**Yellow means medium difficulty, red highest, blue lowest, white comments not needing changes All author comments are green**

Your Article has now been evaluated by 4 referees. You will see from their comments copied below that, although they find your work of considerable interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication in the current form, but we would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.  
  
We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.  
  
As well as addressing all of the other points raised, it will be important to address Reviewer #3's point about the identification of influencers, and to re-run this analysis using Reviewer #3's suggested approach in order to avoid over-representation of certain types of influencer e.g. extreme left. Also, it will be important to convincingly address Reviewer #4's concern over whether retweeting behaviour is a valid measure of ideology.

* **Alex**: I think he [Reviewer #3] meant doing the influencer classification on the whole network instead of per media categories. I think we have to do it to show how it looks, but it will clearly be dominated by left/center news. So I don't think it's a good idea to use it for ideology estimation. Actually, the method of the latent ideology classification requires to start from a set of "elites" that represent all ideologies equally. Then, as we use all the retweets in the ideology estimation analysis, the balance between important and less important influencers is taken into account. I think we need to explain all this better.
* **James**: I agree with Alex that we should do the analysis of top overall influencers, with a follow-up analysis of the news category classification make-up of maybe the top 100 for both election years. We can do this for both CI and PageRank too, and then put it in the supplemental materials followed by an appropriate, more detailed explanation of our influencer extraction methodology in the main manuscript. This way we can show how it looks, but not integrate it into our analysis. As Alex said, doing this overall influencer extraction approach would not be appropriate for ideology estimation, or the modularity analysis, as this would give a lop-sided representation of specific news categories, which would make our results biased.

IMPORTANT: we have two editorial points to raise:  
  
- Reviewer #1 suggests that the full dataset for the study should be published alongside this paper if it is accepted. While we are in favor of data sharing, owing to possible privacy concerns we feel that the full data should not be published in this case, but it should be made available to researchers upon reasonable request, if and when this paper is published.

* **Bolek**: Solution, do not provide link to the data, and remove existing link, what was published before is not via NHB so legally they cannot complain
* **Hernan**: Too late. I have already released all datasets and codes. Alex/Matteo: is this correct?
* **Alex**: Yes, we have released the Tweet IDs, not the full data. I guess we can make the full data available to other researchers upon reasonable request, but I'm not sure. We should check Twitter policies.

- The revised manuscript should remove all identifying information about particular Twitter accounts. This should be removed not only from the main text, but also from the Supplementary Information. We recognize that this is not the same policy as was applied to a Nature Communications paper published by your group a few years ago ("Influence of fake news in Twitter during the 2016 US presidential election"), and we understand that you may have questions about this request. If so, please get in touch with me and I'll be happy to discuss this further over the phone or on Zoom.

* **Bolek**: Talk to editor because Twitter allows account ID but not names of the owners
* **Hernan**: This is would be catastrophic. How can we talk about influencers if we cannot mention @realDonaldTrump. We should think about whether it makes sense to publish this paper under this condition, and thinking about withdrawing the paper now, before we keep working on the rebuttal letter. Also, the fact that they expect us not to release the data does not make sense. Perhaps, we should discuss it personally with the editor. Jamie Horder is quite reasonable, and there is precedent in our previous NatComm.
* **Alex**: It seems that they are very worried about repercussions... I don't know if it repercussions from the holders of the Twitter accounts or if it is from Twitter itself. There may be an argument to be made that accounts with more than X followers are public figures. He mentioned that it was a different policy than for our NatComm paper. We should discuss with them to better understand what the problem is.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.   
  
If you wish to submit a suitably revised manuscript we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.   
  
With your revision, please:  
  
• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.  
  
• Highlight all changes made to your manuscript or provide us with a version that tracks changes.   
  
  
REVIEWER COMMENTS:   
  
Reviewer #1:  
Remarks to the Author:  
LITTLE THINGS  
REPLICATION DATA  
The links the authors provide in the section “data and code availability” do not resolve to full replication data.   
  
