Zhao et al. *EPJ Data Science* (2020) 9:7
https://doi.org/10.1140/epjds/s13688-020-00224-z

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Fake news propagates differently from real news even at early stages of spreading

Zilong Zhao[1,2†], Jichang Zhao[3†], Yukie Sano[4†], Orr Levy[5†], Hideki Takayasu[6,7], Misako Takayasu[7], Daqing Li[1,2*], Junjie Wu[3,8*] and Shlomo Havlin[5,7]

*Correspondence:
daqingl@buaa.edu.cn;
wujj@buaa.edu.cn
[1]School of Reliability and Systems Engineering, Beihang University, Beijing, China
[3]School of Economics and Management, Beihang University, Beijing, China
Full list of author information is available at the end of the article
†Equal contributors

Abstract

Social media can be a double-edged sword for society, either as a convenient channel exchanging ideas or as an unexpected conduit circulating fake news through a large population. While existing studies of fake news focus on theoretical modeling of propagation or identification methods based on machine learning, it is important to understand the realistic propagation mechanisms between theoretical models and black-box methods. Here we track large databases of fake news and real news in both, Weibo in China and Twitter in Japan from different cultures, which include their traces of re-postings. We find in both online social networks that fake news spreads distinctively from real news even at early stages of propagation, e.g. five hours after the first re-postings. Our finding demonstrates collective structural signals that help to understand the different propagation evolution of fake news and real news. Different from earlier studies, identifying the topological properties of the information propagation at early stages may offer novel features for early detection of fake news in social media.

Keywords: Fake news; Social network; Early detection

1 Maintext

Social networks such as Twitter or Weibo, involving billions of users around the world, have tremendously accelerated the exchange of information and thereafter have led to fast polarization of public opinion [1]. For example, there is a large amount of fake news about the 3.11 earthquake in Japan, where about 80 thousand people have been involved in both diffusion and correction [2]. These fake news, which can be fabricated stories or statements yet without confirmation, circulate online pervasively through the conduit offered by on-line social networks. Without proper debunking and verification, the fast circulation of fake news can largely reshape public opinion and undermine modern society [3]. Even worse, fake news can be intentionally fabricated, leading to diverse threats to modern society including turmoil or riot. The later fake news is identified and corrected the greater the damage it can make, due to its fast propagation. Thus, detecting fake news at their early stages, in order to effectively avoid further risks and damages, is crucial.

Different from the age of word of mouth, identification of fake news in the online social network by experts is generally labor-intensive with low efficiency [4], which has at-

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Springer

tracted much research attention to provide alternative solutions. One intuitive idea for understanding fake news spreading is inspired by epidemic models. In the 1960s, Daley and Kendall proposed the so-called DK model [5] in which agents are divided into ignorant, spreader and stifle. Its later extensions are based on the known epidemic spreading models such as SIS model [6, 7], SIR model [8, 9], SI model [10, 11] and SIRS model [12]. While these studies focus on theoretical modeling of fake news propagation, the availability of real data in online social platforms, as we show here, can provide an opportunity to deepen our understanding of the realistic information cascades. Different kinds of observations have been made in empirical studies of fake news, including linguistic features [13], temporal features of re-postings [14–16] and user profiles [17–19]. Actually, information cascades in online social networks are collective propagation networks of which critical topological features remain yet unknown. This motivates our present study to analyze and compare empirically the propagation networks between fake and real news, especially in their early stage, so as to identify the propagation differences and mechanisms behind. These topological features could help to design machine learning approaches to essentially boost the accuracy of fake news targeting [20–22].

Very recently, based on empirical datasets, it has been found that the propagation network of fake news is different from that of real news [23]. They have found that falsehood propagates significantly farther, faster, deeper, and broader than truth news in many categories of information. While this study provides the possibility to differentiate fake news from real news based on the propagation network, it remains unclear how this difference between fake news and real news emerges and how soon one can separate these two types. Thus, a systematic study for the dynamic evolution of propagation topology is still missing. This motivated us to explore deeper in this direction of how the propagation evolves topologically in different scenarios. With collected real data, we identified early signals for identifying fake news, at five hours from the first re-posting, without other information on contents or users. Note that different from considering all the cascade components [23], our finding is valid for even only following the largest cascade component.

