

Anatomy of news consumption on Facebook

Ana Lucía Schmidt^a, Fabiana Zollo^{a,1}, Michela Del Vicario^a, Alessandro Bessi^b, Antonio Scala^{a,c}, Guido Caldarelli^{a,c}, H. Eugene Stanley^d, and Walter Quattrociocchi^{a,2}

^aLaboratory of Computational Social Science, Networks Department, IMT Alti Studi Lucca, 55100 Lucca, Italy; ^bIUSS Institute for Advanced Study, 27100 Pavia, Italy; ^cISC-CNR Uos "Sapienza," 00185 Rome, Italy; and ^dDepartment of Physics, Boston University, Boston, MA 02115

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved January 31, 2017 (received for review October 14, 2016)

The advent of social media and microblogging platforms has radically changed the way we consume information and form opinions. In this paper, we explore the anatomy of the information space on Facebook by characterizing on a global scale the news consumption patterns of 376 million users over a time span of 6 y (January 2010 to December 2015). We find that users tend to focus on a limited set of pages, producing a sharp community structure among news outlets. We also find that the preferences of users and news providers differ. By tracking how Facebook pages “like” each other and examining their geolocation, we find that news providers are more geographically confined than users. We devise a simple model of selective exposure that reproduces the observed connectivity patterns.

computational social science | Facebook | news consumption | misinformation

A large body of research has addressed news consumption on online social media and its polarizing effect on public opinion (1–5). Social media and microblogging platforms have changed the way we access information and form opinions. Communication has become increasingly personalized, both in the way messages are framed and how they are shared across social networks. Furthermore, according to a recent study (6), ~63% of users acquire their news from social media, and these news are subject to the same popularity dynamics as other forms of content. Recent works (7) provide empirical evidence of the pivotal role of confirmation bias and selective exposure in online social dynamics. Users, indeed, tend to focus on specific narratives and join polarized groups (i.e., echo chambers) (8–10), where they end up reinforcing their worldview [even if pieces of content are deliberately false (11, 12)] and dismissing contradictory information (13). Discussion and elaboration of narratives in such a segregated environment elicits group polarization and negatively influences user emotion (14–17). Therefore, in this paper, to better understand how echo chambers emerge, we explore the anatomy of news consumption on Facebook. We focus on how Facebook posts from news outlets are consumed and how user activity causes connectivity patterns to emerge. We analyze the interaction of 376 million users with all of the anglophone news outlets on Facebook listed in the European Media Monitor (18) over a 6-y time span, from January 2010 to December 2015. Using quantitative analysis, we find evidence that selective exposure plays a pivotal role in shaping news consumption online. Users tend to focus on a very limited set of pages and thus create a distinct community structure within these news outlets. We also find that the perspectives of the news outlets and the users differ. Our findings suggest that users have a more cosmopolitan perspective of the information space than news providers. Examining how pages “like” each other and taking into account their geolocation, we find geographically confined connectivity patterns. We conclude by devising a simple model of selective exposure that reproduces the observed connectivity patterns. We first analyze user behavior with respect to the information sources. We then analyze how user activity spans across these sources. Finally, we introduce a simple model that reproduces the observed dynamics. Our findings suggest that probably the main driver of misinfor-

mation diffusion is the polarization of users on specific narratives rather than the lack of fact-checked certifications.

Results and Discussion

Users’ Attention. News items on Facebook appear in posts that can be liked, commented, or shared by users. A like is usually a positive feedback on a news item. A share indicates a desire to spread a news item to friends. A comment can have multiple features and meanings and can generate collective debate. The likes, shares, and comments on Facebook posts present a heavy-tailed distribution (*SI Appendix, 2. Attention Pattern*). The lifetime of a post is the time period between the first and the last comment, and it presents a peak at 24 h. User activity is heterogeneous and the number of likes and comments ranges from very few (the majority) to hyperactivity. The Complementary Cumulative Distribution Function of the number of likes and comments for single users exhibits heavy tails (*SI Appendix*). The overall number of likes of each user is a good proxy for their engagement with Facebook news pages and the lifetime of each user can be approximated by the length of time between the date of their first comment and their last comment. These measures could provide important insights about news consumption. Our goal is to quantify the turnover of Facebook news sources by measuring the heterogeneity of user activity, and thus we measure the total number of pages a user interacts with. Fig. 1 shows the number of news sources a user interacts with for the lifetime (i.e., the distance in time between the first and last interaction with a post) and for increasing levels of engagement (i.e., the total number of likes). For a comparative analysis, we standardized between 0 and 1 both lifetime and engagement over the entire user set.