FIGURES  
Many of numbered figure boxes in fact contain two or three. Several 1 and 5 will be unreadable. Make a pass through and label figures correctly, separate numbering, raise in resolution, and decide which ones to demote to appendix to strengthen the punchlines.  
  
SAMPLING  
Twitter data has known sampling issues that need to be acknowledged here. The probability of non-response from API queries is not provided by Twitter. And They have acknowledged that the 100% firehose is not actually a 100% sample, the 10% is not a randomly distributed 10%, the 1% is not randomly distributed 1%, etc but declined further details on sampling. Indeed a growing number of researchers moving away from Twitter data because the usual sampling details we taught to provide in Stats 101 class are not provided by the company. Need acknowledgement, even if in footnote, of the sampling parameters we do know—and the ones we don’t. Authors could look here for the professional norms and guidance on this: <https://www.aapor.org/Standards-Ethics/Standard-Definitions->(1).aspx. Even getting as close as possible to this is due diligence.  
  
HIGHLY AUTOMATED ACCOUNTS  
There doesn’t appear to be an effort to deal with the problem of highly automated accounts. Certainly we no longer focussing on “bots” the way we once did and now have to use the clumsy phrase “highly automated accounts”. But some of these accounts would be highly automated, might be bots, and the organizational ones almost certainly use some forms of automation. If the authors don’t feel the need to clean for such accounts—other researchers often do—need to explain why not.

* **James**: In the first subsection of the results, we mention that we use Tweet client information to identify tweets from unofficial clients that could come from bots. While we initially used this for volume estimation, I believe we also filtered unofficial clients from all our influencer and polarization analyses. We should explicitly state it as part of our discussion on unofficial clients.

A FEW EXTRA NUMBERS  
“In contrast, fake and extremely biased news are sent mainly by influencers whose accounts are unverified or deleted, with deceptive profiles and much shorter life-span on Twitter than traditional media influencers.” Please provide some descriptive stats for this – what proportion of the accounts captured in 2016 were gone by 2020? What proportion of the accounts capturd in 2020 didn’t exist in 2016?  
  
FIGURE 3 – I think there are 40 chosen across the years, not 30.  
  
I am not confident that allsides and mediabias are “widely used”, or the best choices (<https://en.wikipedia.org/wiki/Media_Bias/Fact_Check>#Criticism) I think best to admit the criticism and affirm that there is no better—perhaps have a colleague in Communication or Journalism research review and offer suggestions.

* **James**: In our methodology outlining our use of AS and MBFC, we reference multiple sources that implement this approach to identifying news categories and misinformation. However, these reviewers seem hesitant to accept even this as validation. I suppose the best approach is the end this section on methodology with a quick overview of the criticism aimed toward this approach to “admit the criticism and affirm there is no better”, unless we have some additional validation to reinforce this part of our work.

BIGGER THINGS  
  
MAGNITUDE  
One of the most interesting findings in the paper is expressed early on page 7, in simple percentages, and then the narrative moves on. “The fraction of tweets in the fake and extremely biased category, representing outlets that were most susceptible to sharing disinformation, decreased from 10% to 6% for fake news and from 13% to 6% for extremely right-bias news. The number of users who shared those tweets also decreased for extremely right-biased news (from 6% to 3%) but not for fake news (which remained at 3%) (SM Tab. S3). The fraction of tweets in the extremely-left bias category is very small (2% in 2016 and even less, 0.05% in 2020).” This is a top level, good finding! Fake, biased, extreme right-bias has halved, and the number of users sharing that extremist content also declined! This requires a little more explanation and needs to be in abstract and conclusion as an outcome. It’s a little misleading to simply say that extremist Twitter influencers and  
audiences are more polarized 2016-2020. The finding is actually that extremist Twitter influences and audiences shrank significantly, but became more polarized. If I have this wrong the authors need to do the interpretive work here, because these simple figures suggest the magnitude of the problem has greatly decreased.  
  