Based on realistic traces of real and fake news propagation in both Weibo (from China) and Twitter (from Japan), we use the re-posting relationships between different users to establish propagation networks (see Methods for details). Given similar popularity scales, we find that fake news shows significant different topological features from real news. These novel topological features will enable us to design an efficient algorithm to distinguish between fake news and real news even shortly after their birth.

2 Results

To construct the propagation network of fake and real news, we utilize the re-posting relation between different users participating in circulating the same message (*see Methods and Table* 1). A schematic description of such propagation networks is shown in Fig. 1A. Typical propagation networks of fake news and real news in Weibo and Twitter are demonstrated in Fig. 1B–E. The topology of the propagation network of fake news and real news can be seen to be different. For example, the number of layers in fake news (*Fig.* 1*B and* 1*D*) is typically larger than that of real news (*Fig.* 1*C and* 1*E*). Additionally, from looking at various examples of fake news propagation networks, it is somewhat surprising that for widely distributed fake news, the creator does not usually have the largest degree in the propagation network (*Figs. S1 and S2*). In the following, our analysis considers also real news

Table 1 Number of users and networks for different propagation networks

	Fake news	Real news
Weibo users in the whole dataset	973,391	347,401
Twitter users in the whole dataset	105,335	133,109
Number of Weibo propagation networks (larger than 200 re-postings)	1701	492
Number of Weibo propagation networks which can be studied at five hours (larger than 200 re-postings)	1578	448
Number of Twitter propagation networks (larger than 200 re-postings)	27	28

The Weibo networks are from the first Weibo datasets.

