Measuring Selective Exposure: A Systematic Comparison of Community Detection Algorithms in Coexposure Networks

Abstract

Network analytic techniques such as community detection have seen widespread application in recent years for the purposes of understanding audience fragmentation and selective exposure to information. In this paper, I propose a formal mathematical model for audience co-exposure networks with an aim to understand which community detection algorithm is most suitable for measuring selective exposure in such networks. I do this by simulating audience behavior in an artificial media environment and constructing a large number of synthetic co-exposure networks for various combinations of the model parameters. I then use a variety of community detection algorithms to measure the extent of selective exposure in these synthetic networks, and compare their performances. Finally, I validate these findings using a novel empirical data-set of actual large-scale browsing behavior and demonstrate the model's utility in informing future analytical choices.

Measuring Selective Exposure: A Systematic Comparison of Community Detection Algorithms in Coexposure Networks

A significant body of literature has emerged in recent years that uses network analytic techniques to understand how audiences navigate media environments (Becatti, Caldarelli, Lambiotte, & Saracco, 2019; Del Vicario, Zollo, Caldarelli, Scala, & Quattrociocchi, 2017; Grinberg et al., 2019; Ksiazek, 2011; Mukerjee, Majó-Vázquez, & González-Bailón, 2018; Schmidt et al., 2017; Taneja & Webster, 2016). These techniques typically involve the construction of *co-exposure* (Grinberg et al., 2019) or *audience overlap* (Mukerjee, Majó-Vázquez, & González-Bailón, 2018) networks from behavioral trace data, which allows researchers to effectively model media consumption patterns in a wide variety of contexts. When used in conjunction with methods such as community detection, it enables certain theoretical notions of macro-level audience behavior such as selective exposure to information, polarization in news consumption patterns, and audience fragmentation to be effectively operationalized.

While such methods have seen widespread application in the social sciences for understanding the behavior of media audiences, their implementation rarely follows a strategy that is grounded in network theory or in principles that are agreed upon a priori. Unlike the construction of these networks, which usually entails the conceptual projection of a bipartite network into an audience overlap network, the analytical choices for investigating the network are often left to the discretion of the researcher. A prominent example can be found in the use of community detection algorithms, which are widely used to identify clustering, and in turn, make inferences regarding the level of selective exposure and polarization in media consumption patterns. The lack of any systematic comparison of how different community detection algorithms perform on co-exposure networks however means that it is possible that the inferences drawn from such analyses are the artefact of the analytical choice the researcher makes. In Schmidt et al. (2017) for instance, the authors report differences in the performances of the different algorithms in the Online Appendix even as they base their main finding on one of them. Choices such as these are endemic in the literature, but they are neither motivated by a systematic comparison of the various methods, nor grounded in rigorous theoretical principles. Nonetheless, they are instrumental in furthering our knowledge about the phenomena they are used to investigate. This makes it difficult to draw more general inferences about the theoretical constructs that these methods purport to capture, precluding any possibility for comparison across the contexts in which they are applied.

This paper aims to ground these approaches in network theory by achieving three goals: first, it proposes a formal model of audience behavior in a simulated media environment, and shows how co-exposure networks of macro-level media consumption patterns can be theoretically generated by agents who choose to visit media outlets, part selectively, and part randomly. Second, it compares eight state-of-the-art community detection techniques in their ability to uncover the underlying selective exposure patterns in these simulated networks that are predicated by the model. As an addendum, it also proposes a modification in the application of these algorithms, and demonstrates how their effectiveness in measuring selective exposure can be further enhanced. Finally, the paper replicates the findings from the simulated networks on a novel empirical dataset tracking the web-browsing behavior of a nationally representative population over a long period of time.

Co-exposure Networks: Related Literature

Research on media co-exposure predates the digital age and can be traced back to the 1960s, when advertisers in the United States used metrics capturing co-exposure to television programs to understand advertising audiences (Goodhardt & Ehrenberg, 1969; Goodhardt, Ehrenberg, & Collins, 1987; Webster, 1985). Similar techniques were later used to investigate news exposure patterns and its relationship with political attitudes and behavior of the audiences. With the rise of online audiences these techniques have gradually lent themselves to the behavioral trace data that digital platforms made possible. This has resulted in a large number of studies that use online web-browsing data or social media data to construct co-exposure or audience overlap networks to understand online media consumption patterns.

A co-exposure (or audience overlap) network is defined as a network with nodes representing media outlets, and edges capturing the number of shared consumers between pairs of outlets. Conceptually, it is the *media projection* of a bipartite network that has two kinds of nodes, media outlets and media consumers, with edges between them implying an interaction between the corresponding outlet and consumer. This interaction could be a website visit, a page view, a social media “reaction” or a share. While originally used in conjunction with web-browsing data (Ksiazek, 2011; Majo-Vazquez, Nielsen, & Gonzalez-Bailon, 2019; Mukerjee, Majó-Vázquez, & González-Bailón, 2018; Webster & Ksiazek, 2012), the use of such networks has grown to witness application with social media data as well. Examples of the former include studies that used data sources such as the Nielsen TV/Internet Convergence Panel to construct a network with binary edges depending on whether or not the overlap between two outlets was higher than random (Ksiazek, 2011; Webster & Ksiazek, 2012). More recent studies have constructed similar co-exposure networks at a national (Mukerjee, Majó-Vázquez, & González-Bailón, 2018; Yang, Majo-Vazquez, Nielsen, & Gonzalez-Bailon, 2020) as well as at a global scale (Taneja, 2017; Taneja & Webster, 2016) using other online panel data sources such as ComScore. This methodological avenue has also stirred a debate in the literature regarding the use of statistical benchmarks while constructing the co-exposure networks as well on the value of using weighted ties as opposed to binarized ties (Majo-Vazquez et al., 2019; Mukerjee, Majo-Vazquez, & Gonzalez-Bailon, 2018; Webster & Taneja, 2018) (See also Coscia and Rossi (2019); Mangold and Scharkow (2020)).