  
FIGURE 2  
This figure is easily my favorite, and I wonder if a few tweaks will really bring out the punch. First, by indexing the two years you give up the opportunity to show reader how the population size changed from 2016 to 2020. So instead of fraction from each year do N of each year. Distribute image across full page width need to give 2020 the width of the page. If you don’t want to do this, the N should go into the figure note so the reader can envision. Second, try to position “centre” in 2016 in the centre of the graph and page. Third be sure to check your label capitalization throughout for consistency if “Leaning” is capitalized “Extreme Bias” also please. The takeway here is that the right has become more extremist, and a signicant chunk of the centre is now leaning left, and this block is on the whole much larger than If the image was centred along the narrow strip of centrists that remained centre  
  
USE OF FAR LEFT  
The use of left and extreme left needs to be clarified, and probably changed, in this piece. If indeed extreme left is miniscule (authors say 0.05%) then it is inappropriate to treat that category in the same way as others.  
In figure 2, the labelling obscures some decision about how to include these sites. Perhaps “extreme left” has been included in left? Or perhaps dropped? Same with Figure 5.  
In figure 3, it is very inappropriate to then give this equivalence with the other categories. Choosing the top five users from each category gives the categories parity, greatly exaggerating the role of extreme left and greatly diminishing the impact of the other categories. This also results in a misleading dataviz. A sensible weighting would probably transform the “poles” significantly. For a sample of 40 prominent influences at most one of them would be far left in 2016 and none in 2020 so most researchers would probably decide to drop the category altogether rather than give it equivalence with other categories!  
Similarly with Figure 4. (BTW some of these category labels are different from earlier figures, be sure to always use same or offer new definitions with each figure, and avoid shortcode as labels use full words.) Indeed this figure may be one of the expendable ones, because it hides the relative magnitude that figure 2 communicates. You could use bar width to represent relative proportions. But again Extreme Left would be invisible on the page unless you wanted to log.  
In sum, there are two concept validity issues here: the extreme left category has been greatly inflated in importance and the extreme right category artificially diminished: the extreme left category may not even be large enough to pass a reasonable expectation of statistical significance and so perhaps after first mention should be dropped from further analysis.  
  
CONCEPT USE 1: IDEOLOGY  
Most researchers don’t use right and left in political research, it is liberal and conservative if referring to ideology. Perhaps right and left leaning parties if comparing across countries. I think this is an artifact of using the media bias labels produced by freelancers researchers. Is there a way to transform the categories or apply the more appropriate concepts?  
  
CONCEPT USE 2: INFLUENCER  
The labeled media affiliated influencers isn't quite right because the accounts generating the vast majority of content in this sample are just the base-level organizational accounts of news sources. For example the Twitter handle of CNN or the Reuters is probably a “source”—not an “influencer”. These would be organizational accounts managed by many people that push out everything the organization produces. Some of the other individually curated accounts probably are influencers , but are rarely themselves sources of fresh content. Indeed that explains why so much content from professional news organizations circulate. And it doesn't make sense to speak of change in affiliation with organizations. The CNN Twitter handle isn't going to become more or less affiliated with CNN between 2016 and 2020. If the user accounts of individual journalists were in this sample—stripping out based Twitter organizational accounts—then you would have comparable user types. I think  
this either needs to be about types of media organizations (strip out influencers and individual journalists) or strip out news organizations and just do individual pundits, journalists and other influencers.

* **Alex**: They would like us to choose between organizations (e.g. CNN) and individuals for the influencers. I don't think this is necessary, but we should maybe reframe the context of "influencers" as "influential accounts", so they can be of either type. They say that organizations are rather "sources" than "influencers”, but both are important here, whether they're one person or an organization.

Reviewer #2:  
Remarks to the Author:  
Key results: Please summarise what you consider to be the outstanding features of the work.  
  
This paper presents an analysis of changes in composition, popularity, and polarization of Twitter influencers and their retweeters in the months leading up to the 2016 and 2020 US presidential elections.   
  