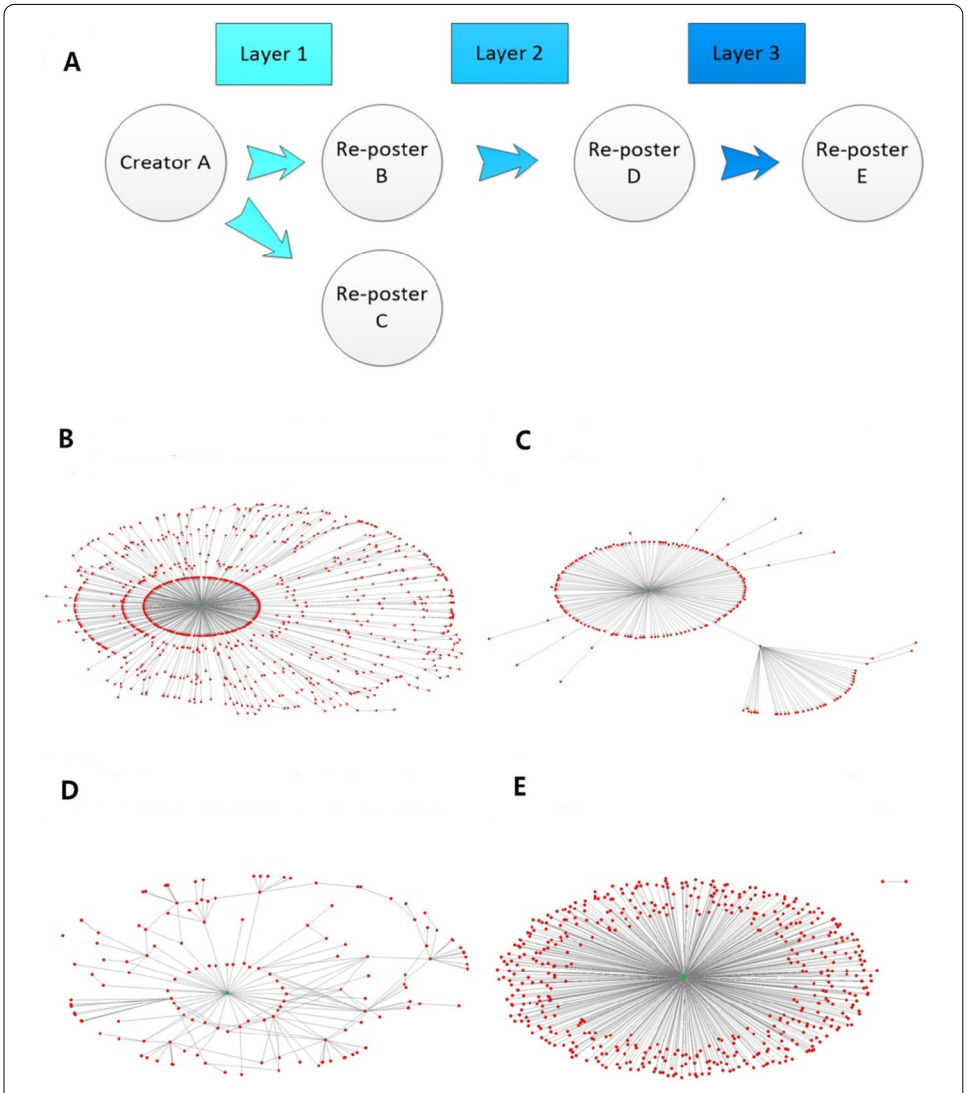

Figure 1 Typical examples of fake and real news networks. (**A**) Schematic diagram of the propagation of a post and its re-posting. The nodes represent the users and the edges are the re-postings. The directionality determines which user is the re-poster among the two users: the origin is the former re-poster and the target is the later re-poster. A layer consists of re-postings whose re-posters have the same distance from the creator. We color the edges according to their layer from light to dark blue. (**B**) A real typical Weibo network of fake news with 1123 nodes. The edge's arrow stands for its direction. This fake news is about health problems due to a milk tea shop. (**C**) A typical Weibo real news network with 215 nodes. This is about a tip of preventing sunstroke. (**D**) A typical Twitter fake news network with 199 nodes. This tweet is about an electric store that raised the price of a battery unreasonably. (**E**) A typical real news network on Twitter with 578 nodes. This tweet is a correction tweet against fake news about Cosmo oil by the Asahi newspaper. We applied the Fruchterman–Reingold layout by using Pajek software here

