 1, and it is positive if the number of edges internal to communities are higher than expected by chance. As table 1 shows, the observed networks, prior to any filtering, do not exhibit this sort of fragmentation, at least as measured by the modularity score, which is very close to 0. However, Figure 2B also shows that the edge weights in these networks span a wide range of values: as already discussed, some edges are substantially weaker than others. This begs the question of whether we could observe fragmentation if we were to filter the least significant ties – i.e., connections that result from a small number of people accessing any two news sources, perhaps because of random browsing behavior (another source of noise in the data). This is where the three thresholding techniques discussed above come in: Do they yield similar insights when applied to the pruning of these two networks? And how sensitive are the results to the choice of a specific threshold? The following two sections give an answer to these questions paying attention, first, to global network properties (e.g., size, density, centralization, transitivity, modularity) and then to local network features (e.g., degree centrality and *k*-coreness).

4. Impact of Threshold on Global Structure

Figure 3 summarizes the effects that filtering has on the set of global network properties. The first row tracks the fraction of edges and the fraction of nodes that remain 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 to 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 more strict 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 for most of their parametric space.

The bottom 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 thresholding techniques. Transitivity, on the other hand, is a measure of triadic closure or clustering aggregated for the overall graph (Wasserman & Faust, 1994, p. 243). 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 – with the exception of 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 its smaller impact on network density, transitivity and centralization remain mostly unchanged under the dyadic filter. The disparity filter is clearly eliminating more ties that contribute to triadic closure, which partly explains why it also contributes the most to increase centralization scores.

In general, global thresholding is the most volatile across its parametric space than the other two methods. This pattern also becomes clear when measuring modularity, which, as explained in the previous section, is a statistic that we can use to assess levels of fragmentation in

the filtered networks. Modularity measures the difference between the fraction of the edges that fall within a given set of communities in a network and the expected fraction if the edges were distributed randomly. While several mathematical formulations of modularity exist, the one we use here requires the optimal partitioning of the network into communities to then calculate how separated those communities are. We obtained the optimal partitioning by implementing a random walk community detection algorithm (Pons & Latapy, 2006). This algorithm operates on the intuition that a perfectly random walker will tend to get trapped in the denser parts of a network, from which we can then derive community boundaries. As Figure 4 shows, modularity scores remain low across all thresholding approaches, which suggests that these networks show no strong evidence of audience self-selection, regardless of the amount of filtering applied. Unlike the two other techniques, however, global thresholding causes again some drastic fluctuations in network topology.

5. Impact of Threshold on Centrality Rankings

In addition to assessing the impact that different thresholding rules have on the structure of networks, it is also relevant to evaluate changes in the local positions of nodes and to determine whether their relative rankings shift with the threshold applied. Figure 5 shows the correlations between the centrality scores in the original, unfiltered network and those in the filtered versions according to the dyadic approach (top row) and the egocentric approach (bottom row). We measure centrality using degree and *k*-coreness. The first metric is just the number of other nodes adjacent to the focal node: in the context of our data, this measures the number of news outlets with which a particular source shares audience. The second metric offers a social definition of centrality: the coreness of a node is defined by the highest *k*-core it belongs to, with a *k*-core

being a maximal subgraph in which all vertices have degree at least *k* (Seidman, 1983). In the context of our data, this measure assigns news outlets to nested layers of connectivity in which all other outlets are equally central, offering a good heuristic to differentiate core and peripheral sites (e.g., those with higher/lower coreness *k*).

As figure 5 shows, the correlation coefficients are in general quite strong, which means that the relative prominence of news outlets does not change drastically in the filtered (sparser) networks. However, there is clearly more disagreement in the periphery of the networks when the egocentric filter is applied: among these peripheral sites, those that were relatively well connected in the raw, unfiltered data become much less central once all ties have to meet a minimum strength or significance (as defined by the *p* value in the disparity filter). This discrepancy might reflect noisy measurement for the sites that have a lower audience reach (where the inferred overlap statistics are weaker) or it might reflect random browsing behavior from the users accessing those sites (which fails to translate into statistically meaningful patterns). Regardless of the reason, Figure 5 suggests that we might overestimate the centrality of news outlets if we do not prune their most irrelevant connections – especially among the smaller, peripheral nodes. The core outlets remain, for the most part, equally central. But not all nodes that were central in the raw, unfiltered data remain central after thresholding, especially so for the disparity filter.