With social media data, similar approaches have been used to understand selective exposure to partisan information on social media platforms. One study, for instance, used the Facebook graph API to build co-exposure networks capturing the interaction of Facebook users with Facebook pages that posted news about Brexit (Del Vicario et al., 2017). The researchers then used community detection algorithms to find two clusters of users, who interacted with Brexit news on Facebook in significantly different ways. A larger study of news consumption on Facebook ($N = 376$ million) used community detection on a similar co-exposure network of 920 media outlets to show that even outside of political news, audiences exhibit selectivity in news consumption patterns (Schmidt et al., 2017).

A more recent study constructed coexposure networks of Twitter audiences to understand the prevalence of exposure to fake news in the build-up to the 2016 Presidential Election (Grinberg et al., 2019). They used community detection algorithms to conclude that fake news reached a very niche and specialized audience, and did not feature prominently in mainstream news consumption behavior, at least among American voters who were active on Twitter. Yet other studies have used coexposure networks of Twitter audiences to understand media consumption in contexts outside the US, such as in the UK (Weaver et al., 2019) and in Asia (Lumban Gaol, Matsuo, & Maulana, 2019).

As the survey of the literature reveals, community detection techniques have been widely used to appraise the extent of selective exposure in co-exposure networks (Del Vicario et al., 2017; Grinberg et al., 2019; Schmidt et al., 2017; Weaver et al., 2019). This is because the intuition behind community detection - that strongly connected or closely clustered media outlets belong to their own community, which is structurally distinguishable from the rest of the network - lends itself to the idea of selective exposure particularly well. It allows researchers to formally capture the phenomenon that most people tend to preferentially choose certain media outlets to others, which itself has a long tradition in media studies and communication research (Katz, Blumler, & Gurevitch, 1973; Price, David, Goldthorpe, Roth, & Cappella, 2006; Zillman & Bryant, 1985).

However, a large number of community detection algorithms exist in the network science literature (Fortunato, 2009; Fortunato & Hric, 2016), and many of them fundamentally differ in how they operate (Lancichinetti & Fortunato, 2009). As a result, they often reveal completely uncorrelated community structures when applied to the same given network. Therefore, for application to co-exposure networks, which of the many community detection algorithms to use, and how their performances compare to each other remains an open question. This paper seeks to answer that question by simulating the behavior of news seekers (“agents”) in an artificial media environment, and generating a range of theoretical co-exposure networks. Then, it seeks to appraise the effectiveness of the most widely used community detection algorithms by comparing the community structures they reveal with a “ground truth” or expected community structure that is determined before the simulation is run. As is expected, the performances of the various algorithms vary greatly, with many of the algorithms failing to reveal any community structure at all. However, some algorithms work demonstrably better than the others. This paper also proposes an alteration to the method used for constructing the co-exposure networks, and shows how it significantly enhances the performances of a majority of these algorithms. Finally, it shows how the performances of these algorithms, as predicted in the theoretical networks, are favorably replicated on a large real world co-exposure network data set as well.

Agent Based Modeling

Agent based modeling is a useful modeling technique that has seen considerable application in the social sciences in recent years. Operationally, it involves modeling a social system as a “collection of decision making entities called agents” who may “execute various behaviors appropriate for the system they represent” (Bonabeau, 2002, pg. 7281). It entails a shift in thinking about social processes not in terms of factors that can cause the process, but actors or agents who constitute the process (Macy & Willer, 2002). This allows researchers to create artificial systems and societies, where agents only behave according to a limited number of predetermined rules. These rules ensure that all statistical noise from the system is effectively removed, thereby allowing the researcher to focus only on the specific aspects of the agents that they deem worthy of investigating.

Agent based models have largely been used in the social sciences for formalizing causal hypotheses and building social scientific theory. In this paper however, I use an agent-based model to test the effect of methodological choices within the confines of a sanitized artificial environment. More specifically, I use such a model to generate *theoretical* co-exposure networks which serve as the substrate for my comparative analysis of community detection algorithms. I then validate the findings from these theoretical networks using a real world network dataset thereby demonstrating the utility of the model in informing analytical choices.

Model Specification

The universe in this model comprises a set of n_1 media outlets $\mathcal{M} = \{m_1, m_2, ...m_{n_1}\}$, and a set n_2 agents (media consumers) $\mathcal{A} = \{a_1, a_2, ...a_{n_2}\}$ who have the option to visit any or none of the n_1 media outlets. Corresponding to the set of media outlets is a reputation vector $\mathcal{R} = \{r_1, r_2, ...r_{n_2}\}$ with r_i capturing the reputation of the media outlet m_i The elements in this vector are drawn from a power-law distribution with exponent α. This serves two purposes: first, the elements of $\mathcal{R}$ capture the inequality in prior popularity and reputation of the various outlets. and two, $\mathcal{R}$ serves as a weight vector, when probabilistically assigning outlets to an agent to visit, allowing agents to be more likely to visit a few reputable outlets than a large number of less reputable ones. Further, there is set of n_3 *types* of media outlets and agents $\mathcal{T} = \{t_1, t_2, ...t_{n_3}\}$, and a population level parameter ρ that determines the extent of randomness in the agents' media consuming behavior. When ρ is 0, agents behave in a fully selective manner: all the media outlets they visit are of the same type as themselves, indicating a hypothetical condition of complete selective exposure When ρ is 1, agents behave in a completely random manner, and can visit any media outlet in the universe, irrespective of the type. For any value of $\rho \in (0, 1)$, the agents visit the corresponding fraction of outlets randomly, and the rest selectively. Thus, for $\rho = 0.6$, 60% of the outlets the agents visit are random, 40% are selective. Finally, because people's media consumption habits aren't normally distributed but skew positive, the number of outlets each agent i visits, v_i is drawn from a right skewed normal distribution $N(\mu, \sigma, k)$ where the skewness k is yet another model parameter. This ensures, that in line with prior findings, a minority of the agents visit the most number of outlets, while majority visit few of them. The full model is therefore, completely specified by the parameter set $\{\mathcal{M}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \alpha, k\}$.