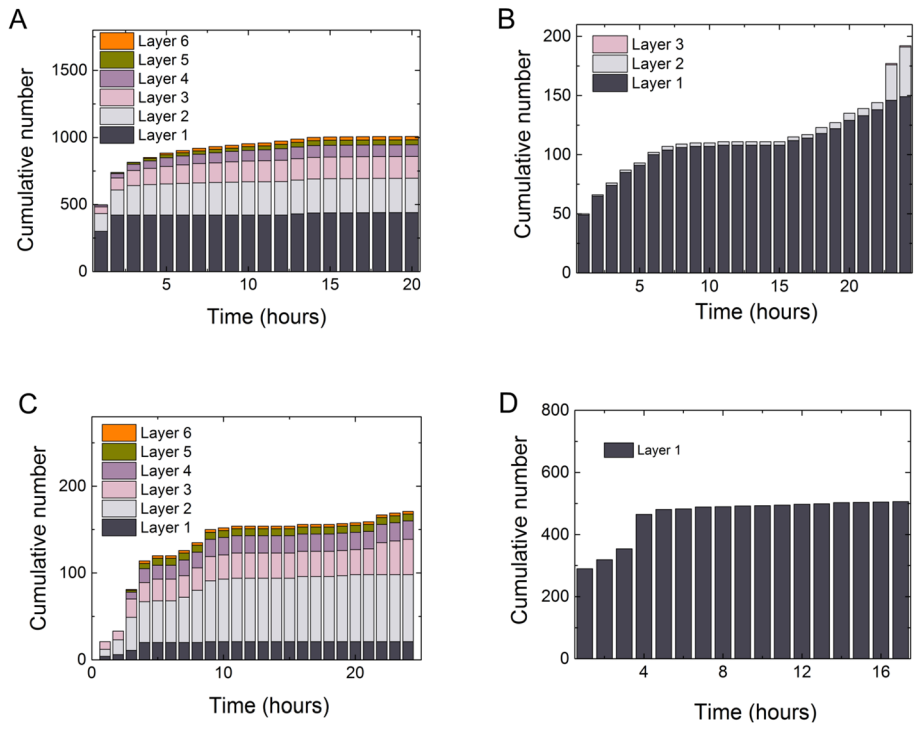

Figure 2 Different layer sizes as a function of time in typical networks. The *y* axis is the cumulative number of re-postings at different layers of typical networks in Weibo and Twitter. The *x* axis is the time (in hours) from the time of news creation and the different colors stand for different layers. Shown examples are (**A**) fake news and (**B**) real news in Weibo, as well as (**C**) fake news and (**D**) real news on Twitter. These four typical networks are the same networks shown in Fig. 1. In Fig. 2A, the fraction of nodes located in the first layer is around 45% of all the nodes at the end of propagation. However, in Fig. 2B, the total number of nodes in the first layer occupies about 78% of all the nodes at the end of propagation. If the total number of nodes does not change much after 20 hours, we ignore the re-posting after 20 hours in order to clearly see the layers in the figure. It is seen that the layer sizes of real news and fake news are significantly different in both Weibo and Twitter. Real news networks tend to have a relatively larger first layer, while fake news networks are relatively uniformly distributed in different layers

created by non-official sources, to avoid the artificial differences due to different types of information creators (official or non-official accounts).

Layer ratio. The layer number is defined as the number of hops from the creator to a given node for a given propagation network. The cumulative numbers of nodes at different layers as a function of time for four typical networks of fake news (*Fig.* 2*A for Weibo and* 2*C for Twitter*) and real news (*Fig.* 2*B for Weibo and* 2*D for Twitter*) are demonstrated. The fraction of re-postings in the first layer of fake news network is found significantly smaller than that of real news, while the fraction in other layers for fake news is significantly larger than that of real news. Early adopters re-posting the message shortly after the creator play a dominant role in circulating real news comparatively. These different roles lead to distinctive landscapes of propagation networks.

The investigation of layer sizes in propagation networks demonstrated in Fig. 2, are systematically extended to all the available messages. As shown in Fig. 3A and 3B, fake news networks tend to possess a relatively smaller first layer, while other layers are larger comparatively. Therefore, we can define the ratio of layer size as the ratio between the size of the second and the first layer. As shown in ratio distribution (Fig. 3C and 3D), the ratio in