6. Discussion

Weighted networks are useful data devices because they encode not just the structure of relationships but also the strength or bandwidth of those ties. Empirical data often requires determining a threshold to define edge relevance, both to identify the most important parts of the

network and to allow the most informative features to become more visible. However, it is not always obvious what the best filtering threshold should be or how that choice will affect the resulting structure. Here, we have compared three techniques used in past work to determine how different the outputs are under different filtering rules. In the background of this comparison lies the trade-off created by the need to eliminate weak signals while preserving the defining features of the observed network. There is also the question of how to find standardized filtering rules to facilitate comparison across networks with different weight distributions and degree correlations.

What our results suggest is that defining a null model at the egocentric level offers the most compelling way to meet those competing goals. First, it offers a probabilistic rationale to select a particular threshold, offering a benchmark that can be used to compare networks mapping different social systems. Second, it is stricter in the elimination of the least important ties while taking into account node diversity in the connections and bandwidth they can build. And third, it preserves the core features of the network, having a greater impact on the connectivity of peripheral nodes – while still preserving the most relevant ties in the periphery as well. This approach, in other words, is better designed to deal with the multiscale nature of observed data than global thresholding rules; and it is stricter with noisy connections than the dyadic approach, which filters much fewer ties.

The question of what counts as a weak tie is contingent on the properties of the network analyzed and on the goals driving the research. Using a specific *p* value is, in the end, also a subjective decision that relies on a conventional understanding of what counts as statistically significant. In this sense, the decision to focus on a particular percentile of the stronger ties (e.g. the top 2%, as in Grinberg et al, 2019) is as good as choosing another threshold – as long as there is some understanding of how sensitive the results are to that choice. The use of a null model

helps define a baseline to compare observed patterns with what should be expected by random chance alone. However, the definition of minimum strength that derives from such null model is a necessary but not a sufficient condition to uncover the most important connections in a network. Imposing that condition can help identify the ties that are more relevant at creating bridges or spanning structural holes (Burt, 2004); but, depending on the density of the network, a more stringent criterion than the minimum threshold for statistical significance might be necessary – for instance, being within a particular percentile within the set of significant connections. The advantage of using a null model is that by relating that criterion to the associated *p* value (0.05 or lower) it is possible to locate the analyzed network within the larger universe of possible, filtered networks.

The choice of an appropriate null model is, in any case, 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. It would also be possible, however, to define a null model on the original bipartite structure, randomizing, for example, the connections across the two sets of nodes (illustrated in Figure 2A). 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; they just randomize how audiences are allocated (i.e., they randomize the bandwidth of the ties). Future research should consider how using other null models defined on the bipartite structure might

influence the outcomes – and the implications of choosing a specific baseline for how we think about the factors that drive the emergence of the observed networks.

The abundance of data associated to the use of digital technologies (like the web logs used to reconstruct the networks analyzed here) offers new research possibilities that have already started to materialize in theoretical advances (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 irrelevance. All three techniques have been used in past work, but (to the best of our knowledge) they have never been compared in a systematic fashion – which poses difficulties for cumulative research due to the absence of basic baselines and commonly agreed standards. By running sensitivity analyses on two empirically observed networks, we identified the changes in network topology that result from different filtering rules. We identified the technique that best solves the trade-off between the need to reduce data while preserving the relevant structural features. On the basis of our findings, we encourage future research to conduct similar sensitivity analyses when analyzing weighted networks so that their findings can be contextualized and more easily integrated in comparative work.

