Reviewers' comments:

Reviewer #1 (Remarks to the Author):

This work addresses an important topic of migration flux in Italy, and authors analyze public discourse around this subject. They are explicitly monitoring the content popularity and the role of social bots for the conversation.

Definition of interactions with verified users in the paper is not clear. Do you consider only retweeting activity or also consider mentions and replies? Depending on the conversation, different mechanisms can be preferred, such as retweets can be used to support opinions of verified accounts, while replies or mentions can be used to discuss or argue. Especially in politics, both of these means of interaction are important to analyze interactions and network structure.

Authors present a simple bot detection method to complement with their entropy-based network analysis. They employ null-models to extract signals from conversations by eliminating base-level of activity. Most of the analysis presented in this paper depends strongly on the network construction procedures and the methodology to filter those edges. What types of interactions are relevant to connect accounts and posts in the bipartite graph? Do you obtain different outcomes if you incorporate mentions and replies to the network?

Differences between histograms in Fig3. and Fig5. should be quantified by using statistical tests to make any claim about the changes of the fraction being significant.

How does the "contagion of polarization" method you described differently than label propagation?

Some minor notes:
- Twitter's documentation use term "retweeting" instead of "retwitting"
- SI Fig1 can lead to more information of authors to try different layout algorithm and wait for convergence.
- I suggest changes in colors in Fig1. Authors use 'tomato red' to represent the Italian Democratic Party. To have a coherent naming scheme, I also suggest using 'basil green', 'eggplant purple', and 'Mediterranean blue' :)

Reviewer #2 (Remarks to the Author):

This is an interesting manuscript that mixes machine learning methods for bot detection to network models drawn from statistical physics to study the impact of social bots in political conversations. The authors study the Twitter conversations around the migrant crisis in Italy and find evidence of groups of bots amplifying the messages of a few verified users, notably some prominent italian right-wing politicians.

One major area of potential improvement is in the presentation, which is weak. Another concern is about the overall novelty of the contribution. In the following I give more detailed feedback:

- Right now the manuscript is really hard to read and structured in a very confusing manner. For example, throughout the manuscript the authors talk about various "validated" entities (as in "validated links", "validated followers", etc.). It turns out that the "validated" network is the network obtained from filtering the data with the statistical model. Given the importance of this step, one would assume that the authors devoted some space to describing it. This is not the case, unfortunately, as nowhere in the main text the validation step is really spelled out in detail. Instead, the authors mention the term in passing in a small footnote at the beginning of the document, and scatter the rest of the details throughout the text and in the supplementary

material. For example, in the supplementary it is said that the filtering is done via multiple hypothesis testing, using the FDR technique.
Footnotes are easy to miss, so first of all I would move that footnote in the main text. As for the scattered details, I would suggest the authors devoted its own section to the validation step, separate from the description of the statistical model, and gave more information about it. (For example, it would be interesting to see a plot of the p-values used in the multiple hypothesis testing).
- Aside from improving the particular issue with the "validation", I would also suggest the manuscript to be edited for clarity and structure. There are several typos (e.g. "retwitters" -> "retweeters"), undefined terms (e.g. "grey literature"), and long, hard to read sentences (an example of a long sentence is the one that starts with "Methodologically," in section 4.3). But, aside from those, I think that the structure of the manuscript should be completely overhauled. For example, section 2.1 should be moved in the Methods section, Finally, often the language is not crisp enough, with vague passages that engender confusion and potential misunderstanding on the part of the reader (for example, the sentence: "Smaller communities regards Malta prime [sic!] Minister [...]" -> here it sounds like the PM is part of multiple communities, which cannot be the case, given the use of Louvain. The authors probably meant that among the smaller communities, there is one that includes the Maltese PM.)
- As a separate issue with presentation, the authors could be more kind to the reader when referencing the cited literature. Right now, there are multiple sentences where the numerical references are used as parts of speech, e.g. "In [33], by implementing the method of [66],...". (This one stands out as one of the worst offenders but several other examples can be found throughout the text.) This is a sloppy writing practice that leads to text that is ugly to read and hard to understand. For one, it forces the reader to interrupt the flow and go in the bibliography to actually lookup the citation, to understand what the authors are talking about. A much better alternative would be to include the name of the first cited author, as in: "By implementing the method of Saracco et al. [66], Becatti et al. [33] used this feature ...". Readers who have already encountered before those particular works will not need to interrupt their reading flow; those who have not will be able to associate a name to the number, which is easier to remember.
- Methods section should be rewritten as multiple independent sections; the main text should provide more obvious pointers to the actual sections. Right now it is not immediately clear where to go when the authors refer the reader to it.
- Data availability statement is not acceptable, it does not state why the authors cannot make the data available by default. I note that there are ways to share tweets that respect both user privacy and Twitter terms (e.g. sharing only tweet/user IDs + bot labels).
- Code is missing, making the results hard to replicate and inspect.
- The main finding is that some hubs are being retweeted by the same bot accounts. This why I presume the term "bot squad" is used, which is otherwise never properly defined anywhere else. It would be interesting to know more about these squads, for example show an example of their re-tweeting activity triggered by the hubs.
- Related to the previous point, one of the main claims is that the presence of "bot squads, retweetting [sic] the messages of two or more strong hubs, increases the visibility of their tweets". This is an interesting finding, however the authors do not show any data to back it up. I would assume that they meant that bot squads increase the popularity over what would normally be expected from the popularity of the hubs alone. Otherwise the finding sounds a bit trivial (more users -> more tweets).
- Supplementary materials: in the expression for the conditional probability of a bi-graph given the constraints, the symbol a_{i\alpha} is never defined. I think the letter should be m, not a.
- In my opinion the novelty of the contribution is mostly in the methodological innovation of using bot detection with the bipartite configuration model, but as far as findings go I am not entirely sure there is a whole lot of novel or striking empirical observations. Perhaps the authors could defend more the novelty of their findings.