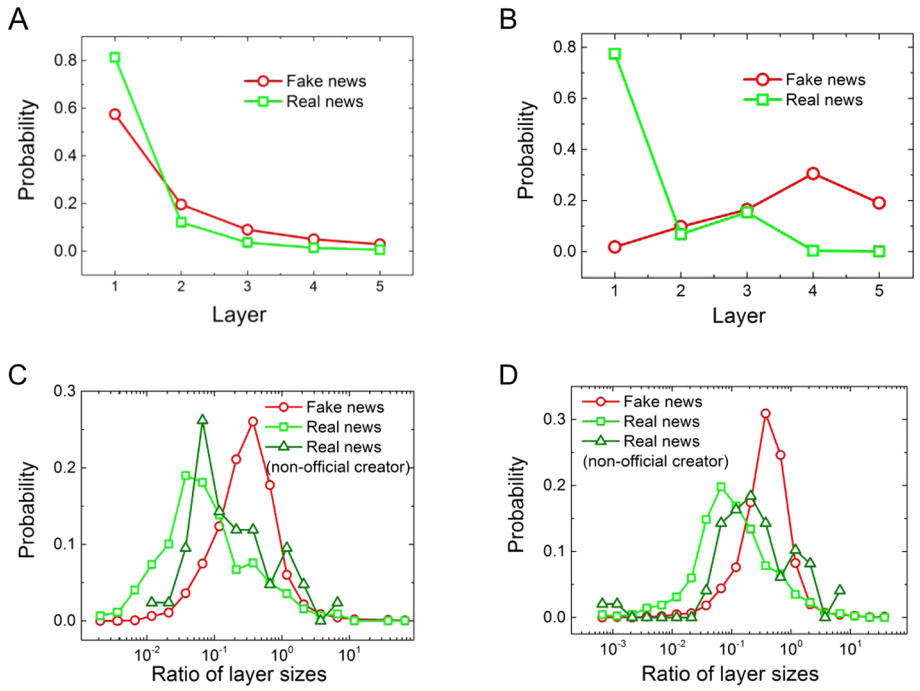

Figure 3 Ratio of layer sizes differentiates fake news from real news. The distribution of the ratio of layer sizes and its development after a period of time can differentiate fake news from real news. These differences appear already after a few hours. (**A**) The PDF of all re-postings in the first five layers averaged over all of the Weibo propagation networks. The *p*-value of Mann–Whitney is below 0.01. (**B**) The PDF of all re-postings in the first five layers in all networks of Twitter. The *p*-value of Mann–Whitney is below 0.01. (**C**) Distribution of the ratio of layer sizes at five hours from the first re-posting. The ratio of layer sizes is the size of the second layer divided by the size of the first layer. The *p*-value of Mann–Whitney between fake and real news is below 0.01. This figure considers all 1701 fake news, all 492 real news and 51 real news with non-official creators at five hours from the first re-posting. (**D**) Distribution of the ratio of layer sizes of all re-postings for the whole lifespan. The *p*-value between fake and real news is below 0.01. Here we consider all available Weibo propagation networks (all 1701 fake news, all 492 real news and 51 real news with non-official creators)

fake news is significantly larger than that of real news. The distribution for the ratio of layer sizes separates fake and real news well with only a small overlapped area. Furthermore, it is seen in Fig. 3C that this difference is already significant only at five hours since the first re-posting. In Fig. 3D, it is seen that, for the whole lifespan, the separation of the fake and the real is also significant. In the circulation of fake news, the success of the propagation depends highly on the branching process creating different layers, which show different evolution paths between fake and real news. We further investigate the probability difference between fake and real news based on distributions of layer ratio from the time of first re-posting (*Figs. S3 and S4*). Note that the layer size distribution has a peak around layer four on Twitter in Fig. 3B, probably due to secondary outbreaks.

It should be noted that real news is more likely to be created by official accounts such as government agencies or mass media agencies. In order to eliminate the possible effects of official creators, we also investigate the distribution of the ratio of layer sizes in real news from only non-official creators. While official news and non-official news have different sample sizes here, we found they both have different propagation patterns from fake news. For example, in Fig. 3C and 3D, the non-official real news and the fake news are found to have different distribution of layer size ratio. To verify our results, we also analyze data

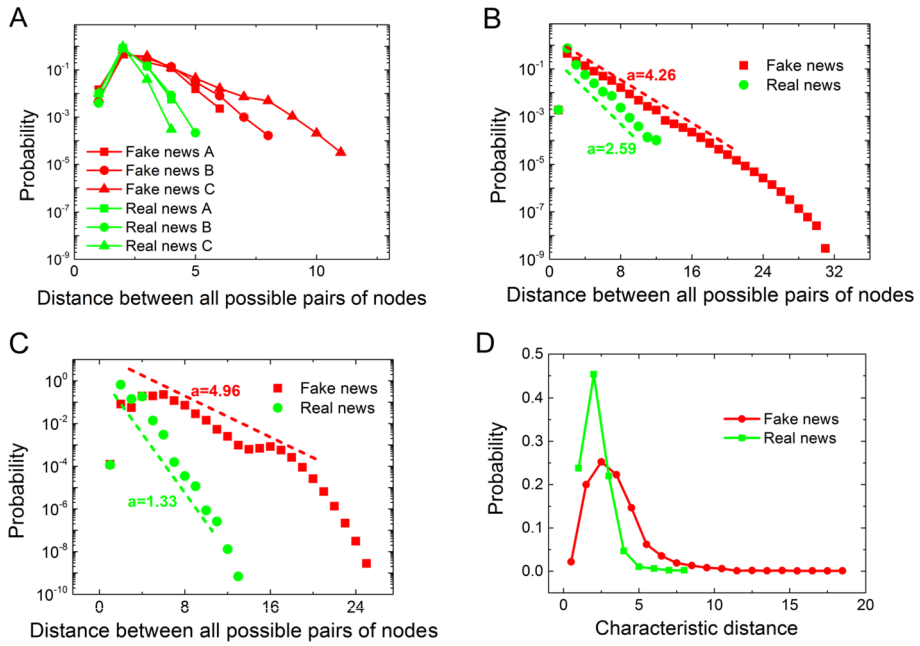

Figure 4 Characteristic distance differentiates fake news from real news. (**A**) The PDF of distances for three typical examples of networks for both fake and real news in Weibo. (**B**) The PDF of distances for all real and fake news networks in Weibo. (**C**) The PDF of distances of all real and fake news networks in Twitter. (**D**) The PDF of the characteristic distances (*details in Methods*) for fake news and real news. The *p*-value between fake and real news is below 0.01

of 2000 more real news from non-official accounts in a more recent dataset from 2016 to 2018 shown in Figs. S5 and S6. The distributions of this real news dataset are also distinct from that of fake news.