Significance

Social media heavily changed the way we get informed and shape our opinions. Users’ polarization seems to dominate news consumption on Facebook. Through a massive analysis on 920 news outlets and 376 million users, we explore the anatomy of news consumption on Facebook on a global scale. We show that users tend to confine their attention on a limited set of pages, thus determining a sharp community structure among news outlets. Furthermore, our findings suggest that users have a more cosmopolitan perspective of the information space than news providers. We conclude with a simple model of selective exposure that well reproduces the observed connectivity patterns.

Author contributions: A.L.S., A.S., and W.Q. designed research; A.L.S., F.Z., M.D.V., H.E.S., and W.Q. performed research; F.Z., M.D.V., and A.B. contributed new reagents/analytic tools; A.L.S., F.Z., M.D.V., A.S., G.C., H.E.S., and W.Q. analyzed data; and A.L.S., A.S., H.E.S., and W.Q. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

¹Present address: Dipartimento di Scienze Ambientali, Informatica, e Statistica (DAIS), University of Venice, 30172 Venice, Italy.

²To whom correspondence should be addressed. Email: walterquattrociocchi@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617052114/-DCSupplemental.

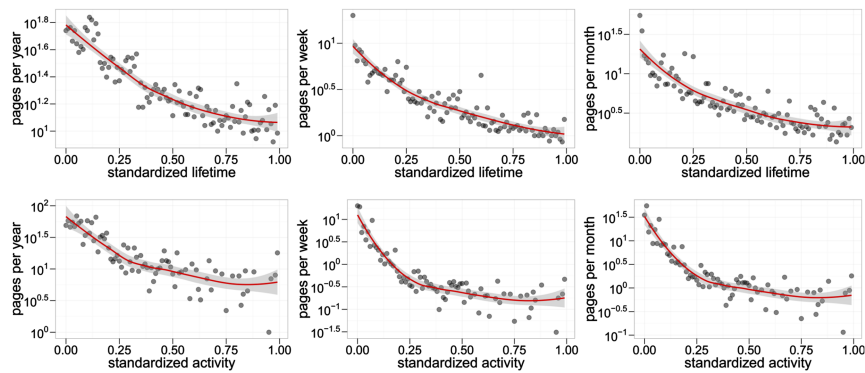


Fig. 1. Users' attention patterns. (Top) Maximum number of unique news sources that users with increasing levels of standardized lifetime interacted with monthly, weekly, and yearly. (Bottom) Maximum number of unique news outlets which users with increasing levels of standardized activity interacted with monthly, weekly, and yearly.

Fig. 1 shows the results for the yearly time window (first column) and for the weekly (second column) and monthly (third column) rates. Note that a user usually interacts with a small number of news outlets and that higher levels of activity and longer lifetime correspond to a smaller number of sources. There is a natural tendency of the users to confine their activity on a limited set of pages. According to our findings, news consumption on Facebook is dominated by selective exposure.