Overall I think that this is an important an interesting paper, which makes an important contribution to the area concerned with the study of d/misinformation on social media. I think that

the manuscript is promising, but in its current format it cannot be published, and I have tried to suggest many potential improvements. If these are addressed, I would be happy to recommend it for publication.

Reviewer #3 (Remarks to the Author):

The authors study a data set of messages from the microblogging platform Twitter. The messages considered here in the paper discuss the flow of migrants from Africa to Italy.

It is not perfectly clear what the purpose of this study is. Certainly, most of the conclusions drawn are about the role played by bots in driving such conversation. As stated in the abstract the main conclusions of the paper are: 1. The strongest hubs have a higher than expected number of bots among their followers. 2. There is at least one example of a couple of political leaders of the same political orientation that share a large number of bots among their followers. There is little, from the point of view of conclusions, about the flow of migrants or how it is perceived through the lens of Twitter. It is unclear therefore why this specific dataset has been chosen. If the meant topic was the "bots and politics" then certainly the dataset considered seem a very restrictive choice.

I am unsure about the fit of the topic with Communications Physics. The editor, no doubt, will have a much more fact-based opinion than mine.

The paper is sufficiently well written (with few caveats). Results are clearly stated and discussed. The introduction and methods section can, instead, be improved.
For example, the introduction starts with a summary of previous and related work about bots and misinformation on Twitter. Suddenly – it is almost a non-sequitur – the discussion turns to "entropy based null-models for the analysis of complex networks". It remains unclear why the reader needs that (at least for several pages). The author should declare as soon as possible what is the purpose of the paper and explain why one needs to know about the null-models. As a matter of fact, such "null models" are a methodological expedient that can be discussed elsewhere. A focus on matters related to the main conclusions of the paper should be mentioned in the introduction.

In several places, starting with the caption of Fig. 1, continuing in sec. 2.3.1, and then in the discussion section, one finds references to political parties, personalities, and their mutual relations relative. I found those parts uninformative (maybe it would be different for somebody well acquainted with Italian politics). Given the little context given, they don't serve neither the purpose of convincing the reader that the analysis performed is meaningful, nor serve the purpose of clarifying aspects of Italian politics or something about the flow of migrants. I believe that the ignorance of the typical reader on this subject will not be dissipated by reading these portions of the paper.

The paper is not well referenced. The authors are very generous in self-citations (more than 20 among the 60+ in the paper) but at the same time seem to ignore a lot of recent and relevant literature on bot detection, polarization, methods to establish the political orientation of users, and other areas germane to this paper.

The results, in my opinion are meager, and not really surprising or novel. I don't think they deserve the long section that is devoted to their discussion. A relatively large space is devoted to methodological issues, also if they are not novel. Apparently all major relevant "tools" have been introduced elsewhere, either by the same authors (the entropy based null-models) or by others (e.g. the approach to the hub-authority) analysis.

The paper is not well referenced. The authors are very generous in self-citations (more than 20 among the 60+ in the paper) but at the same time seem to ignore a lot of recent and relevant

literature on bot detection, polarization, methods to establish the political orientation of users, and other areas germane to this paper.

The procedures followed to collect, filter and analyze the data are discussed at length in the paper, but many of the assumption underlying the choices made remain obscure to me. In particular:
1. what are the benefits of considering "validated" accounts?
2. What is the random noise one "needs" to eliminate by using the entropy-based models? 3. Why, in adopting such null models, does one want to "discount" the virality of messages, the retweeting activity of users and their productivity in writing tweets? The fact that "the study is the first investigation that merges bot detection and entropy-based analysis" is not a merit in itself.