References

Allesina, S., Bodini, A., & Bondavalli, C. (2006). Secondary extinctions in ecological networks: Bottlenecks unveiled. *Ecological Modelling*, *194*(1–3 SPEC. ISS.), 150–161. https://doi.org/10.1016/j.ecolmodel.2005.10.016

Aral, S., & Van Alstyne, M. (2011). The Diversity-Bandwidth Trade-off. *American Journal of Sociology, 117*(1), 90-171. doi:10.1086/661238

Bajardi, P., Barrat, A., Natale, F., Savini, L., & Colizza, V. (2011). Dynamical patterns of cattle trade movements. *PLoS ONE*, *6*(5), 1–51. https://doi.org/10.1371/journal.pone.0019869

Barthélemy, M., Barrat, A., Pastor-Satorras, R., & Vespignani, A. (2005). Characterization and Modeling of weighted networks. *Physica A: Statistical Mechanics and Its Applications*, *346*(1), 1–16. https://doi.org/10.1016/j.physa.2004.08.047

Burt, R. S. (1978). A Structural Theory of Interlocking Corporate Directorates. *Social Networks, 1*(1), 415-435.

Burt, R. S. (2004). Structural Holes and Good Ideas. *The American Journal of Sociology, 110*(2), 349-399. doi:citeulike-article-id:3483879

Butts, C. T. (2009). Revisiting the Foundations of Network Analysis. *Science, 325*(5939), 414-416. doi:10.1126/science.1171022

Centola, D., & Macy, M. (2007). Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology, 113*(3), 702-734. https://doi.org/10.1086/521848

Conover, M. D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., & Flammini, A. (2013). The Geospatial Characteristics of a Social Movement Communication Network. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0055957

De Choudhury, M., Mason, W. A., Hofman, J. M., & Watts, D. J. (2010). Inferring relevant social networks from interpersonal communication. *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 301. https://doi.org/10.1145/1772690.1772722

Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., & Quattrociocchi, W. (2017). Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*, *50*, 6–16. https://doi.org/10.1016/j.socnet.2017.02.002

Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., & Apkarian, A. V. (2005). Scale-Free Brain Functional Networks. *Physical Review Letters*, *94*(1), 018102.

https://doi.org/10.1103/PhysRevLett.94.018102

Fowler, J. H., Grofman, B., & Masuoka, N. (2007). Social networks in political science: Hiring and placement of Ph.D.s, 1960-2002. *Ps-Political Science & Politics, 40*(4), 729-739. doi:10.1017/s104909650707117x

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.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*, 1360-1380.

Golder, S.A. and M.W. Macy, *Digital Footprints: Opportunities and Challenges for Online Social Research.* Annual Review of Sociology, 2014. **40**(1): p. 129-152.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6425), 374-378. doi:10.1126/science.aau2706

Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics Reports, 519*(3), 97-125.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577-8582. doi:10.1073/pnas.0601602103

Karimi, F., Bohlin, L., Samoilenko, A., Rosvall, M., & Lancichinetti, A. (2015). Mapping bilateral information interests using the activity of Wikipedia editors. *Palgrave Communications*, *1*, 15041. https://doi.org/10.1057/palcomms.2015.41

Katz, E., & Lazarsfeld, P. (1955). *Personal Influence, The part played by people in the flow of mass communications*. New York, NY: Free Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … Van Alstyne, M. (2014). Computational social science. *Science*, *323*(February), 721–723. https://doi.org/10.1126/science.1169410

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577-8582. doi:10.1073/pnas.0601602103

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

Olson, R. S., & Neal, Z. P. (2015). Navigating the massive world of reddit: using backbone networks to map user interests in social media. *PeerJ Computer Science*, *1*, e4.

Onnela, J.-P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., . . . Barabási, A. L. (2007). Structure and Tie Strengths in Mobile Communication Networks. *PNAS, 104*(18), 7332-7336.

Park, P. S., Blumenstock, J. E., & Macy, M. W. (2018). The strength of long-range ties in population-scale social networks. *Science, 362*(6421), 1410-1413. doi:10.1126/science.aau9735

Perra, N., Gonçalves, B., Pastor-Satorras, R., & Vespignani, A. (2012). Activity driven modeling of time varying networks. *Scientific Reports, 2*(1). https://doi.org/10.1038/srep00469

Pons, P., & Latapy, M. (2006). Computing Communities in Large Networks Using Random Walks Pascal. *Journal of Graph Algorithms and Applications*, *10*(2), 191–218. https://doi.org/10.7155/jgaa.00124

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

Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., . . . Quattrociocchi, W. (2017). Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences, 114*(12), 3035-3039. doi:10.1073/pnas.1617052114

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks, 5*, 269-287.