Clusters and Users' Polarization. User tendency to interact with few news sources might elicit page clusters. To test this hypothesis, we first characterize the emergent community structure of pages according to the users' activity. We project the user page likes to derive the weighted graph G_l^p (and G_c^p) in which nodes are pages and two pages are connected if a user likes (or comments on) both of them. The weight of a link on a projected graph is determined by the number of users the two pages have in common. Fig. 2 shows the backbone of G_l^p and G_c^p . Colors indicate node membership, as detected by the Fast Greedy (FG) algorithm (see *Methods* for further details). A histogram of community sizes is provided in *SI Appendix, Fig. S5*. To validate the community partitioning, we compare the membership of other community detection algorithms using the Rand method (19, 20) and find a high level of similarity (see *SI Appendix* for further details). We also compared the communities of G_l^p and G_c^p against each other using different community detection algorithms and find some level of similarity (see *SI Appendix, Tables S3 and S4* in for further details). By examining the activity of users across the various clusters and measuring how they span across news outlets, we find that most users remain confined within specific clusters. To understand the relationship between page groupings and user behavior, we quantify the fraction of activity of user u in the largest communities $w_k^u, k \in \{1...5\}$, and the fraction of activity of user u in any other community $w_0^u = 1 - \sum_{k>5} w_k^u$. Fig. 3 shows the activity of users across the five largest communities (Fig. 3, Left) and compares this with a null model (Fig. 3, Right) in which user activity is randomly distributed. We find that users are strongly polarized and that their attention is confined to a single community of pages. User interaction to Facebook news outlets indicate a dominant community structure with sharply identified groups. Because users tend to focus on a small number of pages, the news sphere of Facebook is clustered and dominated by a precise community structure and users tend to focus their attention on a single group of news outlets.

Users' and News Outlets' Perspectives. Facebook pages can like each other. We use this pattern of favorite pages to create a graph of page preferences. In news outlets, these preferences can

be used to compare the perspectives of users and news providers. We use the bipartite projection of the pages that users like G_l^p and define N_p as the network of new outlets that like each other (i.e., N_p is the network in which nodes are pages and links are pages liking each other). We analyze both networks by determining the geographical location of each page. Thus, each node is identified by its region and country, information provided by the European Media Monitor. To determine the community structure of both N_p and G_l^p , we compare the outputs of different community detection algorithms (see *Methods* for further details). Fig. 4 shows the communities N_p and G_l^p represented by taking into account the geographical location of the pages. In the graph, the external bundle groups pages by region, the middle bundle by nation, and node colors indicate community membership as identified by the FG algorithm. Note that in the plot, we use the backbone structure of the networks for visualization purposes (see *Methods* for further details). As in the previous section, we validate the partition found by the FG algorithm by comparing it with results from other community detection algorithms (see *SI Appendix* for further details). Using refs. 19 and 20, we compare the FG community structure of N_p and G_l^p and do not find significant differences. When comparing the FG communities of N_p and G_l^p (the projection of user likes) the similarity index is 0.67, for N_p and G_c^p (the projection of user comments), the similarity index is 0.69 (for the comparison with other community detection algorithms, see *SI Appendix*). Fig. 4 shows that in N_p , the community structure is more

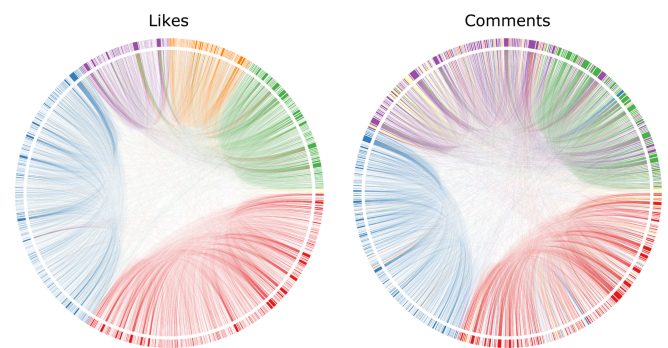


Fig. 2. Community structure. (Left) Backbone of the projections on pages of the users' likes (G_l^p). (Right) Comments (G_c^p). The color of the nodes indicate the Fast Greedy community. Nodes in G_l^p are ordered according to the detected communities, whereas in G_c^p , the nodes follow the same order as in G_l^p .