To conclude. The paper is, in my opinion, an odd mix of:
1. Facts about some specific topic in Italian news/politics, which don't seem to be the focal point of the paper
2. Discussion of methodological tools that are not novel and whose use is not sufficiently justified in the present context
3. and findings about that do not represent a big addition to the present literature on the subject.

Based on these considerations, I am afraid, I cannot recommend publication.

REBUTTAL LETTER

Dear Reviewers:

In this revised version of our manuscript we have accurately taken care of all your questions, as thoroughly described in the remainder of this document. We appreciated all the insightful comments we received in response to our submission. These useful suggestions and observations helped us improve and strengthen our manuscript. Please, find below your original comments, together with a thorough explanation of how we addressed them. In order to support the revision process, we have also highlighted in blue all the modified paragraphs in the revised manuscript.

Thank you in advance for your time and consideration.

Yours faithfully,

Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco

Reviewer #1

Summary:
This work addresses an important topic of migration flux in Italy, and authors analyze public discourse around this subject. They are explicitly monitoring the content popularity and the role of social bots for the conversation.

Comment 1:
Definition of interactions with verified users in the paper is not clear. Do you consider only retweeting activity or also consider mentions and replies? Depending on the conversation, different mechanisms can be preferred, such as retweets can be used to support opinions of verified accounts, while replies or mentions can be used to discuss or argue. Especially in politics, both of these means of interaction are important to analyze interactions and network structure.

Reply: Thanks to the reviewer's comment, we have now specified, in Section 2.1, that the interactions between the users of the bipartite network that we consider are retweets. In the same section, we motivate the choice, commenting on the nature of replies, mentions, and quoted tweets and on what would have meant to consider them in the analysis.

Comment 2:
Authors present a simple bot detection method to complement with their entropy-based network analysis.

Reply: We would like to better explain the work behind the definition of the bot detector used in this manuscript. Details about the bot detector can be found in 'S. Cresci, R. di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi: Fame for sale: Efficient detection of fake Twitter followers. Decision Support Systems, 80, 56-71 (2015)'. The bot detector uses all the features of an account

profile, and has been developed after testing known detection methodologies on a baseline dataset of fake and genuine accounts. In particular, the authors tested the Twitter accounts in the reference set against algorithms based on: (i) single classification rules proposed by Media and bloggers, and (ii) feature sets proposed in the literature for detecting spammers. The results of such preliminary analysis suggested that fake accounts detection needs specialized mechanisms. They classified rules and features according to the cost required for gathering the data needed to compute them and showed how the best performing features are also the most costly ones. Then, building on the cost of crawling analysis, they implemented a series of lightweight classifiers using less costly features, while still being able to correctly classify more than 95% of the accounts of the baseline dataset. They also validated performances of the classifiers over two other sets of human and fake accounts, disjoint from the original training dataset. The enriched description can be found in Section 4.2 of our manuscript.

Comment 3:
They employ null-models to extract signals from conversations by eliminating base-level of activity. Most of the analysis presented in this paper depends strongly on the network construction procedures and the methodology to filter those edges. What types of interactions are relevant to connect accounts and posts in the bipartite graph?

Reply: In the bipartite network between accounts and tweets, we consider exclusively all retweets and their original messages. So, if user A retweets tweet T originated by user B, then, in the bipartite directed network we have a link from B to T and a link from T to A.

In order to connect two users A and B in the validated projection, the number of retweets has to be much greater than expected by considering only the degree sequences of the nodes. More in details, in the directed monopartite projection, B links to A if A retweets an amount of tweets by B that cannot be explained by the expectation of the users activities, namely: 1) the writing and retweeting activities of, respectively, B and A and 2) the popularity (i.e., the number of retweets) of B's tweets. In the revised version of the manuscript, Section 2.2, we added a detailed description of the network construction. Actually, we are not considering the relevance of a single tweet, but the number of messages written by B that are shared by A: if this number is not compatible with the information contained in the degree sequence of both layers of the bipartite network, we draw a link in the projected network.

Comment 4:
Do you obtain different outcomes if you incorporate mentions and replies to the network?

Reply: Actually, retweets are intended to share messages one agrees with. Adding mentions and replies would not permit to have a clear interpretation of the measured overlaps; thus we avoided such analysis. We added more details about the choice of retweets as acts of endorsement in the revised Section 2.1.

Comment 5:
Differences between histograms in Fig 3. and Fig 5. should be quantified by using statistical tests to make any claim about the changes of the fraction being significant.

Reply: Please note that in the revised manuscript we added some figures. Thus, original Figures 3 and 5 are now Figures 4 and 11, respectively. We ran a Pearsons Chi Square test on the distribution of political tags on both the original and the validated network. Beside showing

several differences, the associated p-value associated is close to 1, thus the differences are not statistically significant. We mentioned this extra test in the caption of the Figure 11.