Characteristic distance. While the ratio of layer sizes can be regarded as a local feature of the network structure, we further inspect a global feature in terms of characteristic distance in a propagation network. As seen in Fig. 4A, distances between pairs of nodes in fake news are longer than those of real news, implying that later adopters foster the penetration of fake news in social networks. In order to quantify this finding for all the networks, we propose a second measure called characteristic distance (a) shown in Fig. 4B (*see Methods*). Considering the distance of all the networks as in Fig. 4B, fake news possesses a significantly longer characteristic distance (4.26) than that of real news (2.59). Similar results can also be observed in Twitter propagations (*Fig.* 4*C*). The distributions of characteristic distances for all networks are shown in Fig. 4D, where the two curves of fake and real news are well separated. Different from the results in [23], we show that the size distributions of fake and real news are similar (Fig. S7). This suggests that with similar levels of popularity, the characteristic distance is significantly different in fake news compared to real news. We also verified that the propagation size has less correlation with the characteristic distance (*Fig. S8*). To verify our results, we also analyze data of 2000 more real news from another dataset shown in Fig. S5.

Structural heterogeneity. Network topology describes the geometry of connections, with more information embedded than the scale statistics in [23]. Here we measure the Heterogeneity (*see Methods*) between propagation networks in fake and real news. The parameter h reflects the difference between a given propagation network and its counterpart of a star

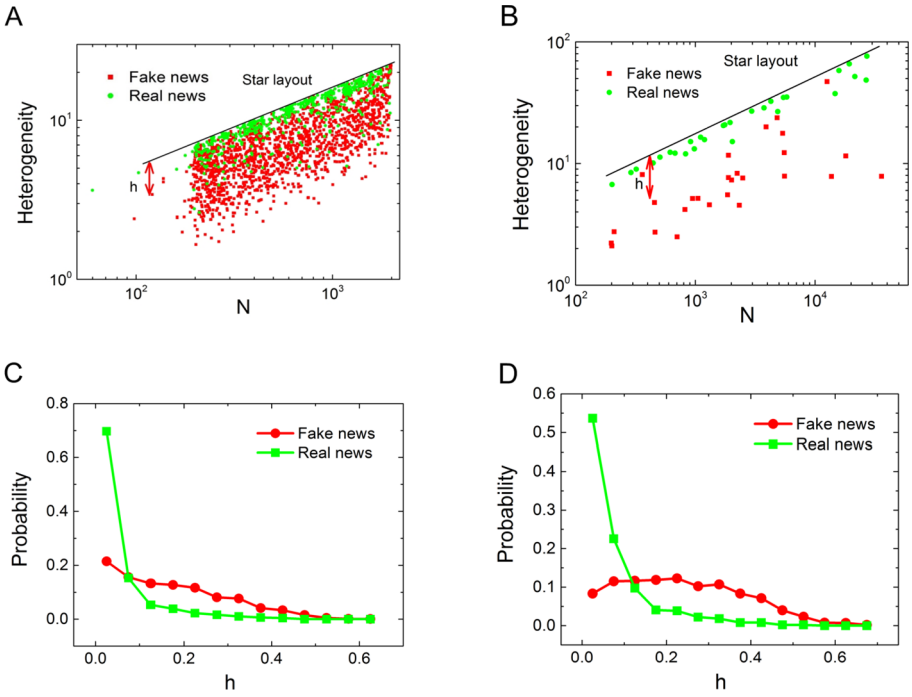

Figure 5 Heterogeneity measure for fake and real news in Weibo and Twitter. (**A**) The *x* axis is the size of the propagation network, and the *y* axis is the heterogeneity measure of the networks. The black line is the value of the star layout. The *h* is the difference of heterogeneity value between a real network and the corresponding value of star layout. (**B**) The scatter plot like in (**A**) for Twitter. (**C**) Distribution of *h* at five hours from the first re-posting of the Weibo propagation networks. The *p*-value here is below 0.01. (**D**) Distribution of *h* of all re-postings in Weibo for the whole lifespan. The *p*-value is below 0.01