Simulation

The model is simulated by iterating through every agent $a_i \in \mathcal{A}$ and assigning to them $v_i \in N(\mu, \sigma, k)$ media outlets as the ones they visit. The assignment is done by using weighted sampling, where each outlet m_i is chosen with probability r_i. This is to ensure that every agent has a greater chance of visiting media outlets with higher values of r_i, so that final audience size of the media outlets resemble a power-law distribution.

Every simulation of this model yields a bipartite graph $\mathcal{G}(\mathcal{M}, \mathcal{A}, \mathcal{E})$ over the disjoint sets of $\mathcal{M}$ and $\mathcal{A}$ where an edge $e \in \mathcal{E}$ between $m_i \in \mathcal{M}$ and $a_j \in \mathcal{A}$ exists if agent a_j visited media outlet m_i. Projecting this graph onto the vertex set $\mathcal{M}$ gives the weighted co-exposure network $\mathcal{G}'(\mathcal{M}, \mathcal{E}')$. Each weighted edge $e' \in \mathcal{E}'$ in $\mathcal{G}'$ between two media outlets captures the number of agents who visited both outlets. Each such weighted co-exposure network serves as my unit of analysis, on which I apply the various community detection algorithms and measure selective exposure.

Formally, the adjacency matrix of this weighted network (one-mode projection) $\mathcal{G}'(\mathcal{M}, \mathcal{E}')$ is given by $\mathbf{B}^T\mathbf{B}$, where $\mathbf{B}^T$ is the transpose of the incidence matrix $\mathbf{B}$ of the original bipartite graph $\mathcal{G}(\mathcal{M}, \mathcal{A}, \mathcal{E})$. The usual definition for one-mode projections of bipartite networks (Newman, 2018) requires setting the main diagonal elements of $\mathbf{B}^T\mathbf{B}$ to 0. However, Arenas et al. (2008) have shown that the community structure an algorithm reveals can be extremely sensitive to the values of this main diagonal of the adjacency matrix. In the case of co-exposure networks, these values for $\mathbf{B}^T\mathbf{B}$ represent weighted self-loops capturing the number of agents who visited each outlet. Therefore, I report two sets of results for each community detection algorithm I use. The first, with the main diagonal elements of $\mathbf{B}^T\mathbf{B}$ set to 0 (I call this the “baseline” network). And second, *without* setting the main diagonal elements of $\mathbf{B}^T\mathbf{B}$ to 0 (I call this, the “augmented” network). The logic of constructing the “baseline” as well as the "augmented" networks are visually depicted in Figure 1.

– Figure 1 here –

Analysis

For each simulation, I apply eight different community detection algorithms on both, the baseline as well as the augmented network, thereby yielding sixteen resultant community substructures per simulation. The eight algorithms I choose are available in the "igraph" package, a popular network analysis library with implementations in R, Python, and C++. The eight algorithms along with their brief descriptions are as follows:

1. **Edge-betweenness** (EB) (Newman & Girvan, 2004): this algorithm works by iteratively removing the edge with the highest edge betweenness centrality, thereby isolating network clusters that would otherwise be connected to each other.

2. **FastGreedy** (FG) (Clauset, Newman, & Moore, 2004): this algorithm operates by trying to find a partition of vertices by optimizing the modularity score of the network in a greedy manner.

3. **Infomap** (IM) (Rosvall, Axelsson, & Bergstrom, 2009): an information-theoretic technique that identifies network communities by simulating the flow of information through the network structure.

4. **Louvain or Multilevel** (ML) (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008): an iterative algorithm that assigns vertices to communities in order to progressively increase the modularity score of the network.

5. **Leading Eigenvector** (LE) Newman (2006): this algorithm operates in two steps. First, it calculates the eigenvector of the modularity matrix for the largest positive eigenvalue. Then, it separates the vertices into communities based on the sign of the corresponding element in the eigenvector.

6. **Label Propagation** (LP) (Raghavan, Albert, & Kumara, 2007): an iterative algorithm that first randomly assignms labels to every vertex, and then keeps reassigning labels to every vertex based on the labels of its nearest neighbors, till reaching convergence.

7. **Spin-Glass** (SG) Reichardt and Bornholdt (2006): this algorithm is rooted in statistical mechanics and identifies communities using a spin-glass model.

8. **WalkTrap** (WT) (Pons & Latapy, 2006): a widely used algorithm that simulates the path of a random walker on the network over a long period of time, under the intuition that the walker will tend to get trapped within dense communities.

To evaluate the performance of a community detection algorithm, the revealed community structure is usually compared with a "ground truth" that is known *a priori*. In most cases, this ground truth is informed by the use the LFR benchmark (Lancichinetti, Fortunato, & Radicchi, 2008), which allows one to generate artificial networks with pre-determined community structures, that are comparable to the networks one is analyzing.