Comment 6:
How does the 'contagion of polarization' method you described differently than label propagation?

Reply: Thanks to the reviewer's doubt, we have added further details in Section 2.1.3. The contagion of polarization is indeed a label propagation procedure in which the labels are those obtained in the previous step. We hope to have better clarified how polarization have been given to unverified users which have no direct interactions with verified ones.

Comment 7:
Some minor notes: - Twitter's documentation use term retweeting instead of retwitting.

Reply: The term has been corrected throughout the manuscript.

Comment 8:
Fig. 1 can lead to more information of authors to try different layout algorithm and wait for convergence.

Reply: Please note that original Figure 1 is now Figure 2. Since the network is extremely dense, we tried several layouts in order to better represent the network: actually, all alternatives were not satisfactory, since they were not able to properly highlight the community structure of the system. We think that the original one is the most readable figure we have been able to obtain.

Comment 9:
I suggest changes in colors in Fig 1. Authors use 'tomato red' to represent the Italian Democratic Party. To have a coherent naming scheme, I also suggest using 'basil green', 'eggplant purple', and 'Mediterranean blue' :)

Reply: Done :) (Figure 1 → Figure 2)

Reviewer #2

Summary:
This is an interesting manuscript that mixes machine learning methods for bot detection to network models drawn from statistical physics to study the impact of social bots in political conversations. The authors study the Twitter conversations around the migrant crisis in Italy and find evidence of groups of bots amplifying the messages of a few verified users, notably some prominent italian right-wing politicians.

Comment 1:
One major area of potential improvement is in the presentation, which is weak. Another concern is about the overall novelty of the contribution. In the following I give more detailed feedback:

Right now the manuscript is really hard to read and structured in a very confusing manner. For example, throughout the manuscript the authors talk about various 'validated' entities (as in 'validated links', 'validated followers', etc.). It turns out that the 'validated' network is the network obtained from filtering the data with the statistical model. Given the importance of this step, one would assume that the authors devoted some space to describing it. This is not

the case, unfortunately, as nowhere in the main text the validation step is really spelled out in detail. Instead, the authors mention the term in passing in a small footnote at the beginning of the document, and scatter the rest of the details throughout the text and in the supplementary material. For example, in the supplementary it is said that the filtering is done via multiple hypothesis testing, using the FDR technique. Footnotes are easy to miss, so first of all I would move that footnote in the main text. As for the scattered details, I would suggest the authors devoted its own section to the validation step, separate from the description of the statistical model, and gave more information about it. (For example, it would be interesting to see a plot of the p-values used in the multiple hypothesis testing).

Reply: We thank the reviewer for the precious comments. In this revised version of the manuscript, we have clarified that throughout the document will be used the entropy-based null models, and motivated this choice. Specifically:

- The footnote in the introduction has been removed.
- The entire Results section has been reviewed. In Section 2.1, it is described how the model has been used to attach a polarization label to the verified accounts and (in the second to last paragraph) and the difference between the 'verified' and 'validated' terms is clarified .
- Section 2.2 now contains details about the application of the null model, in its 'directed' version (BiDCM), to obtain a 'validated' network where only statistically significant messages are exchanged among users.
- Sections 4.3 and 4.4. of Methods now provide technical details about the application of:
  (1) the BiCM model to link pairs of verified users who have a statistically significant number of common retweets with unverified users, and
  (2) the BiDCM model to extract the validated network where there are only those nodes that contribute in a statistically relevant way to the message flow.
- Section 4.3.2 in Methods discusses the application of FDR and the multiple test hypothesis.

Finally, we would like to say that p-values from different distributions cannot be compared (the approach of FDR is indeed different, since it considers every p-value on its own and then calculates the expected rate of False Positives). While we understand the rationale of the reviewers request - getting a deeper understanding of the p-values -, we think that the suggested plot could be misleading for the reader. Instead, we have added further details, about the ratio of the validated p-values and the effective threshold (see Section 4.4 of the main text).

Comment 2:
Aside from improving the particular issue with the 'validation', I would also suggest the manuscript to be edited for clarity and structure. There are several typos (e.g. retwitters → retweeters), undefined terms (e.g. grey literature), and long, hard to read sentences (an example of a long sentence is the one that starts with Methodologically, in section 4.3). But, aside from those, I think that the structure of the manuscript should be completely overhauled. For example, section 2.1 should be moved in the Methods section, Finally, often the language is not crisp enough, with vague passages that engender confusion and potential misunderstanding on the part of the reader (for example, the sentence: 'Smaller communities regards Malta prime [sic!] Minister [...]' → here it sounds like the PM is part of multiple communities, which cannot be the case, given the

use of Louvain. The authors probably meant that among the smaller communities, there is one that includes the Maltese PM.)