The data used in the analysis consists of all tweets from the months of Jun-Nov of each election year that mentioned at least one of the presidential candidates and contained a url to a site that is tracked by [allsides.com](http://allsides.com) (AS) or [mediabiasfactcheck.com](http://mediabiasfactcheck.com) (MBFC). This method yielded a corpus of 30.7M tweets from 2.3M users in 2016 and 72.7M tweets from 3.7M users in 2020. An expanded set of tweets that are not limited to links to AS and MBFC tracked sites is used for the analysis of polarization of Twitter users. The authors define “news media” to the set of sites that are tracked by AS and MBFC. The Collective Influence (CI) algorithm is used to identify “news media influencers” – Twitter accounts that play the most influential role in spreading links to news media content.  
  
The contribution of the work is a longitudinal descriptive analysis that surfaces system-wide shifts in the Twitter data related to ideological polarization. The main findings of the study are:  
  
Among users that are active across the 2016-2020 span, a major shift towards retweeting more left leaning media content (Figure 2)  
A significant increase in polarization and fragmentation between left and right leaning media sources (Figure 3)  
A large turnover of the specific accounts that are most influential in spreading political media from 2016 to 2020 (Figures 4 & 5)  
Increased polarization of media sharing by Twitter users (Figure 6)  
  
As far as I am aware, this kind of longitudinal analysis of Twitter network change is rare and important, and certainly the focus on two pivotal US presidential elections makes it even more important.   
  
Validity: Does the manuscript have flaws which should prohibit its publication? If so, please provide details.  
  
I did not see any methodological issues that would invalidate specific descriptive results. My concerns are primarily in the lack of sufficient context to know how to interpret the significance of the various findings as detailed elsewhere in this review.  
  
Originality and significance: What are the major claims of the paper? Do you think that they represent a significant advance in the field? If the conclusions are not original, please provide relevant references. On a more subjective note, do you feel that the results presented are of immediate interest to many people in your own discipline, and/or to people from several disciplines?  
  
The longitudinal comparison across 2016 and 2020 is of great interest given the dramatic shifts in the American political landscape and the undeniable importance of social media in general, and Twitter in specific given that it was President Trump’s platform of choice. I feel that with better contextualization, some elements of this analysis might be of great interest across many disciplines (e.g., computational social science, political communication / science, media studies).  
  
Data & methodology: Please comment on the validity of the approach, quality of the data and quality of presentation. Please note that we expect our reviewers to review all data, including any extended data and supplementary information. Is the reporting of data and methodology sufficiently detailed and transparent to enable reproducing the results?  
  
I was not previously familiar with CI and worry about the fact that it ignores weights on edges. Why not just use PageRank? Can the authors better justify their choice of the CI algorithm?

* **James**: We already show in supplemental materials (directly referenced in the manuscript) that the influencers extracted via CI and the via PageRank are quite similar. This is done visually and using the metric ranked biased overlap. I believe we should stick with CI because the work that we are comparing against in 2016 used CI, and the comparisons are critical. Furthermore, Alex’s 2016 work also establishes that the characteristics of the retweet networks have the same patterns in both the weighted and unweighted versions, further justifying our use of CI. I don’t think we need to change anything here, other than maybe doubling down and being more explicit on our justification of CI.

As far as I am aware, filtering based on client is not sufficient to sort humans from bots. The most common thing bots do is use Twitter’s web API. Those bots appear like a person using the Twitter webpage. It does not seem that the present analysis accounts for this kind of (most prevalent) form of bot.

* **James**: this is a relevant point, but I don’t know if it would be feasible to deploy even more bot detection methods on such a massive dataset as ours. Most ML and NLP bot-detection models that I have used previously would not be able to handle it, so I think using the built-in client data from the API is still the best approach we have with the sheer size of the data. Unless anyone else has any suggestions?

Conclusions: Do you find that the conclusions and data interpretation are robust, valid and reliable?  
  