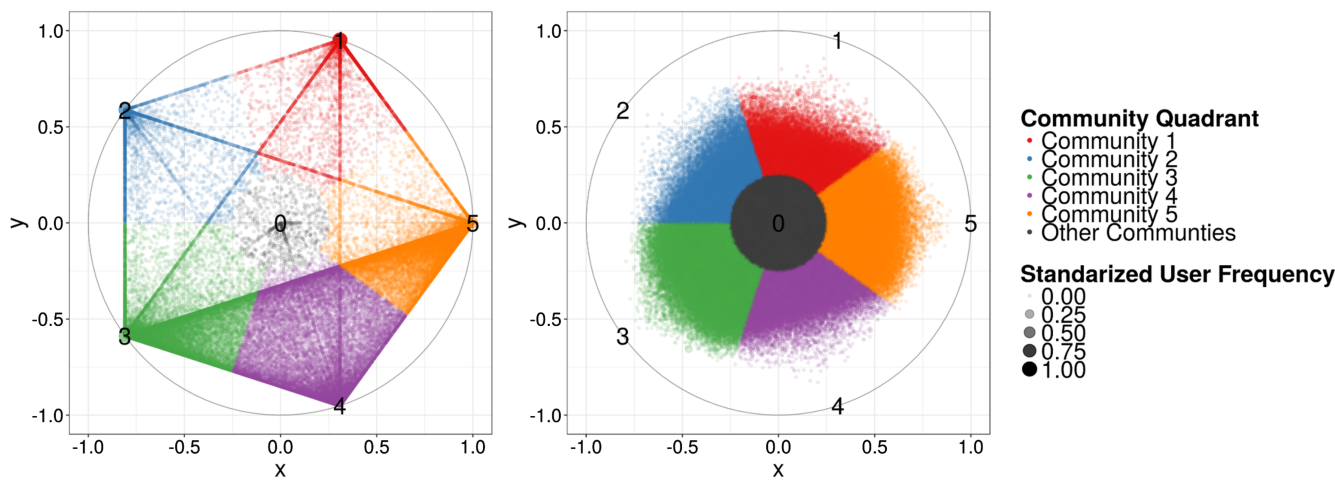


Fig. 3. Users Polarization. (Left) Activity of users across the five largest communities. (Right) Null model where users' activity is randomly distributed. Vertices of the pentagon represent the five largest communities and the central point all of the remaining ones. The position of each dot is determined by the number of communities the users interacts with. The size and transparency indicate the number of users in that position.

confined within geographical boundaries than in G_i^p . Because the geographical location of pages also defines a community partitioning, we compare pages in G_i^p , G_c^p , and N_p according to their community partitioning as detected by the FG algorithm and to their grouping according to the geographical location. The Rand similarity index among the partitions obtained by community detection algorithms and the partitions based on the geographical location is 0.71 for G_i^p , 0.72 for G_c^p , and 0.84 for N_p (for the comparison with other community detection algorithms, see

SI Appendix). This finding suggests that page communities are more locally confined than user interaction communities, which can span across nations and continents.

The Model. Users on Facebook tend to focus on a limited set of news sources, on a macro scale. This mechanism of selective exposure generates a clustered and polarized structure. The community structure that emerges when analyzing the likes among the pages is different from the community structure defined by