Reply: We did our best to follow the reviewer's instructions, correcting typos and streamlining the narrative. With the term 'Grey Literature', we meant classification rules for bot detector defined by online blogs. We have rephrased the sentence. Section 2.1 has been moved to the Methods section, as its first subsection. And yes, we meant that the prime minister is part of one of those communities. Thanks to the reviewer's comment we have reworked that part.

Comment 3:
As a separate issue with presentation, the authors could be more kind to the reader when referencing the cited literature. Right now, there are multiple sentences where the numerical references are used as parts of speech, e.g. "In [33], by implementing the method of [66],...". (This one stands out as one of the worst offenders but several other examples can be found throughout the text.) This is a sloppy writing practice that leads to text that is ugly to read and hard to understand. For one, it forces the reader to interrupt the flow and go in the bibliography to actually lookup the citation, to understand what the authors are talking about. A much better alternative would be to include the name of the first cited author, as in: "By implementing the method of Saracco et al. [66], Becatti et al. [33] used this feature ...". Readers who have already encountered before those particular works will not need to interrupt their reading flow; those who have not will be able to associate a name to the number, which is easier to remember.

Reply: We agree with the reviewer and we apologise for the inconvenience. In the revised version, we have avoided as much as possible to refer to past work without giving the name of the authors.

Comment 4:
Methods section should be rewritten as multiple independent sections; the main text should provide more obvious pointers to the actual sections. Right now it is not immediately clear where to go when the authors refer the reader to it.

Reply: The section has been reorganized, with appropriate cross-references. Moreover, we added to the main text more precise references to the subsections of the methods.

Comment 5:
Data availability statement is not acceptable, it does not state why the authors cannot make the data available by default. I note that there are ways to share tweets that respect both user privacy and Twitter terms (e.g. sharing only tweet/user IDs + bot labels)

Reply: We apologize to the reviewer for the miscommunication. Whether and when the article will be published, we will be happy to distribute the data, under the limits of Twitters Developer Terms. We have now modified the Data Availability Statement. Please note that the statement is not at present tense, as it would be the final version of the manuscript. We will load the data on the webpage as soon as the paper will be accepted for publication.

Comment 6:
Code is missing, making the results hard to replicate and inspect.

Reply: We apologize for not having released a code statement. However, we are working on releasing the bot classifier, as well as a Python package for filtering out the noise from the network. In view of this, we have added a Code Availability Statement in the manuscript.

The bot detector model is replicable because the training dataset is public (available from `http://mib.projects.iit.cnr.it/dataset.html`), and 'S. Cresci, R. di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi: Fame for sale: Efficient detection of fake Twitter followers. Decision Support Systems, 80, 56-71 (2015)' provides details about the used learning algorithms.

Comment 7:
The main finding is that some hubs are being retweeted by the same bot accounts. This why I presume the term "bot squad" is used, which is otherwise never properly defined anywhere else. It would be interesting to know more about these squads, for example show an example of their re-tweeting activity triggered by the hubs.

Reply: The term "squad" refers to those accounts that follow and retweet the same hubs. We have introduced its definition at the end of the Introduction.

Concerning the activities of the squads, at the end of the Results section, we have introduced an analysis about the number of tweets, retweets, mentions, and quoted tweets operated by the biggest bot squad. Also, we have shown the most retweeted hubs and the most mentioned accounts (please see brand new Figure 10).

Comment 8:
Related to the previous point, one of the main claims is that the presence of 'bot squads, retweeting [sic] the messages of two or more strong hubs, increases the visibility of their tweets'. This is an interesting finding, however, the authors do not show any data to back it up. I would assume that they meant that bot squads increase the popularity over what would normally be expected from the popularity of the hubs alone. Otherwise the finding sounds a bit trivial (more users → more tweets).

Reply: We absolutely agree with the reviewer that, if the contribution of the article was 'the more users, the more tweets', it would not be a real contribution at all. Our work confirms that automated accounts are used to retweet the messages of those who, already on their own, are able to achieve a great visibility; but it also suggests that the hubs and their bot squads join efforts to further increase the visibility of the hubs messages. To the best of our knowledge, the existence of formations of bots in common with different hubs has never been noticed in the literature before. We have expressed our viewpoint about this at the end of the introduction.

Comment 9:
Supplementary materials: in the expression for the conditional probability of a bi-graph given the constraints, the symbol $a_{i\alpha}$ is never defined. I think the letter should be m, not a.

Reply: Thanks to the reviewer for noticing it. We corrected the typo, indeed, it is m.

Comment 10:
In my opinion, the novelty of the contribution is mostly in the methodological innovation of using bot detection with the bipartite configuration model, but as far as findings go I am not entirely

sure there is a whole lot of novel or striking empirical observations. Perhaps the authors could defend more the novelty of their findings.