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

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

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

Xie, W. J., Li, M. X., Jiang, Z. Q., & Zhou, W. X. (2014). Triadic motifs in the dependence networks of virtual societies. *Scientific Reports*, *4*, 1–13. https://doi.org/10.1038/srep05244

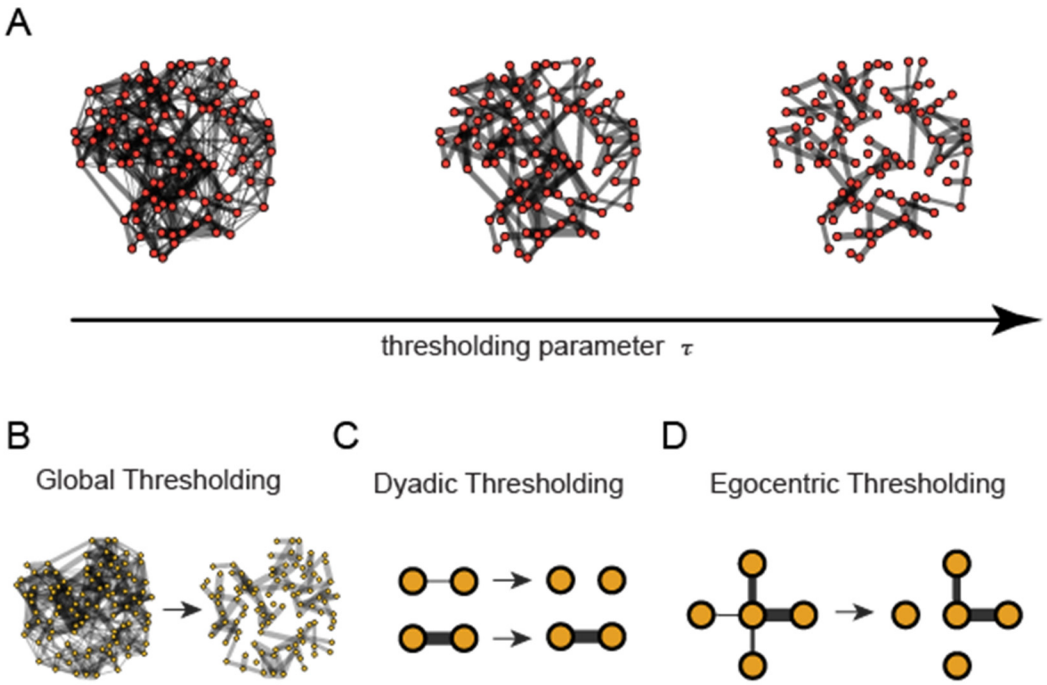
Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research, 33*, 452-473.

Zhang, Q., Perra, N., Gonçalves, B., Ciulla, F., & Vespignani, A. (2013). Characterizing scientific production and consumption in Physics. *Scientific Reports*, *3*, 1–36. https://doi.org/10.1038/srep01640