In general, as I read this manuscript, I kept trying to understand whether the specific quantitative findings were surprising or revealing beyond what is known from other sources about ideological polarization in the US (not just on Twitter), and how changes in news media leaning (which is noted but not sufficiently addressed) or changes in the make up of Twitter users (not addressed at all) shape the results of this work. Specifics below.  
  
The shift shown in Figure 2 raises a number of questions:  
How much of this shift is due to shifts in the classification of media sources? If the users in 2016 did not change their retweeting behavior at all, but the media sources they retweeted did change political lean, what would this figure look like?  
What is the political distribution of Twitter users that were active in 2016 but became inactive in 2020? What is the political distribution of Twitter users that were inactive or not on the platform in 2016 but became active in 2020? This would help us understand how to interpret the massive shift towards the left. Perhaps right leaning users left the platform, and meanwhile more left leaning users were attracted to Twitter? Or was there an actual shift of the Twitter user based towards the left as this figure suggests?  
Did the political balance overall of Twitter uses shift from 2016 to 2020 (Pew and other research orgs track this)? Is this overall shift (if any) partly to account for the analysis findings?  
  
How are we to interpret the big drop in tweets with media links from 18% to 10%. To what degree is this due to new media sites that were created or gained traction between 2016 and 2020 that AS and MBFC did not manage to track? The definition of “news media” is being tied to these two services which might be problematic. For example, if AS or MBFC had budget limitations, they may simply have chosen not to add certain sources to their tracking because of resource limitations, not because they did not consider the source a news media source.  
  
The authors report that 58% of the two 25 influential users in 2020 were not part of the top 100 from 2016. This seems to suggest high turnover, but should we be surprised by this specific quantitative result? Is 58% high? Compared to what? Is the turn over of influencers in other domains (e.g., music, sports) on Twitter over this same period? We need some kind of benchmark.  
  
The fact that there is increased polarization from 2016 to 2020 (visualized in Fig 3, and quantified using modularity analysis) is to be expected given all the other indicators about rising polarization more generally. It would be helpful to interpret the significance of these specific results within Twitter with more contextualization. How does the degree of increased polarization compare against other indicators of polarization that have been extensively studied by political scientists in 2016 and 2020? How does the degree of media polarization measured by [allsides.com](http://allsides.com) compare to what the authors find on Twitter? Is Twitter polarizing at the same rate as other parts of our communication environment or is it more extreme, akin to the distorting effect of social media noted in other studies (e.g., Chris Bail, Breaking the Social Media Prism)?  
  
Suggested improvements: Please list additional analyses, experiments or data that could help strengthening the work in a revision.  
  
The questions raised above point towards the various ways in which the authors might put their findings into context for a much more impactful study.  
  
References: Does this manuscript reference previous literature appropriately? If not, what references should be included or excluded?  
  
The authors should cite prior work by Benker et al (Yochai Benkler, Robert Faris, and Hal Roberts (2018). Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.) which reports similar findings of political fragmentation (they analyze both Twitter sharing patterns, and citation patterns between news organizations and find similarities across both methods). There is significant overlap in the methods and findings of this work, I believe, and provides a good foundation for comparison of the current study’s 2016 results, and of course the current work then makes an important contribution by contrasting to the situation in 2020.

* **Alex**: comparing and contextualizing the findings may be a bit tricky, but that is a good point. They give us a work to compare with (Yochai Benkler, Robert Faris, and Hal Roberts (2018). Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.), so that is a starting point.

Clarity and context: Is the abstract clear, accessible? Are abstract, introduction and conclusions appropriate?  
  
An expository problem with the paper is that it reads somewhat like a laundry list of findings. Perhaps a focus on the change in polarization of news media influencers and of Twitter users, analyzed against a background of Twitter-external ideological polarization survey data would create a more focused and impactful paper?  
  
Please indicate any particular part of the manuscript, data, or analyses that you feel is outside the scope of your expertise, or that you were unable to assess fully.  
  