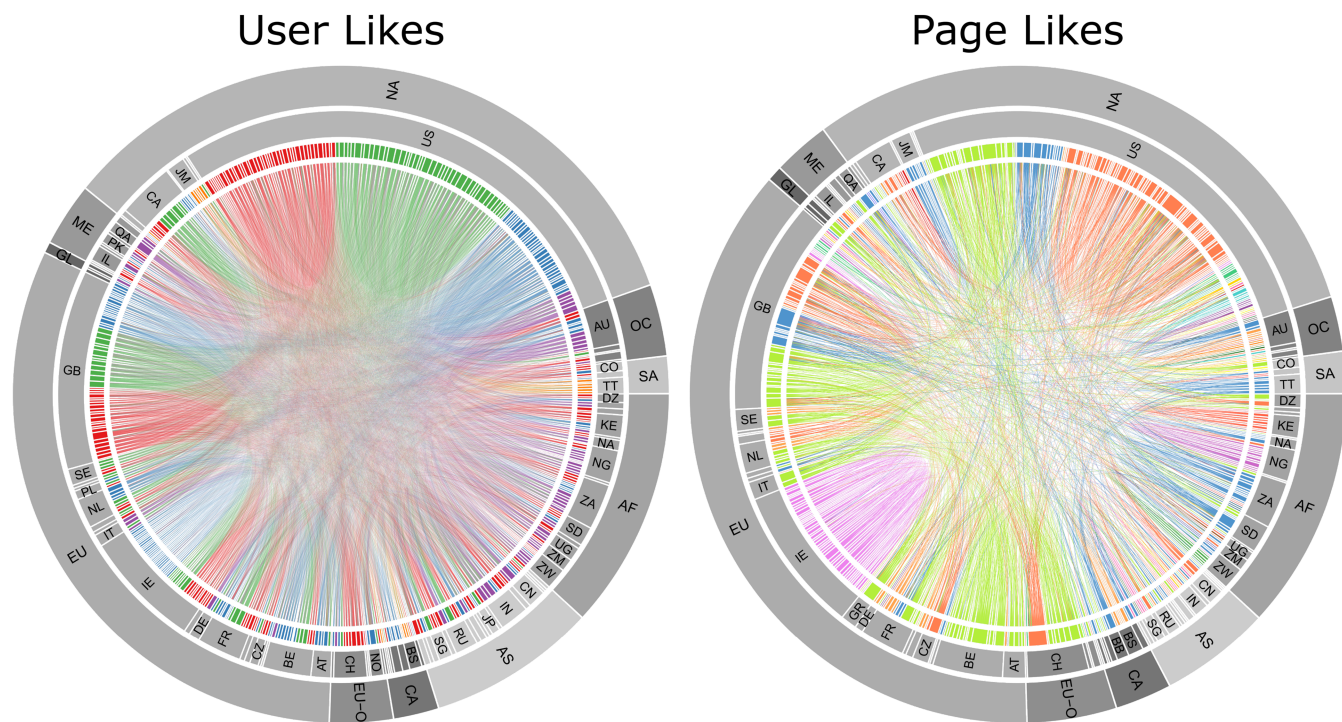


Fig. 4. Pages and users' communities and locations. Backbone of the projections on pages of the user likes reduced to the pages that appear in N_p (Left) and the network of pages liking each other (Right). Inner nodes represent the pages and their color indicates the Fast Greedy community, middle track marks the country, and outer track the region as established by the European Media Monitor. Order of the inner nodes in both plots is done by region, country, and community, in that order. AF, Africa; AS, Asia; CA, Central America; EU, European Union; EU-C, EU Candidate; EU-O, EU Other; GL, Global; ME, Middle East; NA, North America; OC, Oceania; SA, South America.

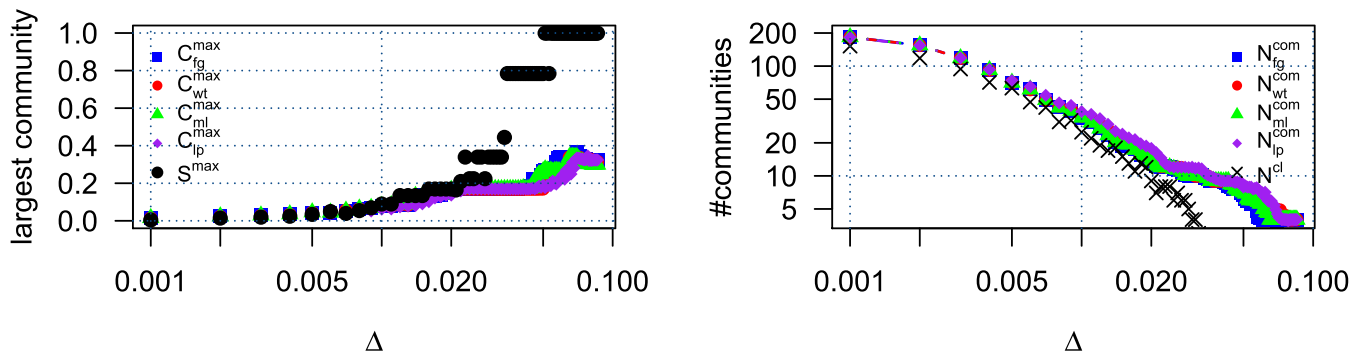


Fig. 5. Analysis of the synthetic pages-to-pages graphs G_{sim}^p generated according to our extension of the BCM model. (Left) We show the size of the largest component S_{max} and the size of the largest community $|C_{max}^{alg}|$ versus the tolerance Δ . (Right) We show the number of connected components N_{cl} and the number of detected communities N_{com}^{alg} ; notice that by definition, the number of communities must be $N_{com}^{alg} \geq N_{cl}$.

the users' interaction with the pages. In this section, we provide a simple model of users' preferential attachment to specific sources that reproduces the observed community structure.