Table 1. Descriptive Statistics for the Two Observed Networks

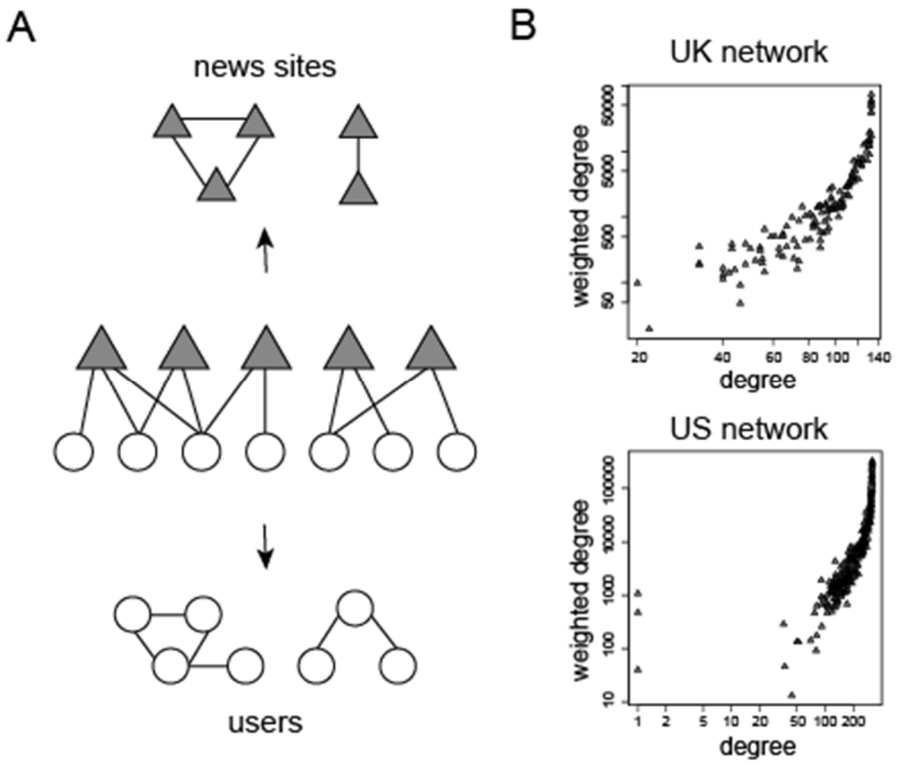
	UK	US
Number of nodes	134	322
Number of edges	6,105	33,353
Density	0.69	0.65
Centralization	0.31	0.35
Transitivity	0.80	0.77
Degree correlation	-0.27	-0.30
Modularity	0.03	0.02
Minimum degree	20	1
Maximum degree	133	318

Figure 1. Approaches to Filtering Weighted Networks



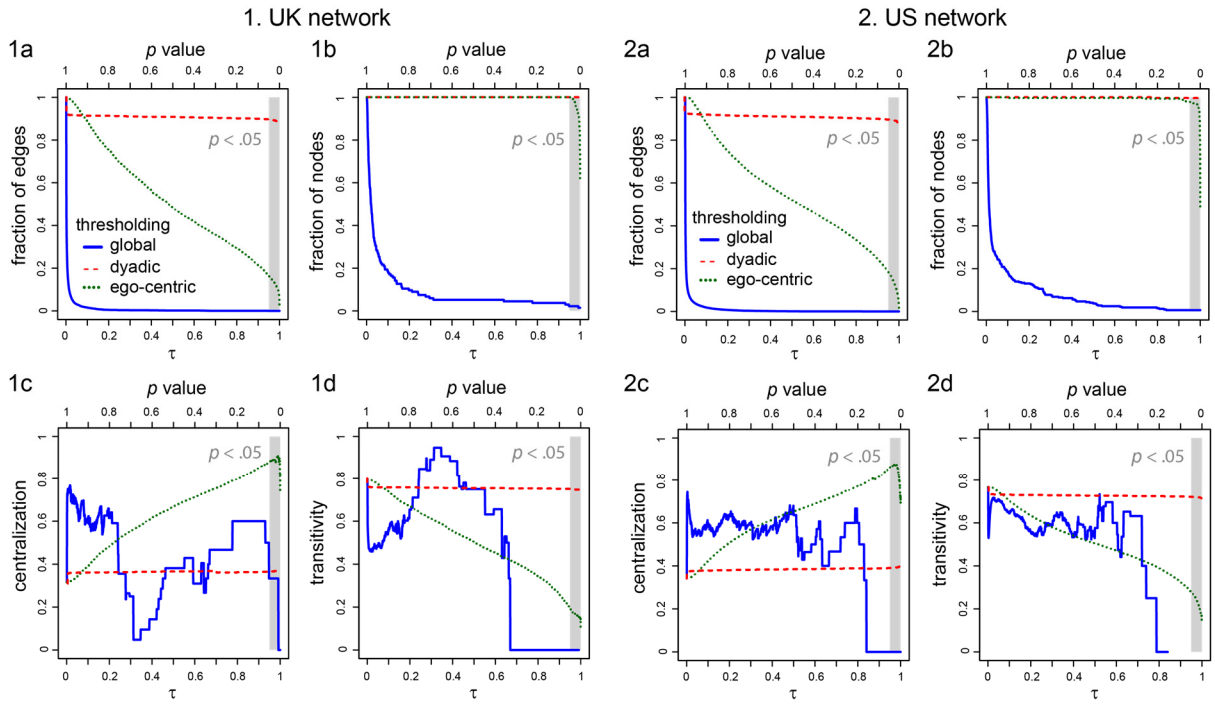
Note: The edges of weighted networks can be filtered according to a threshold parameter τ (panel A). There are different approaches to select the specific value of τ: a global approach that selects a cutting point in the overall weight distribution (panel B): a dyadic approach that measures the difference between observed edge weight and the randomly expected strength for a given pair of nodes using the φ coefficient (panel C); and an egocentric approach that compares observed and expected strength given the weight distribution around a focal node (panel D).