I did not conduct a detailed review of the latent ideological estimation method but trust that it is sound given previous work cited by the authors that use this method.  
  
  
  
Reviewer #3:  
Remarks to the Author:  
The manuscript offers an interesting and informative examination of Twitter influencers in the political conversation around the 2016 and 2020 presidential elections in the US. As a social media platform heavily used by political actors, media sources, and activists, Twitter is worth exploring to understand the information and communication flows it facilitates.  
  
While I enjoyed reading the manuscript, there are some challenges that make it difficult to evaluate the results presented in this work. One key concern I had was related to the fact that data was split into categories of varying relevance to the overall political conversation on Twitter, which categories were then treated as equal parts of the discourse. Another related issue was the subsetting of the dataset for various analyses in ways that would clearly affect the results, but which were not explained or justified.  
  
(1) Media/influencer sample  
  
The paper offers a lot of details, but its main focus seems to be on examining the influencers who shaped the 2016 and 2020 flows of information into Twitter. In this case, those are the actors best able to spread news articles (via links to news sites) on the platform. That is an important group to examine (though the authors may want to note explicitly that the analysis would not capture any influencers who spread opinions/information that does not involve any links to external sources).  
To identify those influencers, the authors focus on a relatively short list of media sources. The sources are selected from an initial list of news domains (not sure how many or what was included?) and split into 8 categories based on their political leaning and quality. News outlets that got less than 1% of tweets received by the top outlet in their category were discarded. This means we are retaining outlets in some categories (e.g. “extreme left-leaning”) which are less popular than outlets we have discarded in other categories (e.g. “left-leaning”).  
The study then identifies the key influencers \*within each category\* and retains the top 25 per category. It would be helpful to know what the distribution of retweets or influence scores looked like and why 25 (or, for other analyses, 30) would be a good cut-off. Either way, based on the outlet size imbalance and examining the influencer names, it is clear that some of them are extremely prominent, while others are very minor participants in the general political conversation. For instance, the lowest scoring 2020 influencer in “extreme left” (@amberofmanyhats) seems to currently have 106 followers, while the lowest scoring influencer in “center news”, VP Kamala Harris has about 20 million followers.  
You can see where I’m going with this – by selecting top outlets/users by category, we’re including some very marginal accounts that were not a key part of the overall conversation, while excluding others which are much more influential but fall in a popular category. So, when all follow-up analyses look at polarization, clusters, “echo chambers”, shifting types of influencers across categories, etc. they are always working with this somewhat weird influencer sample. This makes it difficult to know whether, as the authors say, “the news propagated by Twitter [is] less influenced by traditional media and political organizations”, or whether that finding is an artifact of the categorization scheme that was used. I am, for example, not sure that media are losing their majority influencer status on Twitter, as the authors suggest on p.18.  
A better way to examine the global Twitter conversation may be to identify top influencers overall, then categorize those key actors into partisan and quality categories and work with that group to examine how polarized the conversation gets, or how skewed to one side or the other. If there are no high-level influencers to the extreme left in this group, for instance, that’s fine – it would just be a reflection of what the Twitter political world looks like.

* **Alex**: there is a good point (also made by Reviewer 1 about the far left) that because of our methodology, an important influencer from a not important media category is actually not important compared to influencers of other categories. This was done on purpose because we wanted to treat each category fairly for some of the analysis. But it's true that this gives an outsized importance to the far left for example in some of the analysis. We would need to make this clear and we would possibly need to compare our influencer list with another classification of influencers based on the entire retweet network.
* **James**: see my comment on the editor’s notes (page 1).

(2) For many analyses, the study uses a subsample of all data – for example, p.8 “users that are present in both election datasets”. To help the readers understand what that looks like, some brief descriptives of these samples would be helpful – if nothing else, what proportions of the total users are included here?  
  
(3) The paper discusses the polarization of the influencer retweet network (p.11) based on the position of the clusters after visualizing the network with a force-directed algorithm. While the visualizations are helpful, I would suggest explicitly describing the over-time change -- for instance by using modularity scores as the authors do later for other networks.  
  