network with the same-size. Network with smaller h means similar to a star network. Although the out-degree distribution demonstrates only a minor difference between fake news and real news (*Fig. S9*), it is interestingly found here that the topology heterogeneity is significantly distinguishable. Note that the relationship between heterogeneity and N for star networks is power-law as seen in Fig. 5A. The h is the difference between the logarithm of a real network heterogeneity value H_r and the logarithm of heterogeneity value of the same-size star network H_s. The parameter h of fake news is significantly larger compared to that of real news. Consistent findings can also be observed on Twitter (Fig. 5B). In order to quantify the heterogeneity systematically, two distributions of h considering different time intervals are calculated. In Fig. 5C, it shows a significant difference at five hours from the first re-posting. For the whole propagation lifespan in Fig. 5D, h of fake news is also significantly larger than that of real news. Fake news networks have typically lower heterogeneity (larger h) since their propagation involves few dominant broadcasters. On the contrary, real news demonstrates higher heterogeneity (smaller h) and a more star-like layout. The ability to distinguish fake news from real ones is also valid for real news posted by non-official users (*Fig. S10*). This implies that the indicator based on structural heterogeneity is independent of the creator type. Additionally, another measure named the Herfindahl–Hirschman Index (HHI [24]) shows also a distinction between fake news and real news (*Fig. S11*).

The distinction between fake and real news of the heterogeneity measure is the highest among the above three indicators as seen in Fig. 6 and Table 2. For a given Weibo network,

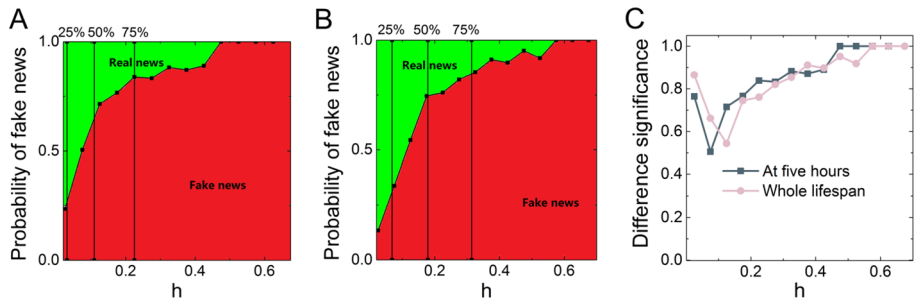

Figure 6 The heterogeneity measure shows a high difference between fake news and real news of Weibo in its early stage of propagation. (**A**) Probability of being fake news at five hours. The three vertical lines divide the figure into four parts with an equal number of networks. For example, the area on the left of the first left line has 25% of all the Weibo propagation networks. (**B**) Probability of being fake news for the whole lifespan. (**C**) The difference significance between fake news and real news

Table 2 Comparison between three methods

Measure	Difference significance of Weibo		
	Whole lifespan	Five hours	Non-official creator
Ratio of layer sizes	75.0%	74.4%	74.0%
Characteristic distance	74.6%	73.0%	70.1%
Heterogeneity (h)	78.2%	74.1%	71.3%

The Weibo networks are from the first Weibo datasets.

measuring its h provides a clear difference between fake news and real news, even only considering re-postings at five hours from the first re-posting (Fig. 6A). This identification becomes even sharper in Fig. 6B, when we consider all re-postings. We show in Fig. 6C the difference significance (*see Methods*) between fake news and real news for different h. The differences are about 76% and 79% respectively for re-postings at a relatively short time (five hours) and all re-postings. Note that the probability of being fake news at five hours is already very similar to that for the whole propagation lifespan. The verification analysis (shown in Figs. S5 and S6) also demonstrates the difference significance between fake news and real news from another dataset, which is fully published by non-official accounts. Our results suggest that even without sophisticated features like texts or user profiles, direct and understandable topological features can offer high significance for developing early detections.