In this case however, the "ground truth" is woven into the model specification itself. We know *a priori* that there are n_3 types of media outlets. Each media outlet should, for all values of $\rho < 1$, on average, share stronger overlap with outlets of the same type, than with outlets of a different type, giving us a clear idea of the theoretical community structure that the algorithms should ideally reveal. This community structures would in turn, reflect the underlying selective exposure patterns that the simulation would have created. I leverage this knowledge and assess the performance of the algorithms using this pre-determined classification of media outlets as the benchmark. To formalize the performance metric, I use the Normalized Mutual Information (NMI) score (Danon, Diaz-Guilera, Duch, & Arenas, 2005) to compare the similarity between the vector of labels produced by the algorithm $C = \{c_{m_1}, c_{m_2}, ...c_{m_{n_1}}\}$ and the vector of outlet types, determined prior to the simulation $T = \{t_{m_1}, t_{m_2}, ...t_{m_{n_1}}\}$. The NMI value is given by equation (1):

$$NMI(C,T) = \frac{-2\sum_{i=1}^{C}\sum_{j=1}^{T} N_{ij}\log\left(N_{ij}N/N_{io}N_{oj}\right)}{\sum_{i=1}^{C} N_{i\mathrm{o}}\log\left(N_{io}/N\right) + \sum_{j=1}^{T} N_{oj}\log\left(N_{oj}/N\right)} \tag{1}$$

Here, N is the confusion matrix with rows corresponding to the outlet types assigned prior to the simulation, columns corresponding to the community labels produced by the algorithm, and N_{ij}, the number of outlets of type i assigned to community j. The row sums and columns sums are denoted by N_{io} and N_{oj} respectively (Lancichinetti & Fortunato, 2009). $NMI(C,T) = 1$ when the algorithm produces a community label vector C identical to T, and 0 when they are independent and orthogonal.

For a given set of model parameters $\{\mathcal{M}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \alpha, k\}$, I run 100 simulations for different values of ρ ranging from 0 to 1 in steps of 0.1. For each simulation, I compute eight pairs of NMI values corresponding to the eight algorithms, applied first to the baseline network, and then to the augmented network (i.e. with the self-loops).

Results

Comparison on Artificial Networks

The community structure that should *theoretically* emerge upon running the simulation depends crucially on the value of ρ that controls the extent of randomness in the behavior of the agents. For $\rho = 0$, because there is no randomness in the agents' behavior, all agents only visit the same type of outlets as themselves. In other words, there is no co-exposure between outlets of different types as no agent visits outlets of more than one type. As ρ is increased, the behavior of the agents become less and less selective till at $\rho = 1$, their behavior is completely random. Thus, theoretically, the community structure will be most prominent at $\rho = 0$, with every community (corresponding to every type) completely separated from the others. The community structure would gradually become less and less prominent as ρ is increased as more and more co-exposure occurs between outlets of different types, till, eventually, no structure would emerge owing to the high degree of randomness in the agents' behavior.

Figure 2 shows the results of simulating the model with the following parameter values: $n_1 = 100$ outlets, $n_2 = 1000$ agents, $n_3 = 5$ types, $\alpha = 3$, and $k = 3$. For each value of ρ, the simulation was run 100 times. The mean NMI scores along with their standard deviations are visualized. The pink lines and ribbons correspond to the NMI values on the "baseline" networks, while the cyan lines and ribbons correspond to the "augmented" networks with self-loops. When $\rho = 0$, all the algorithms are able to to reveal the underlying community structure perfectly ($NMI = 1$). This is because, in the absence of any random behavior, the outlets of different types share no co-exposure with each other. In other words, the only edges that exist in the network are between outlets of the same type, producing a network with $n_3 = 5$ isolated and clearly identifiable and communities. As ρ is increased, the increased randomness in the agents' behavior reduces the ability of the algorithms to reveal the perfect community structure and the NMI values gradually decline. At $\rho = 1$ almost all the NMI values drop to 0. The ones that still remain positive are owing to the algorithms overfitting the data, and often (for example, in the case of WalkTrap on the "baseline network") assigning each node to its own community. It is between these two extreme values of ρ that the algorithms reveal interesting insights.

– Figure 2 here –

The algorithms that perform the best on the "baseline" networks are Fast Greedy (FG), Multi-Level (ML), and Spin-Glass (SG) while Edge-Betweenness (EB), Infomap (IM), and Label Propagation (LP) perform the poorest. This is evidenced by the drastic drop in NMI values for relatively low levels of randomness in the agents' behavior ($0.1 < \rho < 0.5$) with EB, IM, and LP. FG, ML, SG on the other hand, are able to achieve relatively high values of NMI scores even when ($0.5 < \rho < 0.7$).

Their performances on the "augmented" network reveals a further nuance. ML and FG, which were already out-performing the other algorithms, report even higher NMI scores when the network is augmented with self-loops. Non-parametric Wilcoxon's Rank Sum test between the NMI scores on the "augmented" networks and the NMI scores on the "baseline" networks demonstrate that the former are significantly higher ($p < 10^{-3}$) than the latter for $0.7 < \rho < 0.9$[1]. For SG, LP, IM, and Leading-Eigenvector (LE), "augmenting" the network with self-loops has no effect, while for WalkTrap (WT), the addition of self-loops significantly *lowers* the NMI scores.

The analyses demonstrate that at least on the artificially generated networks, not only do FG and ML out-perform the rest, but that their performances can be further enhanced by adding self-loops to the nodes of the network with weights equal to the size of the audiences of the nodes. Thus, contrary to usual the definition of one-mode projections Newman (2018), *not* setting the main diagonal elements of $\mathbf{B}^T\mathbf{B}$ to 0 while projecting the bi-partite network and constructing the co-exposure network, allows us to uncover a more accurate community structure. These results are robust to the values of skewness paramter k, being qualitatively identical when $2 < k < 4$. For α, however, the performances of all the algorithms reduce substantially for values < 2. However, their qualitative order remains very similar, even as they all under-perform.

In the next section, I show that these findings are favorably replicated in the case of a real-world empirical network as well.