Reply: We agree that the application of the two methodologies together is innovative and it is able to highlight which genuine accounts and bots are more successful in conveying messages on Twitter. The revised version of the introduction describes more in detail why we have decided to use them jointly. In addition, we found bunches of bots attached to bunches of hubs. Although we cannot make a claim about the mind and the strategy behind this organization, we find the result noteworthy, and consider interesting to look for the phenomenon in other datasets; the new Discussion section considers this aspect.

Comment 11:
Overall I think that this is an important an interesting paper, which makes an important contribution to the area concerned with the study of d/misinformation on social media. I think that the manuscript is promising, but in its current format it cannot be published, and I have tried to suggest many potential improvements. If these are addressed, I would be happy to recommend it for publication.

Reply: We thank the reviewer for appreciating our work and for having given us detailed suggestions for improvements. We sincerely hope that the reviewer's doubts have been dispelled by the enriched descriptions in the revised version of the manuscript.

Reviewer #3

Summary:
The authors study a data set of messages from the microblogging platform Twitter. The messages considered here in the paper discuss the flow of migrants from Africa to Italy.

Comment 1:
It is not perfectly clear what the purpose of this study is. Certainly, most of the conclusions drawn are about the role played by bots in driving such conversation. As stated in the abstract the main conclusions of the paper are: 1. The strongest hubs have a higher than expected number of bots among their followers. 2. There is at least one example of a couple of political leaders of the same political orientation that share a large number of bots among their followers. There is little, from the point of view of conclusions, about the flow of migrants or how it is perceived through the lens of Twitter. It is unclear therefore why this specific dataset has been chosen. If the meant topic was the bots and politics then certainly the dataset considered seem a very restrictive choice.

Reply: We apologize with the reviewer because we have not entirely clarified the motivations and results of the article in its first version; we have now restructured abstract and introduction. In the revised Introduction, we have separated background and results. We have also added a paragraph to motivate the use of both a bot detector and a filtering technique based on entropy-based null models. Finally, we have clarified that, to the best of our knowledge, it is the first time that bunches of bots have been found attached to bunches of hubs. Although we cannot make any claim about the strategy behind this organization, we find the result worthy of dissemination.

Actually, the political subject of managing the migration flux captured the whole Italian political debate. On newscasts, newspapers and online news outlet, the discussion had a strong daily focus, due to the communication strategy of the Minister of Internal Affairs at that time. As

the number of collected tweets testifies, the subject of the analysis is not restrictive. The period under observation regards events that have taken place around the, first denied and then granted, landing of one ship operated by NGOs, rescuing migrants fleeing from North Africa to Italy. The fate of the migrants has caused a stir in public opinion, and has involved a lively debate between the political forces of the country. The boat called 'Diciotti' had 177 migrants on board rescued on August 16, 2018 off Lampedusa island. The Diciotti received a veto from the Italian government to bring down the migrants who were on board and Mr. Matteo Salvini, as Italian Deputy Prime Minister and Minister of the Interior, was investigated for kidnapping and abuse of office as well as other crimes. The case was stopped on February 19, 2019, when the Italian Senate decided not to trial Mr. Salvini. Straddling the judgement, there was a very intense debate on social networks, characterised by supporting, or criticising, the migrants, the NGOs, the Italian Government, and the European Union. Information about the analysed dataset has been now inserted in Section 4.1.

Comment 2:
The paper is sufficiently well written (with few caveats). Results are clearly stated and discussed. The introduction and methods section can, instead, be improved. For example, the introduction starts with a summary of previous and related work about bots and misinformation on Twitter. Suddenly it is almost a non-sequitur the discussion turns to entropy based null-models for the analysis of complex networks. It remains unclear why the reader needs that (at least for several pages). The author should declare as soon as possible what is the purpose of the paper and explain why one needs to know about the null-models. As a matter of fact, such null models are a methodological expedient that can be discussed elsewhere. A focus on matters related to the main conclusions of the paper should be mentioned in the introduction.

Reply: We thank the reviewer for her/his appreciation of the presentation of the results. We have reworked the introduction, that now ends with a description of the main results and motivates the joint use of bot detection techniques and network normalization based on entropy-based null models. In particular, we have clarified that although there is a wide literature on methods for bot detection, rarely it has been studied how effectively the messages conveyed by bots penetrate the network (in our case, in terms of retweets). In this work, we have applied an entropy-based null model to strip off data so-called noise, i.e. those activities not distinguishable from a purely random activity. This is important for both genuine accounts and bots.