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



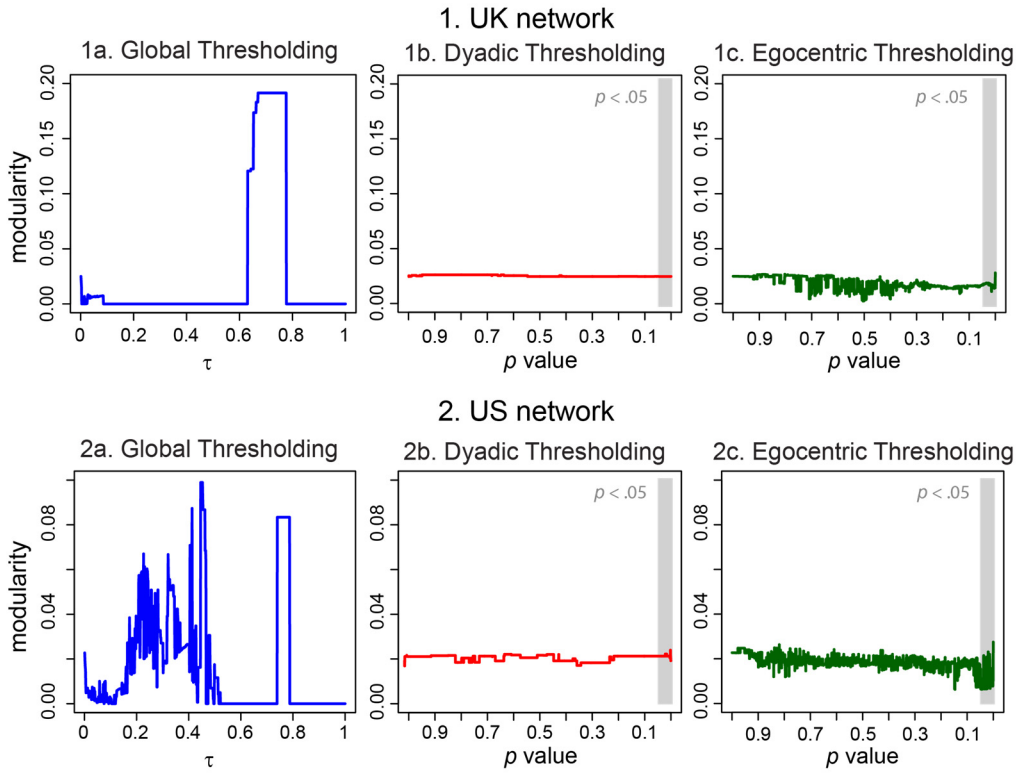
Note: We analyze the news sites projection of a bipartite structure in which unique web users are associated to the domains they visit: nodes are news outlets and the strength of edges measures the number of users co-exposed to any two outlets (panel A). The degree centrality and weighted degree centrality distributions are highly skewed and they are correlated in the two networks we analyze, although the correlation is weaker in the periphery of the network (panel B).