The entities of our model are pages $p \in P$ and users $u \in I$. Each page p is characterized by a set of opinions (an editorial line) modeled as a real number c_p that ranges $[0..1]$. We assume that the c_p values are uniformly distributed. Each user u has an initial opinion that is modeled as a real number θ_u , which ranges between $[0..1]$ and is uniformly distributed. We suppose c_p and θ_u to be homogeneous such that the quantity $|c_p - \theta_u|$ is the distance between the opinion of user u and the editorial line of page p . We mimic confirmation bias by assuming that if user u interacts with a page p and the opinion distance $|c_p - \theta_u|$ is less than a given tolerance parameter Δ , the preference of user u will converge toward the editorial line of page p according to the Bounded Confidence Model (BCM) (21) equation $\theta_u' = (1 - \mu) \cdot \theta_u + \mu \cdot c_p$ where μ is a simple convergence parameter. To mimic user activity, we give each user u an activity coefficient a_u that represents the number of pages a user can visit. Thus, the final opinion of a user will average the editorial lines of the pages the user likes. If Ω is the set of $|\Omega|$ pages that matches the preferences of user u , then the average opinion will be $\theta_u = (1 - \mu) \theta_u + \mu |\Omega|^{-1} \sum_{p \in \Omega} c_p$ (i.e., $\bar{\theta}_u = |\Omega|^{-1} \sum_{p \in \Omega} c_p$). To mimic the long tail distribution of our data, we set the activity distribution to be power law distributed $p(a) \sim a^{-\gamma}$ with exponent $\gamma = 3$. We use numerical simulation to study our model. A user randomly selects a subset of P with which to interact. The user likes a page only when $|c_p - \theta_u| < \Delta$. When this occurs, the feedback mechanism of the BCM equation reinforces the user's page preference. Thus, the final opinion of a user will be the average of the editorial lines of the pages the user likes. When a user's opinion converges, we build in the bipartite graph $\mathcal{B}_{sim} = (I, P, E_{sim})$, where the set of edges E_{sim} are the couplings (u, p) with which user u likes page p . Hence, \mathcal{B}_{sim} represents users interacting with their favorite pages, and from \mathcal{B}_{sim} , we can build the projected graph G_{sim}^p that links the pages according their common users. Fig. 5 shows an analysis of G_{sim}^p as a function of the tolerance parameter Δ . Each point of the simulation is averaged over 50 iterations. Fig. 5, Left shows both the size of the largest connected component S_{max} and the size of the largest community $|C_{max}^{alg}|$ detected by several algorithms on G_{sim}^p . Fig. 5, Right shows the number N_{clu} of connected components and the number of communities N_{com}^{alg} detected by several algorithms of G_{sim}^p . At $\Delta \sim 0$, G_{sim}^p is broken down into disconnected pieces (N_{clu} is of the order of the number of nodes $|P|$ of G_{sim}^p), and the size S_{max} of the largest component is extremely small. Although this regime is unlike real online social networks that are usually dense and strongly connected, S_{max} increases rapidly to $|P|$ as Δ departs from zero, indicating that G_{sim}^p becomes a sin-

gle connected graph at $\Delta \sim 0.03$. On the other hand, the size of the largest communities detected by the various algorithms are consistently smaller than $|P|$ (Fig. 5, Left), and the number of communities is consistently greater than one and decreases slowly with increasing Δ (Fig. 5, Right). Thus, the page-page graph G_{sim}^p shows a stable, nontrivial community structure induced by user preferences even when it is a dense, connected graph like real online communities.