Comparison on a Real-world Network

To test whether the comparative analysis using the artificial networks mimic the findings when used with real world data, the primary challenge lies in getting a co-exposure network dataset with a clear community structure that is known a priori. For validation purposes, I therefore used a large scale dataset of desktop web-browsing behavior of Indian internet users, obtained from the media analytics firm, ComScore. ComScore maintains representative panels of online users in over 40 different countries, and they use a robust methodology called Unified Digital Measurement to integrate their panelists' data that they collect using a passive tracking software, with server side data that they capture by inserting specific tags in the source code of the web-pages. Their overall estimates, which are de-duplicated and then pre-processed to control for

[1] For $\rho < 0.7$, there is no significant difference because both ML and FG often yield perfect NMI scores of 1 on both the baseline and augmented networks

bot activities, are then made available monthly. The Indian dataset provided the perfect context to appraise the performance of the algorithms because of the high degree of linguistic diversity in the country [2] - that creates large and dedicated news reading audiences that are defined along socio-cultural and linguistic dimensions. In other words, India is a country that is characterized by a high-degree of selective exposure to media because of the large number of languages that exist in the country, and the large number of people who are fluent in their own regional vernacular. This makes large sections of Indian society selectively consume news in their own language (their own "regional" media) (Athique, 2012; Chakravartty & Roy, 2013). English media, on the other hand, while more widespread *across* the country, doesn't penetrate the rural swathes within every state, and remains the choice of only the upper-class elites across the nation.

For the analysis, I used two statistics that ComScore provided: the first, is the *total audience* of every news website (the unique number of visitors to a given website in a month), and the second is *cross visiting* for every *pair* of outlets (the unique number of visitors who visit a pair of websites in a month). These statistics are available over a 45 month period starting from October 2014 to June 2018. I use these data to first construct a "baseline" co-exposure network with media outlets constituting the nodes, and the average monthly shared audience between the outlets capturing the weight of the edge between the corresponding nodes. I also construct the "augmented" network, by adding self-loops to every node of the "baseline" network where the weight of the self-loop is the average monthly audience for the node. Further, each node is manually annotated with the language in which the outlet primarily publishes. These annotations serve as the "ground truth" for appraising the performance of the algorithms, as we would expect outlets of the same type (i.e. ones that publish in the

[2] A 2009 World Report by UNESCO, puts India at 9^{th} in the world in terms of linguistic diversity with a Linguistic Diversity Index of 0.930. In contrast, the United States, France, Germany, and the UK rank 116^{th} (LDI: 0.353) 128^{th} (LDI: 0.272), 139^{th} (LDI: 0.189), and 155^{th} (LDI: 0.139) respectively (UNESCO, 2009).

same language) to share heavier overlap with other outlets of the same type, than with outlets of different types. We would therefore expect the algorithms to assign outlets of the same type to the same community. In all, the network has 174 nodes and $14,932$ edges [3]. The network is therefore, extremely dense with an edge density (without self-loops) of 0.9921.

Figure 3 shows the results of running the eight algorithms on the "baseline" network and the "augmented" network in pink and cyan respectively. The y-axis tracks the Normalized Mutual Information score. When compared to the synthetic networks under conditions of moderate to high randomness, all the algorithms perfectly replicated their performances on the real world network as well. FG and ML out-performed the rest yet again, both working even better when the real world network was "augmented" with self-loops - achieving the two highest NMI scores respectively.

– Figure 3 here–

This analysis achieves two goals: one, by perfectly replicating the performances of a generic family of graph algorithms, it lends credence to the generative model for constructing co-exposure networks. Second, it corroborates the model's ability to inform future methodological choices, thereby giving researchers a toy model to try out various analytical approaches on, before choosing the most suitable one for analyzing real world audience data.

Discussion

Co-exposure networks and community detection algorithms are powerful analytical tools that help reveal insights into a variety of socio-political

[3] This is not a network of all possible media outlets in India, but of all the outlets that occurred every month for the 45-month period in the ComScore dataset. A website needs to register a minimum threshold of page views to be included in each month's data. Because the edge-weights of the network are monthly averages, including all the websites in the dataset would have resulted in biased estimates of the averages, as many websites did not occur in some of the months. The absence of these outlets in some months is because simply they did not have enough traffic to meet ComScore's minimum inclusion threshold.

phenomena (Becatti et al., 2019; Del Vicario et al., 2017; Grinberg et al., 2019; Ksiazek, 2011; Mukerjee, Majó-Vázquez, & González-Bailón, 2018; Schmidt et al., 2017; Taneja & Webster, 2016). Not only is the use of such techniques to model consumption intuitive, they allow for the precise operationalization of various social-scientific phenomena that had hitherto defied formal measurement. They help add richness to the quantitative description of media exposure patterns, and allow various macro-level and often emergent phenomena to be effectively modeled. While their wide usage in recent years in diverse contexts has helped cement their place in social science methodology, the manner of their application has often lacked formal justification. This has prevented prior findings from being contextualized within a robust methodological framework, as well as prevented findings from different studies in different contexts, which make different analytical choices, from being compared and consolidated systematically.

In this paper, I proposed a formal model for constructing co-exposure networks with a view to identify the most suitable community detection algorithm that could best reveal the underlying community structure reflecting selective exposure patterns. This model worked by simulating the media outlet visiting behavior of a set of pre-programmed agents in an artificial media environment. This allowed me to construct large numbers of synthetic co-exposure networks, for various combinations of the model parameters. Some of these controlled the properties of the media environment (such as the number of media outlets, number of agents, number of types of media outlets, and the overall randomness in audience behavior), while others controlled the individual behavior of the agents (such as the their own type, or the number of outlets they visit). Eight community detection algorithms were then applied to these synthetically generated networks, and their performances were appraised with the use of the Normalized Mutual Information metric. I also proposed a methodological modification, informed by prior work (Arenas et al., 2008), to further enhance the performance of these algorithms. This modification - which entailed adding weighted self-loops to every node in the network - enabled two of the eight algorithms, Fast Greedy and Multilevel, to significantly outperform the others in being able to better unveil the underlying selective exposure patterns that the model predicated. However, because a real world network may behave very differently from how artificial networks behave in theory, I repeated the comparison of the algorithms on a novel, large web-browsing network data set of a nationally representative sample of Indian internet users. Comparison of the algorithms on this empirical data showed that their performances on the artificial networks were faithfully replicated, with Fast Greedy and Multilvel significantly outperforming the others and yielding a high NMI score.

