

Response to Webster and Taneja's Response to

"Networks of Audience Overlap in the Consumption of Digital News"

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

*Annenberg School for Communication, University of Pennsylvania

** Reuters Institute for the Study of Journalism, University of Oxford

We appreciate the chance to respond to Webster and Taneja's comments to our article. We take this as a welcome opportunity to engage in the always pertinent discussion on why measurement and methods are so important to reach meaningful theoretical conclusions from empirical work.

For the sake of transparency, we would like to begin by stating that we sent the manuscript of our article to Webster and Taneja at the same time we submitted it for review to the *Journal of Communication* (back in May 2017), asking for comments or clarifications in case we had misinterpreted their analyses or findings. Both authors amiably acknowledged receipt, but they never sent us any written feedback until we received their response from the editor (March 2018).

We would also like to state that in our article we acknowledge what we genuinely think is an important contribution: their work was amongst the first to apply network methods to the analysis of audience data. As we write in page 30, this past research "shows that the analysis of audience overlap data can offer relevant insights on exposure to information and media diets" and that the "network analysis of audience behavior can offer a powerful methodological approach to uncover the characteristics of media diets and the venues where audiences concentrate more clearly". However, having carefully read their response, we still stand by the three main claims we make in our article: (1) that past work did not offer a satisfactory benchmark to assess significant overlap between outlets that are very different in reach; (2) that disregarding the strength of the overlap misses crucial information when analyzing audience networks; and (3) that once overlap strength is taken into account, the conclusions that can be drawn from the analysis of audience networks are not consistent with the idea of a "massive overlap culture".

First, we disagree with Webster and Taneja's claim that our method for testing significance mishandles sampling error. Statistical modeling is not only about inferring from samples to populations; there are other sources of noise that affect estimates, for instance measurement error, and standard errors are useful to account for these. We take comScore estimates of audience overlap as reflecting a great deal of uncertainty: we see these estimates as the outcome of an underlying stochastic process that could have produced different results, even if the data had been collected in the same conditions. This means that we need a test of significance to determine how different the observed numbers are from other outcomes that could also have

resulted from this stochastic process. One way to think of this is seeing the comScore panel as a sample from a larger population of possible panels: standard errors help us assess uncertainty in the overlap metrics that comScore provides with the one panel we have access to. As we note in the article, page 35: “it is important to determine not only how much overlap there is but also whether this overlap is beyond a margin of error as determined by usual probability procedures”. Again, our goal is not to infer from samples to populations but to model the data generating process by explicitly accounting for noise.

It is true that we only have access to population estimates (like everyone else analyzing comScore data); but these include the total number of unique users that access a given site and, of those, how many access a second site (from where we infer the weighted ties). These are the counts that we use to determine whether we should expect the observed overlap by chance, considering how different in absolute reach news sites are. Webster and Taneja mention the “duplication of viewing law”. We have serious doubts that what they refer to as “law” can actually be considered a law, but we also take issue with how they make use of this past research to justify their analytical choice. In their response, they write: “the expected (random) duplication is a probability of a person visiting both outlets in a pair if visiting each is independent of the other”. Again, there is uncertainty attached to these probabilities (as inferred from the aggregated percentages that comScore provides) and we therefore believe it is a better strategy to calculate confidence intervals around estimates of overlap. We then use those confidence intervals to determine where the observed overlap is likely to result from random chance. Needless to say, we could not disagree more with the statement that Webster and Taneja make that “ultimately, deciding when audience traffic constitutes a link is more the province of theory than inferential statistics”.

On a lesser note, according to the information provided to us by comScore the panel size for the US is not a million, as Webster and Taneja claim in their response, but between 200,000 and 270,000, depending on the year. And the minimum time spent on a site required for that activity to count as a unique visit is 3 seconds, not 1. Regardless, the comScore data (like all empirical data) has limitations that we explicitly discuss in the article, especially in the discussion section. Uncertainty in the overlap estimates is one limitation we explicitly deal with in our work, unlike past research.

Second, we also disagree with the statement that analyzing dichotomous ties is just “setting the bar too low”. Ignoring the strength of the overlap in the construction of audience networks just leads to misleading conclusions. In our article, we show that the ranking of most central sites changes drastically when tie weights are taken into account. In their response, Webster and Taneja write that looking at weighted eigenvector centralities is not the only way to reach the conclusion that news consumption is “narrowly concentrated”. But that was not our point. Our point was to illustrate how different the conclusions that can be drawn from audience networks are when weights are disregarded. We focused on centrality because that was the measure used in past work, and we show that when we take edge weights into account, we obtain a very different ranking of centrality. Once weights are taken into account, it is obvious that audiences are highly concentrated in a few outlets; it is also obvious that the relationship between network

centrality and audience reach gets lost if the network is analyzed as unweighted. Webster and Taneja's choice to dichotomize the overlapping ties is, to use an analogy, equivalent to assuming that roads connecting little towns are equivalent to highways connecting cities. It is trivially true that they are all connected, but this choice (which, as we describe in detail in the article, was prevalent in past research) masks important differences in how audiences browse online sites.

And third, once these empirical patterns are taken into account, it is difficult to argue that the observed overlap data can be best described as "massive overlap". What our findings show is that there is a small core of highly accessed sites and a larger periphery of sites that are more loosely connected to the core. It is true that we only focus attention on news sites, so our claims might not be generalizable to other types of online media, but to the extent that audience reach is unevenly distributed, we predict similar core-periphery structures in other audience overlap networks. As we state in the conclusions, this finding is compatible with theories of niche consumption, but the "massive overlap" only happens at the core, not amongst the niche news sources – at least if by "massive" we assume that large numbers of consumers are involved. Webster and Taneja write that we have no empirical basis upon which to refute Webster's characterization of massive overlap. Rather than refuting it, we were trying to qualify the argument (see third paragraph in page 43). We do not argue against the existence of the "massively parallel cultures" or "microcultures" that Chris Anderson discussed in his work (as cited by Webster and Taneja in their response). We do state, on the basis of our data, that the massive overlap is limited to the core of the network and does not characterize consumption behavior in general. In any case, we defer to the reader to decide whether our caveats are warranted on the basis of the evidence provided.

We believe in cumulative research, in the constant improvement of methodological tools to obtain more accurate empirical insights – and get progressively better at theorizing about the patterns observed. In our article, we offer a careful discussion of past work and we offer what we believe is a more appropriate approach to the analysis of audience networks. We also believe that this is just a step in a more ambitious research program that should (a) improve transparency in the choices made when analyzing data and (b) test the implications of those choices for the interpretation of results. Two lines of work that we are currently developing involve testing the impact that different thresholding techniques have on the structure of audience networks, and using those techniques to come up with standardized measures that can be used to compare patterns across countries and in time. This constant attempt to improve the tools and the metrics that we use to analyze data is, we believe, the key to theoretical progress and our article was written in that spirit. Of course, it is ultimately up to the research community to decide if the research we report in our article offers a substantive improvement to how audience networks were analyzed in the past, and a promising step towards future developments.