Conclusions

Using quantitative analysis, we show that the more active a user is, the more the user tends to focus on a small number of news sources. Looking at the page clusters generated by user activity, we find a distinct community structure and strong user polarization. We provide evidence that preferences of users and news outlets differ in that communities established by page creators are more locally confined than communities identified by the users' activity, which can span across international borders. This segregation in distinct communities can be reproduced by a simple model that mimics the selective exposure of users. Content consumption on Facebook is strongly affected by the tendency of users to limit their exposure to a few sites. Despite the wide availability of content and heterogeneous narratives, there is major segregation and growing polarization in online news consumption. News undergoes the same popularity dynamics as popular videos of kittens or selfies. The spreading of fake news and unsubstantiated rumors motivated major corporations like Google and Facebook to provide solutions to the problem. Google news decided to flag fact-checked information and to penalize providers of fake news; others are proposing to use black lists of sources to automatically limit their spread. However, often debates, especially on socially relevant issues, are based upon conflicting narratives. Probably, the main problem behind misinformation is polarization of users online.

Methods

Ethics Statement. The entire data collection process is performed exclusively by means of the Facebook Graph application programming interface (22), which is publicly available. We used only publicly available data (users with privacy restrictions are not included in our dataset). We abided by the terms, conditions, and privacy policies of Facebook. We did not seek ethical approval because the data were preexisting.

Data Collection. The European Media Monitor provides a list of all news sources. We limit our collection to all those pages reporting in English. The downloaded data from each page includes all of the posts made from January 1, 2010 to December 31, 2015, as well as all of the likes and comments on those posts. The European Media Monitor list includes the country and the region of each news source. For accurate mapping on the globe, we also collected the geographical location—latitude and longitude—of each page.

A breakdown of data and the set of pages used in the analysis is provided in *SI Appendix, Table S1*.

Definitions. In this section, we provide a brief description of the main concepts and tools used in the analysis.

Projection of Bipartite Graphs. A bipartite graph is a triple $\mathcal{B} = (A, B, E)$ where $A = \{a_i | i = 1 \dots n_A\}$ and $B = \{b_j | j = 1 \dots n_B\}$ are two disjoint sets of vertices, and $E \subseteq A \times B$ is the set of edges (i.e., edges exist only between vertices of sets A and B). A bipartite graph \mathcal{G} is described by a rectangular matrix M , defined as $M_{ij} = 1$ if there is an edge between a_i and b_j , and $M_{ij} = 0$ otherwise.

We consider bipartite networks in which the two disjointed sets of nodes are users and Facebook pages. Edges represent interactions among users and pages. As an example, a like to a given piece of information posted by a page constitutes a link between the user and the page and $M_{p,u} = 1$ will indicate that user u has liked a post on page p . Indicating with P the set of pages and with U the set of users, we can build the cooccurrence matrices $C^P = MM^T$ and $C^U = M^T M$ that quantify, respectively, the number of common neighbors between two vertices of P or U .

Community Detection Algorithm. A community detection algorithm is used to identify groups of nodes in a network. The strategy relies on the modularity that quantifies the division of a network into separated clusters, and a high modularity corresponds to a dense connectivity between nodes in a community and sparse connections between modules. We use four community-detection algorithms. (i) The fast greedy (FG) algorithm measures the maximum modularity by considering all possible community structures in the network. Every vertex initially belongs to a separate community, and communities are merged iteratively such that each merge is locally opti-

mal (i.e., it yields the largest increase in the current value of modularity). The algorithm stops when it is no longer possible to further increase modularity (23). (ii) The walk trap (WT) algorithm exploits the fact that a random walker tends to become trapped in the denser parts of a graph (i.e., in communities). Hence, WT uses short random walks to merge separate communities (24). (iii) The multilevel (ML) algorithm uses a multilevel modularity optimization procedure. Each vertex is initially assigned to a community. At each step vertices are then reassigned to communities and nodes move to the community in which they have the highest modularity (25). (iv) The label propagation (LP) algorithm (26) gives a unique label to each vertex, which is then updated according to majority voting in the neighboring vertices. Dense node groups quickly reach a consensus on a common label. Finally, to compare the various community structures, we use standard methods that compare the similarity between different clustering methods and consider how nodes are assigned in each community detection algorithm (19, 20).

Backbone Detection Algorithm. The disparity filter algorithm is a network reduction technique that identifies the backbone structure of a weighted network without destroying its multiscale nature (27). We use this algorithm to determine the connections that form the backbones of our networks and to produce clear visualizations.