There are two key takeaways of this research. The first applies to future studies seeking to characterize selective exposure patterns by identifying the underlying community structure of co-exposure networks. My results suggest that Fast Greedy and Multilevel are the two most suitable algorithms for this purpose, that work remarkably well even in the presence of substantial noise, when many of the others fail to identify any community structure at all under similar conditions. For future researchers interested in using community detection in co-exposure networks to appraise selective exposure patterns, Fast Greedy and Multilevel could therefore hold great promise.

The second takeaway is more general: that a formal model can effectively help us make suitable analytical choices, instead of solely informing social scientific theories. Of course, a model, by definition is an over-simplification of reality and there are many ways in which any model can be made better. While this set of simulations yielded promising results, I recognize that there are various parameters that could be further added to make this model even more complete and realistic. For instance, instead of a population level randomizing parameter ρ, one could use an agent-level randomizing parameter. The distribution of this parameter, could in turn be informed by empirical data of how selectivity in news consumption patterns is distributed in an audience population. Next, instead of having the same number of outlets and agents per type, as this model does, a future model could introduce more complexity by adding variability in these distributions, and then assess whether the findings qualitatively change. Finally, my model was limited by the availability of computing power in the number of outlets and agents I could simulate. Future studies can explore if these findings replicate for larger numbers of outlets, and more importantly, large numbers of agents. There is also the need to validate the findings from this model on other real world co-exposure networks. While the ComScore Indian dataset offered a useful empirical test, it remains to be seen if the community detection algorithms tested here behave similarly in co-exposure networks derived from other real world contexts such as social media. I invite future scholars to use this model to inform their own modeling exercises, and build on it in the future.

References

Arenas, A., Fernandez, A., & Gomez, S. (2008). Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, *10*(5). Retrieved from `http://arxiv.org/abs/physics/0703218` `http://dx.doi.org/10.1088/1367-2630/10/5/053039` doi: 10.1088/1367-2630/10/5/053039

Athique, A. (2012). *Indian media global approaches*. Polity Press. Retrieved from `https://books.google.com/books?id=Dc-EKcI-s8MCprintsec=frontcoverdq=indian+media-` `media adrianf=false`

Becatti, C., Caldarelli, G., Lambiotte, R., & Saracco, F. (2019, Aug). Extracting significant signal of news consumption from social networks: the case of twitter in italian political elections. *Palgrave Communications*, *5*(11), 1–16. doi: 10.1057/s41599-019-0300-3

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008, Oct). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. doi: 10.1088/1742-5468/2008/10/P10008

Bonabeau, E. (2002, May). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7280–7287. doi: 10.1073/pnas.082080899

Chakravartty, P., & Roy, S. (2013). Media pluralism redux: Towards new frameworks of comparative media studies "beyond the west". *Political Communication*, *30*(3), 349–370. doi: 10.1080/10584609.2012.737429

Clauset, A., Newman, M. E. J., & Moore, C. (2004, Dec). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111. doi: 10.1103/PhysRevE.70.066111

Coscia, M., & Rossi, L. (2019, Aug). The impact of projection and backboning on network topologies. In *Proceedings of the 2019 ieee/acm international conference on advances in social networks analysis and mining* (p. 286–293). Association for

Computing Machinery. Retrieved from https://doi.org/10.1145/3341161.3342862 doi: 10.1145/3341161.3342862

Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005, Sep). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, *2005*(09), P09008–P09008. doi: 10.1088/1742-5468/2005/09/P09008

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. doi: 10.1016/j.socnet.2017.02.002

Fortunato, S. (2009). Community detection in graphs. Retrieved from http://arxiv.org/abs/0906.0612 http://dx.doi.org/10.1016/j.physrep.2009.11.002 doi: 10.1016/j.physrep.2009.11.002

Fortunato, S., & Hric, D. (2016, Nov). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44. doi: 10.1016/j.physrep.2016.09.002

Goodhardt, G. J., & Ehrenberg, A. S. C. (1969, May). Duplication of television viewing between and within channels. *Journal of Marketing Research*, *6*(2), 169–178. doi: 10.1177/002224376900600205

Goodhardt, G. J., Ehrenberg, A. S. C., & Collins, M. A. (1987). The television audience: patterns of viewing. an update. *The television audience: patterns of viewing. An update.*(No. Ed. 2). Retrieved from https://www.cabdirect.org/cabdirect/abstract/19871848487

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

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *Public Opinion Quarterly*, *37*(4), 509–523.

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

10.1080/08997764.2011.626985

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*(5), 056117. doi: 10.1103/PhysRevE.80.056117

Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008, Oct). Benchmark graphs for testing community detection algorithms. *Physical Review E*, *78*(4), 046110. doi: 10.1103/PhysRevE.78.046110

Lumban Gaol, F., Matsuo, T., & Maulana, A. (2019, Sep). Network model for online news media landscape in twitter. *Information*, *10*(99), 277. doi: 10.3390/info10090277

Macy, M. W., & Willer, R. (2002). From factors to factors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*(1), 143–166. doi: 10.1146/annurev.soc.28.110601.141117

Majo-Vazquez, S., Nielsen, R. K., & Gonzalez-Bailon, S. (2019). The backbone structure of audience networks: A new approach to comparing online news consumption across countries. *Political Communication*, *36*(2), 227–240. doi: 10.1080/10584609.2018.1546244

Mangold, F., & Scharkow, M. (2020, Apr). How do filtering choices impact the structures of audience networks? a simulation study using data from 26 countries. *Communication Methods and Measures*, *14*(2), 125–144. doi: 10.1080/19312458.2020.1724274

Mukerjee, S., Majó-Vázquez, S., & González-Bailón, S. (2018). Networks of audience overlap in the consumption of digital news. *Journal of Communication*, *68*(2), 26–50. doi: 10.1093/joc/jqx007

Mukerjee, S., Majo-Vazquez, S., & Gonzalez-Bailon, S. (2018). Response to webster and taneja's response to "networks of audience overlap in the consumption of digital news". *Journal of Communication*, *68*(June), 15–18. doi: 10.1093/joc/jqx007

Newman, M. E. J. (2006, Sep). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, *74*(3), 036104. doi:

10.1103/PhysRevE.74.036104

Newman, M. E. J. (2018). *Networks* (Second Edition ed.). Oxford University Press.