(4) There are a few points that would be helpful to clarify about the similarity network used to examine polarization. Was the similarity score computed based only on retweets from other influencers, or from all Twitter users (the latter would be preferable)? Was a weighted version of modularity computed? The difference in score between the two networks seems rather small (.39 vs .37). It would be helpful to report the percent within-community edges for 2020 as well (by the way, how are those percentages of within/across edges calculated with weighted data – zero vs. non-zero ties?).  
  
(5) The authors make strong generalizations from their results – e.g. on p.24:  
“We confirm a change in the behavior of Twitter users when selecting to whom to retweet influencers’ content, that subsequently increased the division of influencers and users into opposing echo chambers from one election to the next.”   
It seems important to differentiate (where possible) changes actually resulting from the way users consume/share information on Twitter vs. changes driven by external factors. For instance, a lot of results are driven by the reclassification of media sources which were considered centrist in 2016 to be left-leaning (notably CNN and Politico). This is not something that stems from changes in Twitter user behavior, it is exogenous to the platform. Similarly, shifts in the “fake news” influencers seem likely to be driven by Twitter shutting down some sources and others disappearing from the internet. So, I would also doubt the suggestion that we’re seeing “a shift in the sources of disinformation from informal to formal organizations”.  
  
(6) A few smaller things:  
- On p.2 the authors mention that Twitter users are embedded in a relatively stable network of retweets – how stable is it really? Not crucial to this study perhaps, but an interesting data point nonetheless.

- On p.5: “The 2020 dataset contains 702 million tweets sent by 20 million users. Hence, we observe a significant increase in Twitter involvement in distributing election polarization, since in four years the number of Twitter users nearly doubled and number of tweets per user more than doubled, increasing the total number of tweets more than fourfold.”. One, I think perhaps the authors mean election ‘information’ rather than ‘polarization’. Two, I wonder if the growth is on par with the growth of the platform itself, or whether there is a disproportionate increase in the participation around politics.

* **Alex**: I don't remember that we wrote that in the manuscript, but this is also due to the fact that Zenkhun collected at a higher frequency in 2020 than I did in 2016. I would not compare absolute numbers between 2016 and 2020 because we cannot really control the difference in collections with Twiiter API. Reviewer 1 also asked that we mention the fact that we do not truly know if Twitter is truly random. We should state this as a limitation and probably remove comparisons of absolute numbers.

- On p. 11 “Node size is proportional to its out-degree in the complete combined network, i.e., the number of different users that have retweeted the node at least once” – should that be in-degree?  
  
  
  
Reviewer #4:  
Remarks to the Author:  
The manuscript explores a massive collection of twitter data with presidential candidate names in the runup to the 2016 and 2020 elections. Relying on a couple of websites’ categorizations of media organizations, the authors find shifts between the elections in twitter users greatest number of tweets from one category to the next, including: center and left leaners moving to left leaning, and fake and extremely biased moving to the right category. Subsequently, algorithms find that increased separation in influencers from 2020, and growth in right-biased news. Coupled with the manual coding of users into media vs political groups they find that unaffiliated are more common in the extremes and fake categories, and, surprisingly, that media-affiliated have moved right and fake.   
  
There is much to like about this paper---the impressive data collection, in particular, and the importance of the topic of social media polarization. Though I don’t think the manuscript meets the bar for publication in its current form, I hope the comments and criticisms that follow will be helpful to the authors.   
  
While the authors cite some of the current literature on polarization, they quickly drop any sophisticated engagement with it for media website ([allsides.com](http://allsides.com) and [mediabiasfactcheck.com](http://mediabiasfactcheck.com)) categories. I appreciate the links to the discussion of their methodologies, but should I trust this at all? ---or over the myriad of peer-reviewed categorizations of ideology? And which polarization, if any, can this study speak to if it relies on this extremely broad categorization for these organizations?   
  