ACKNOWLEDGMENTS. The authors thank Sandro Forgione and Pandoros. This work was funded by EU Future and Emerging Technologies (FET) MULTIPLEX Project 317532, FET SIMPOL Project 610704, FET DOLFINs Project 640772, SoBigData Project 654024, and Center of Excellence for Global Systems Science Project 676547. The funders had no role in the study design, data collection and analysis, decision to publish, or manuscript preparation.

1. Anderson AA, Brossard D, Scheufele DA, Xenos MA, Ladwig P (2014) The "Nasty effect." Online incivility and risk perceptions of emerging technologies. *J Comput Mediat Commun* 19(3):373–387.
2. Webster JG, Ksiazek TB (2012) The dynamics of audience fragmentation: Public attention in an age of digital media. *J Commun* 62(1):39–56.
3. Bennett WL (2015) Changing societies, changing media systems: Challenges for communication theory, research and education. *Can the Media Serve Democracy? Essays in Honour of Jay G. Blumler*, eds Coleman S, Moss G, Parry K (Palgrave Macmillan, London), pp 151–163.
4. Ksiazek TB, Malthouse EC, Webster JG (2010) News-seekers and avoiders: Exploring patterns of total news consumption across media and the relationship to civic participation. *J Broadcast Electron Media* 54(4):551–568.
5. Iyengar S, Hahn KS (2009) Red media, blue media: Evidence of ideological selectivity in media use. *J Commun* 59(1):19–39.
6. Newman N, Levy DAL, Nielsen RK (2015) Reuters institute digital news report 2015 (Social Science Research Network). Available at <https://ssrn.com/abstract=2619576>.
7. Quattrociocchi W, Scala A, Sunstein CR (2016) Echo chambers on Facebook. (Social Science Research Network). Available at <https://ssrn.com/abstract=2795110>.
8. Bessi A, et al. (2015) Trend of narratives in the age of misinformation. *PLoS One* 10(8):e0134641.
9. Bessi A, et al. (2015) Viral misinformation: The role of homophily and polarization. *Proceedings of the 24th International Conference on World Wide Web Companion* (International World Wide Web Conferences Steering Committee, Florence, Italy), pp 355–356.
10. Del Vicario M, et al. (2016) The spreading of misinformation online. *Proc Natl Acad Sci USA* 113(3):554–559.
11. Mocanu D, Rossi L, Zhang Q, Karsai M, Quattrociocchi W (2015) Collective attention in the age of (mis) information. *Comput Hum Behav* 51:1198–1204.
12. Bessi A, et al. (2015) Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS One* 10(2):e0118093.
13. Zollo F, et al. (2015) Debunking in a world of tribes. arXiv:1510.04267.
14. Sunstein CR (2002) The law of group polarization. *J Polit Philos* 10(2):175–195.
15. Zollo F, et al. (2015) Emotional dynamics in the age of misinformation. *PLoS One* 10(9):e0138740.
16. Yardi S, Boyd D (2010) Dynamic debates: An analysis of group polarization over time on twitter. *Bull Sci Technol Soc* 30(5):316–327.
17. Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239):1130–1132.
18. Steinberger R, Pouliquen B, Van Der Goot E (2013) An introduction to the Europe media monitor family of applications. arXiv:1309.5290.
19. Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 66(336):846–850.
20. Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218.
21. Deffuant G, Neau D, Amblard F, Weisbuch G (2000) Mixing beliefs among interacting agents. *Adv Complex Syst* 3(01n04):87–98.
22. Facebook (August 2013) Using the graph API (Facebook, Menlo Park, CA). Available at <https://developers.facebook.com/docs/graph-api/using-graph-api>. Accessed January 19, 2014.
23. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. *Phys Rev E* 70(6):066111.
24. Pons P, Latapy M (2006) Computing communities in large networks using random walks. *J Graph Algorithms Appl* 10(2):191–218.
25. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 2008(10):P10008.
26. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E* 76(3):036106.
27. Serrano MA, Boguna M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci USA* 106(16):6483–6488.