Newman, M. E. J., & Girvan, M. (2004, Feb). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. doi: 10.1103/PhysRevE.69.026113

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *10*(2), 191–218. doi: 10.7155/jgaa.00124

Price, V., David, C., Goldthorpe, B., Roth, M. M. C., & Cappella, J. N. (2006). Locating the issue public: The multi-dimensional nature of engagement with health care reform. *Political Behavior*, *28*(1), 33–63. doi: 10.1007/s11109-005-9001-2

Raghavan, U. N., Albert, R., & Kumara, S. (2007, Sep). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*(3), 036106. doi: 10.1103/PhysRevE.76.036106

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. , 1–16. doi: 10.1103/PhysRevE.74.016110

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009, Nov). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. doi: 10.1140/epjst/e2010-01179-1

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). doi: 10.1073/pnas.1617052114

Taneja, H. (2017). Mapping an audience-centric world wide web: A departure from hyperlink analysis. *New Media and Society*, *19*(9), 1331–1348. doi: 10.1177/1461444816642172

Taneja, H., & Webster, J. G. (2016). How do global audiences take shape? the role of institutions and culture in patterns of web use. *Journal of Communication*, *66*(1), 161–182. doi: 10.1111/jcom.12200

UNESCO. (2009). *The unesco world report on cultural diversity | united nations educational, scientific and cultural organization.* Retrieved from `http://www.unesco.org/new/en/culture/resources/report/the-unesco-world-report-on-`

Weaver, I. S., Williams, H., Cioroianu, I., Jasney, L., Coan, T., & Banducci, S. (2019, May). Communities of online news exposure during the uk general election 2015. *Online Social Networks and Media*, *10–11*, 18–30. doi: 10.1016/j.osnem.2019.05.001

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

Webster, J. G., & Ksiazek, T. B. (2012). The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of Communication*, *62*(1), 39–56. doi: 10.1111/j.1460-2466.2011.01616.x

Webster, J. G., & Taneja, H. (2018). Building and interpreting audience networks: A response to mukerjee, majo-vazquez gonzalez-bailon. *Journal of Communication*, *68*(June), 11–14. doi: 10.1093/joc/jqy024

Yang, T., Majo-Vazquez, S., Nielsen, R. K., & Gonzalez-Bailon, S. (2020, Mar). Exposure to news grows less fragmented with an increase in mobile access. *Forthcoming in the Proceedings of the National Academy of Sciences.* Retrieved from `https://papers.ssrn.com/abstract=3564826` doi: 10.2139/ssrn.3564826

Zillman, D., & Bryant. (1985). *Selective exposure to communication.* Routledge.

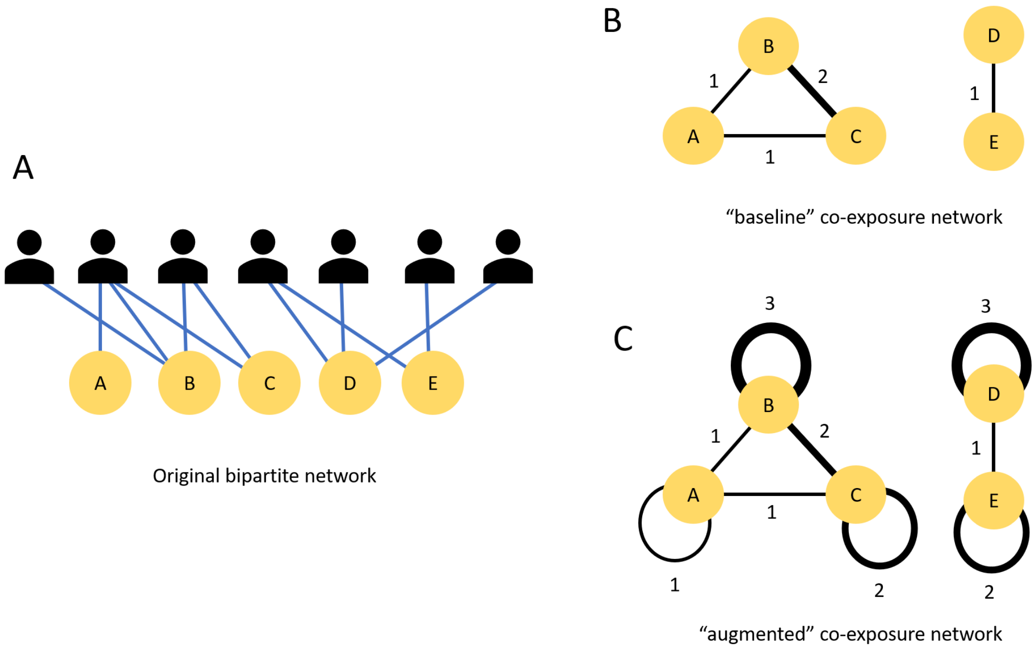
A
A
B
C
D
E
Original bipartite network
B
B
A
C
1
2
1
D
E
1
"baseline" co-exposure network
C
3
B
A
C
1
2
1
1
2
3
D
E
1
2
"augmented" co-exposure network

Figure 1. The construction of baseline and augmented co-exposure networks. Panel A shows the original bipartite network with each edge indicating that an agent visited an outlet. The outlet projection of this bipartite network is shown in Panel B ("baseline"). The edges here indicate co-exposure between pairs of outlets, with the weight of the edge capturing the shared audience between them. Panel C shows how the co-exposure network is augmented by adding self-loops with weights equal to the size of the audience of the respective node.