Moreover, does it make sense to reverse engineer user ideology like this? Placing twitter users into the categories of which organization material they are (re)tweeting the most seems wrong at least some of the time. Many of us don’t agree with what we tweet: e.g., RT /= support. Doesn’t this assume that disinfo orgs cannot give good info and vice versa – and that right leaning cannot provide left leaning and vice versa. Moreover, that a conservative user couldn’t retweet a far left organization? My feed’s filled with users mocking extreme news sites with retweets.

* **Hernan**: I think that we already addressed this point with the Latent Ideology Analysis. In fact we say: “This high correlation indicates that the shift in bias observed at the level of the media outlets is also present at the level of the users’ retweeting pattern and serves as an independent validation of the media outlet classification.” Perhaps we can elaborate on this issue.
* **Alex**: I think this is a different issue. This is about whether a retweet implies endorsement of the original tweet. We actually addressed it in the manuscript by saying that as users can aslo quote, we consider that pure retweet without commenting means endorsement. There is no way to know if someone retweets ironically, but I think we can say that on average the retweets should be a good indication of the ideology.
* **James**: Aside from our discussion on quotes versus retweets, we also directly mention at the beginning of the paper that the 2015 study by Metaxas et al. found that retweeting can be considered an endorsement. Of course, there will be some retweets that could be considered ironic or mocking, but to identify those, we would need a complex user-centric model for flagging these for every user, something well beyond the scope of this project, and mainly unwarranted given that we have past works supporting our assumption. If anything, we could just back up our Metaxas reference with more references and explicitly address the fact that we know there are outliers, but on average the endorsement aspect of retweets holds, especially for such a large dataset.

Relatedly, in the lit review, polarization seems to be misrepresented or misunderstood, for example: partisan polarization is not limited to negative partisanship; there is overlap in ideological polarization and policy based…etc. In general, a deeper dive into the literature and appropriate citations for each category (not just a few general pieces at the start of the paragraph) would demonstrate substantive engagement and help to clarify any potential contribution.   
  
I similarly find the political vs media organization distinction to be atheoretical, subjective and without a link to the literature. For example, direct correspondence with a media outlet would put most of my colleagues into that category at those times, though I wouldn’t classify them as media organizations.   
  
Of the analyses, I enjoyed most the section on the change in top influencers (around figure 5). It was kind of neat to look at ponder the changes in some of these big media groups, interest groups and celebrities over time. Unfortunately, from a substantive contribution angle is that we don’t have any idea from these analyses what’s being tweeted about and the direction of the tweets. Is this policy, party, gossip, flash-in-the-pan events…etc.? With such a big category as presidential names, we’re left to wonder what’s caused the change: growth in twitter among subgroups, bans, real drop in reliance on traditional media, drop in reliance on traditional media on some topics or issues…etc.?   
  
Can the authors speak to the generalizability of their sample? Are 100 twitter influencers representative of trends in polarization? Is this just elite polarization, which we already have a great deal of evidence for?   
  
I much preferred the political actor measure of ideology used in the last section on top influencers. In general, I found this section more methodologically promising. Thus, expanding this to the previous analyses, though nontrivial, could be helpful. Alternatively, writing a full piece with this approach and focusing on the top influencers has the potential for a focused substantive contribution that is lacking in the previous sections. Of course, that paper would require a big reframe and deeper theoretical engagement.   
  
Ultimately, I found this to be a neat though inconsistent collection of results and computationally intensive analyses. What I found most lacking was any sort of clarity on the theoretical mechanism of those results. Yes, polarization is on the rise among elites and the public in various forms, but that hardly speaks to this inconsistent collection of findings. Do we learn anything novel here? Again, there were a bunch of interesting findings but just post hoc guesses as the causes (and effects), not to mention a number of caveats on changes in the platform and samples. The work has great potential but is yet both methodologically and substantively unfocused.

* **Alex**: This is the trap we are in I guess. We are limited to claims that are supported by the results. But the reviewer suggested some points where we could improve the political/social science theoretical framework.