Figure 3. Effects of Filtering on Global Network Properties



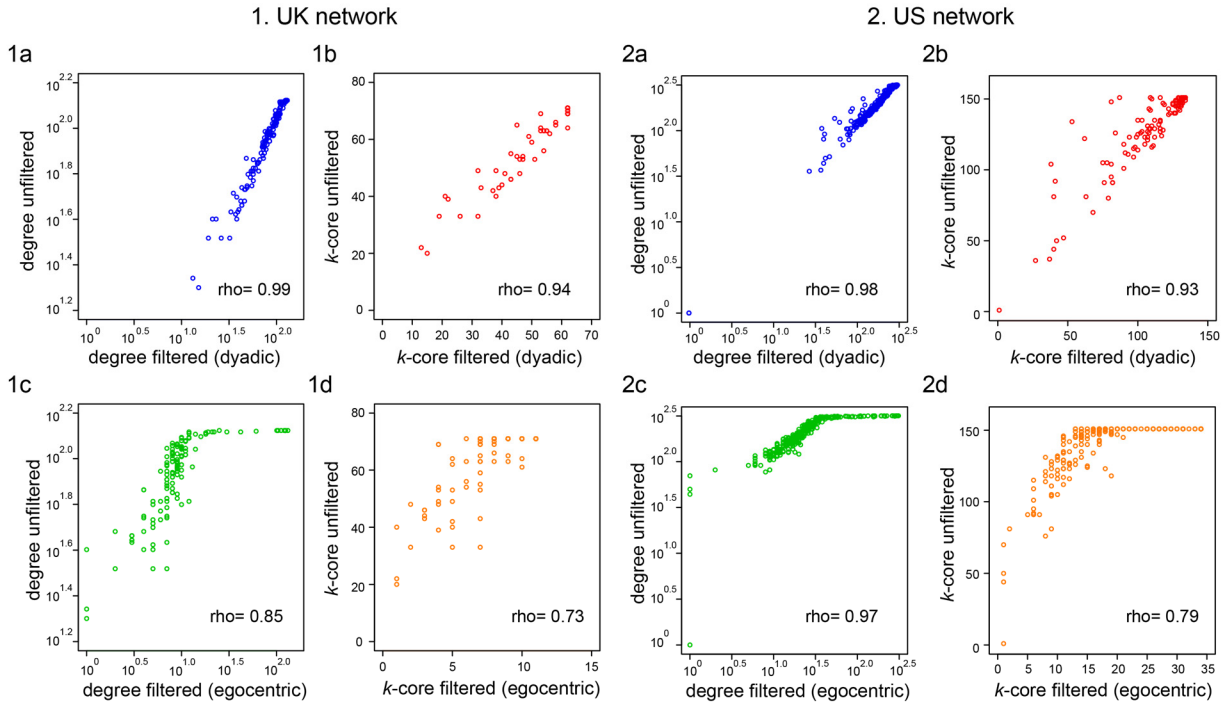
Note: These plots summarize changes in network topology as the filtering threshold becomes more stringent. The bottom axes track the threshold defined globally, the top axes track the threshold as defined by the *p*-values in the dyadic and egocentric approaches. Vertical grey bars highlight the parametric area conventionally defined as containing statistically significant results.

Figure 4. Effects of Filtering on Network Fragmentation



Note: These plots summarize changes in the modularity score of networks as the filtering threshold becomes more stringent. Vertical grey bars highlight the parametric area conventionally defined as containing statistically significant results. The modularity scores were calculated using a community detection method based on random walks.

Figure 5. Effects of Filtering on Node Positions



Note: Correlations between centrality measures (degree centrality and *k*-core) in unfiltered and filtered networks (for a $p < 0.05$). The top row shows correlations with the networks filtered with the dyadic approach; the bottom row shows correlations with the networks filtered with the egocentric approach.