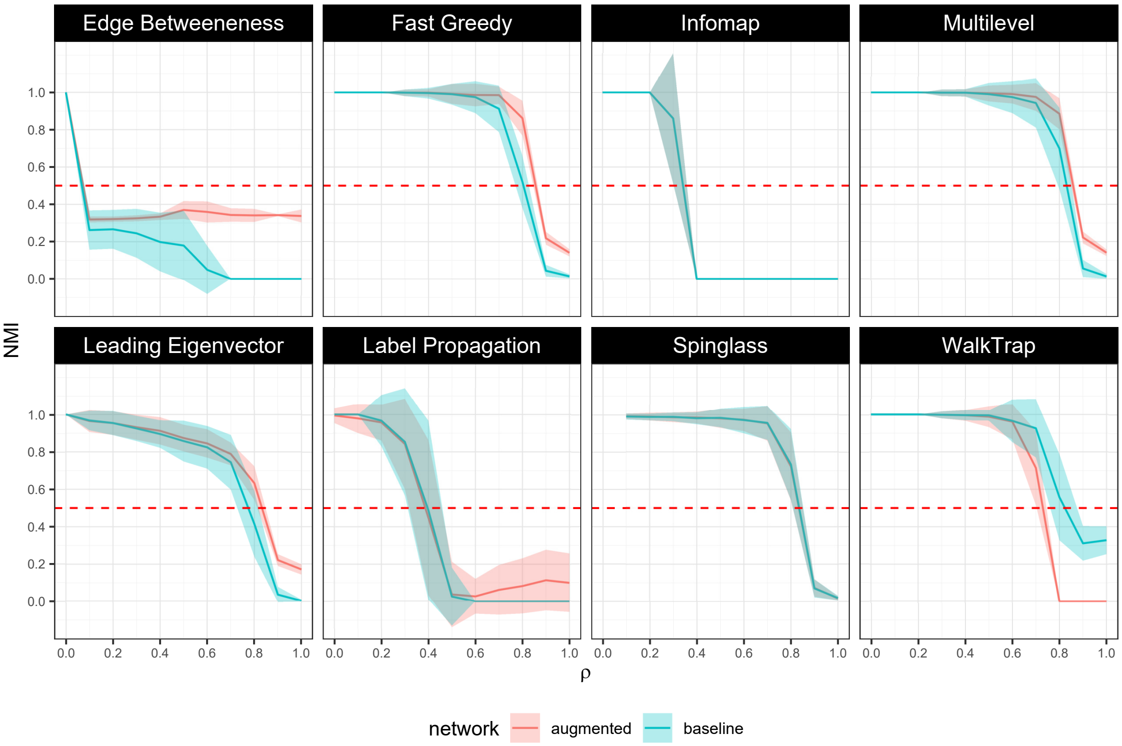

Figure 2. The performances of the 8 community detection algorithms are shown on the "baseline" (pink) and "augmented" (cyan) networks respectively. The y-axis tracks the Normalized Mutual Information value for each algorithm for every value of ρ, that determines the extent of randomness in the agents' behavior. As ρ increases, NMI values drop. Fast Greedy, Multilevel, and Spinglass outperform the others even when the randomness is substantial, and the performances of FG and ML are further enhanced with the augmented network.

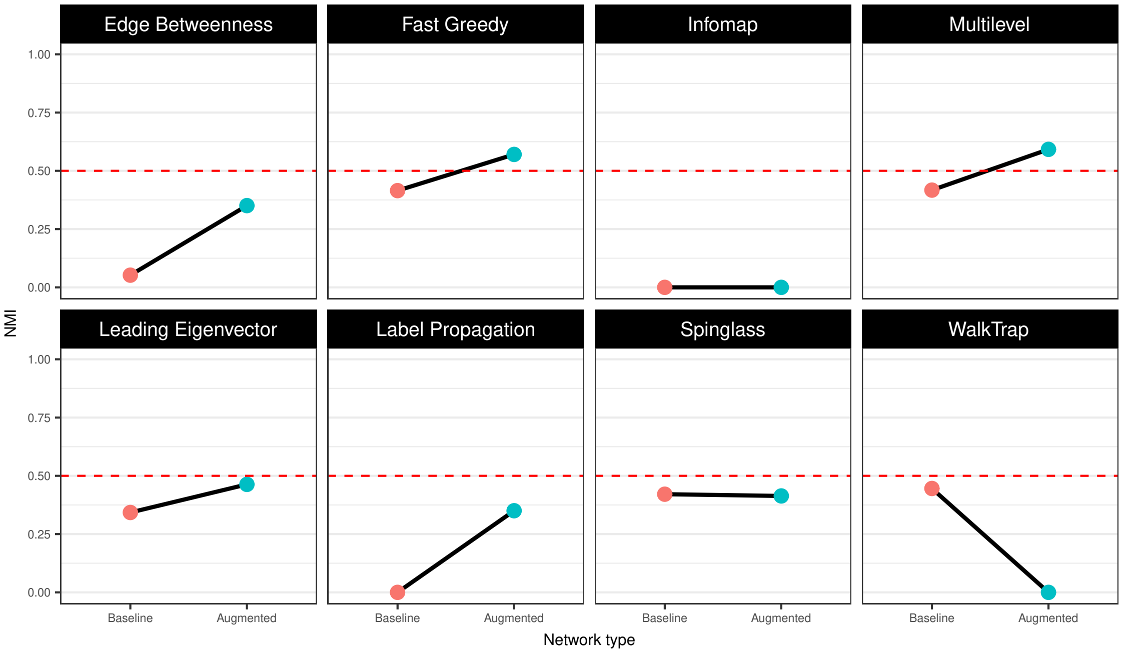

Figure 3. The performances of the 8 community detection algorithms are shown on the “baseline” (pink) and “augmented” (cyan) empirical network. The y-axis tracks the Normalized Mutual Information value for each algorithm. As with the synthetic networks, Fast Greedy, Multilevel, and Spinglass outperform the others on the “baseline” network. The performances of FG and ML are yet againfurther enhanced with the augmented network.