Comment 3:
In several places, starting with the caption of Fig. 1, continuing in sec. 2.3.1, and then in the discussion section, one finds references to political parties, personalities, and their mutual relations relative. I found those parts uninformative (maybe it would be different for somebody well acquainted with Italian politics). Given the little context given, they dont serve neither the purpose of convincing the reader that the analysis performed is meaningful, nor serve the purpose of clarifying aspects of Italian politics or something about the flow of migrants. I believe that the ignorance of the typical reader on this subject will not be dissipated by reading these portions of the paper.

Reply: We have improved the description of the context, ensuring that the various characters, political parties, newspapers and the other mentioned entities are better introduced and contextualised. Specifically for parties and the government in office at the time the dataset was

collected, we have introduced a new Section 2.1.1. Moreover, we have added information to the caption of Figure 2, as well as to the footnotes related to that figure.

Comment 4:
The paper is not well referenced. The authors are very generous in self-citations (more than 20 among the 60+ in the paper) but at the same time seem to ignore a lot of recent and relevant literature on bot detection, polarization, methods to establish the political orientation of users, and other areas germane to this paper.

Reply: Thanks to the reviewer's comment, we have extended the literature review on bot detection with recent work (see final part of Section 4.2). We also added a discussion on prominent recent work concerned with users' polarization and their political orientation at the end of Section 2.1 (just before 2.1.1).

Comment 5:
The results, in my opinion are meager, and not really surprising or novel. I dont think they deserve the long section that is devoted to their discussion. A relatively large space is devoted to methodological issues, also if they are not novel. Apparently all major relevant tools have been introduced elsewhere, either by the same authors (the entropy based null-models) or by others (e.g. the approach to the hub-authority) analysis.

Reply: We hope that by providing a new contextualization and a new introduction we have better explained our results and their value. As for the methodology, although the two main methods - BiCM and BiDCM - were not developed in this article, their contextual use allowed not only to detect the hubs, but also to highlight how the same hubs interact with the same bots. We do not know the strategy pursued by the developers of bots for such behavior. In the text, we added a small paragraph highlighting the behaviour of the greatest bot squad. In particular, we showed that their main activity, as expected, is retweeting genuine accounts among the strongest hubs. Interestingly, when they send original tweets, in the 89% of the case, they share pieces of news from a website. In 97% of the case, the related URL is from voxnews.info, a website present in the black list of more than one fact-checking website as a source of political misinformation. We thank the referee for stimulating deeper investigations that brought us this findings.

Comment 6:
The procedures followed to collect, filter and analyze the data are discussed at length in the paper, but many of the assumption underlying the choices made remain obscure to me. In particular:

(1) what are the benefits of considering validated accounts?
(2) What is the random noise one needs to eliminate by using the entropy-based models?
(3) Why, in adopting such null models, does one want to discount the virality of messages, the retweeting activity of users and their productivity in writing tweets? The fact that "the study is the first investigation that merges bot detection and entropy-based analysis" is not a merit in itself.

Reply: To respond to the above three issues we have to say that:

(1) Validated accounts are those resulting from a (directed or undirected) projection. The resulting network focuses on non-trivial activities only, i.e., those not already encompassed by the degree sequence. Indeed both null models discount the degree sequence of nodes

belonging to both layers, to highlight those patterns that cannot be simply derived by degree sequences.

(2) The entropy-based null-model removes the noise of users acting randomly on the platform, i.e., without focusing their activity on specific targets, as it would be if they would systematically retweet the messages of specific accounts. The noise is automatically discounted by the entropy-based null-model under consideration: it maximises the "randomness" of links, fixing (on average) the degree sequence. In this sense, specific patterns that are not caused by the degree sequence are highlighted.

(3) The validated projection yields a monopartite directed network in which a (directed) link is present if the target node has retweeted the source node a number of times that is greater than expected by the BiDCM. Since the BiDCM discounts both the in- and out- degree sequence of nodes belonging to both layers, the resulting network focuses on the non-trivial sharing activity that cannot be explained by the degrees of the nodes. For example, knowing that user A retweets, say, 42 messages of user B is not sufficient to conclude that A is strongly following B. To infer this, we need to consider also the retweeting activity of A (i.e., its in-degree), the authorship activity of B (i.e., the out-degree) and the virality of messages shared by A (i.e., the out-degree of the tweets).

The benefits for considering validated accounts have now been explained in different parts of the manuscript: Introduction, Results, Methods.

In general we tried to better motivate our choices:

- Section 2.1 now specifies that users polarization has been obtained by discounting the degree sequence of the layers in the bipartite network of verified and unverified users accounts.
- Section 2.2 now specifies that the validated network has been obtained by discounting the number of original tweets posted by every user, the number of retweets of every tweet, and the number of retweets of every account.
- In Section 2.2 of the main text, and more in details, in Section 3.2.1 of the Supplementary Information, we discuss how to discount the contribution of viral tweets and avoid that a strongly retweeted tweet biases the projection and assigns a strong contribution to all its retweeters. The bias is eliminated by properly weighting virality. We discount also retweeting activities of the users to correctly weight 'retweet maniacs', i.e., users whose standard behaviour on the platform is retweeting everything.

Overall, using jointly, for the first time, two different methodologies (bot detection and filtering via entropy-based null models) has allowed us to make an in-depth analysis.

Comment 7:

To conclude, the paper is, in my opinion, an odd mix of:

- *Facts about some specific topic in Italian news/politics, which dont seem to be the focal point of the paper*
- *Discussion of methodological tools that are not novel and whose use is not sufficiently justified in the present context*
- *and findings about that do not represent a big addition to the present literature on the subject. Based on these considerations, I am afraid, I cannot recommend publication.*

Reply: We are very sorry to have conveyed this evaluation with our first submission. We hope that the new version clarifies the context and the motivations for applying the selected methodologies and that the novel contribution is more evident.

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

I want to start by thanking authors for extending their results and rewriting particular sections of the paper for better clarity. I have a few minor comments on the revised version, and I think authors can address them easily.

The distinction between verified and normal Twitter users is important for the analysis. There have been research studies on characterizing verified users on Twitter. I suggest a couple of works that study verified users and their characterizations as well as relations to social bot research.
- Varol, Onur, and Ismail Uluturk. "Journalists on Twitter: self-branding, audiences, and involvement of bots." Journal of Computational Social Science (2019): 1-19.
- Hentschel, Martin, et al. "Finding users we trust: Scaling up verified Twitter users using their communication patterns." Eighth International AAAI Conference on Weblogs and Social Media. 2014.

Please label the x-axis of the Fig4 and Fig11 to help reader.

I also appreciate authors for going along with my humorous colour palette.

Reviewer #2 (Remarks to the Author):

The authors addressed all of my comments, and I am happy to recommend the work for publication. I appreciate the effort to make data & code available. For maximal impact, I highly recommend to make sure both repositories are online before publication, possibly using an institutional repository like figshare (for data) and github (for code).

Reviewer #3 (Remarks to the Author):

I really appreciated the substantial amount of work that the authors have done to clarify both the motivations, the results and some of the methodological aspects of their paper. My concerns are fully addressed, and I suggest publication

REBUTTAL LETTER

Dear Reviewers:

We are pleased for the appreciation of our revised version of the manuscript. We sincerely thank you for your comments, which have allowed us to substantially improve the paper. Please, find below your last comments and our answers. We have highlighted in blue all the modified paragraphs in the revised manuscript.

Thank you for your support and consideration.

Yours faithfully,

Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco

Reviewer #1

Summary:
I want to start by thanking authors for extending their results and rewriting particular sections of the paper for better clarity. I have a few minor comments on the revised version, and I think authors can address them easily.

Comment 1:
The distinction between verified and normal Twitter users is important for the analysis. There have been research studies on characterizing verified users on Twitter. I suggest a couple of works that study verified users and their characterizations as well as relations to social bot research.

- *Varol, Onur, and Ismail Uluturk. Journalists on Twitter: self-branding, audiences, and involvement of bots. Journal of Computational Social Science (2019): 1-19.*
- *Hentschel, Martin, et al. Finding users we trust: Scaling up verified Twitter users using their communication patterns. Eighth International AAAI Conference on Weblogs and Social Media. 2014.*

Please label the x-axis of the Fig4 and Fig11 to help reader.
I also appreciate authors for going along with my humorous colour palette.

Reply:

Following the reviewer's suggestion, we have introduced motivations for taking into account verified users, also citing the suggested papers. The new part can be found in Section 'User Polarization' of the Results, highlighted in blue.

Regarding the figures, we thought a bit about the suggestion. Finally, since they are bar charts, it seemed more appropriate to us to explain more in details, in the captions, what the different colors refer to. Please note that the two figures have been moved. One is noe Figure 10 in the main text, the other one is Supplementary Figure 1 in the Supplementary Information. This has been done for editorial requirements.

We thank the reviewer for the appreciation of the revised manuscript and, yes, we thought it'd be fun to take the advice on the palette :)

Reviewer #2

Summary:
The authors addressed all of my comments, and I am happy to recommend the work for publication. I appreciate the effort to make data & code available. For maximal impact, I highly recommend to make sure both repositories are online before publication, possibly using an institutional repository like figshare (for data) and github (for code)..

Reply: We are really grateful to the reviewer for the appreciation. The data are now available on the TOFFEe project website, and we are preparing the code publication.

Reviewer #3

Summary:
I really appreciated the substantial amount of work that the authors have done to clarify both the motivations, the results and some of the methodological aspects of their paper. My concerns are fully addressed, and I suggest publication

Reply: We are very pleased with the appreciation. Thanks very much.