Classifier. The three features mentioned above, namely the ratio of layer sizes, the characteristic distance, and the heterogeneity parameter could be used to create a Support Vector Machine (SVM) classifier. Here we divide the dataset into training set (60%) and test set (40%) ten times randomly. We find that the average accuracy of this classifier is 79.5% when applying the RBF kernel.

3 Discussion

Being the most vital and popular form of new media, online social networks, fundamentally enhance the creation and dissemination of fake news [25, 26]. Though existing solutions, especially the inspired machine learning approaches, perform impressively on targeting fake news, their black-box style essentially prevents a solid understanding and corresponding method development of debunking or blocking false information. On the

other way, the human-intensive labor approach is time-consuming and expensive. For example, it usually takes at least three days [4] for verification and therefore misses the optimal prevention window before massive spreading. In this sense, novel approaches that could help to identify fake news at early stages are urgently needed in preventing the negative impact of false information propagation on modern society.

We show here that fake news spread with very different network topology, even at early stages, from authentic messages. We focus, in this manuscript on the evolution differences between the propagation topology of two types of information at early stages rather than providing a comprehensive prediction approach [22]. Even taking only one feature, the difference between fake news and real news is significant. The propagation mechanism, which essentially couples information dynamics and collective cognition in social networks, results in a distinctive landscape of circulations between fake and real news. In this way, several early signals can be derived, including the layer-ratio, the characteristic distance and the heterogeneity. Varol et al. study early detection of promoted campaigns by using supervised machine learning, which contains features about diffusion patterns, content information, sentiment information, temporary signals, and user data [27]. Moreover, Vicario et al. study fake news by identifying polarizing content, which contains structural features, semantic features, user-based features and sentiment-based features [28]. In contrast, our suggested measures focus on structural features which are simple, without text analysis, and time efficient. For example, the weak heterogeneity of fake news might be the result of opinion competition from weak ties between social communities. As stated that "bad" is usually more influential than "good" [29], the unconsciousness of "negative-bias" might result in a late burst of fake news, which essentially differs from the spread of real ones. Disclosing intelligence factors that generate the specific topological features we found here can be a promising research direction in the future. Moreover, once we identify fake news, it is possible to study the nodes that participated in many networks. These nodes are much more active in the permeation of fake news, and as a result, they are more likely to be bots. The study of these vital nodes in the fake news propagation will play an important role in identifying and analyzing bots.

Note that our study has several major differences from Vosoughi et al. [23]. We focus more on the topological features (shape of a network), rather than on scale measures of propagation networks (depth or width). Furthermore, we focus on the largest cascade component of the propagation network, while all the cascade components are considered in [23]. As both manuscripts confirm the difference between fake news and real news in different aspects, we find surprisingly that this difference can be very significant even at the early stages of propagation.

4 Methods

Weibo data preprocessing. We analyze 1701 fake news of Weibo propagation networks (with 973,391 users) and 492 real news of Weibo propagation networks (with 347,401 users) that spread on Weibo from 2011 to 2016. We choose here large networks with more than 200 tweets. More details are given in Table 1. The topics of these Weibo propagation networks include political fake news, economic fake news, fraudulent fake news, tidbit fake news and pseudoscience fake news (*Fig. S12*).

Fake news is officially investigated and confirmed by the platform of Weibo [30]. Regarding real news, we collect them directly from reliable Weibo accounts. Creators of the

real news can be official accounts, for example, government accounts and on-line newspaper accounts. All these real news accounts have been officially verified by the platform of Weibo. On the other hand, we also select manually 51 out of 492 real news networks whose creators are not official accounts. To verify our results, we also analyze another dataset (2000 more recent real news) from Weibo in Figs. S5 and S6. These 2000 real news networks are from more recent records that has been collected in the same way as above, and from non-official accounts.

In order to create the network, in which nodes are users of Weibo and links are re-postings, we first mine the following data both for fake and real news:

(a) Users: the unique serial number of users who participate in the same network. We also mark the node of the network creator.
(b) Re-postings: the unique serial number of directed re-posting activities, and the serial number of source users and reposted users of this re-posting.

Twitter data preprocessing. Twitter data was collected from Japanese tweets posted during the period between March 11th and March 17th in 2011, which is the Great East Japan earthquake period. During this period, a lot of fake news propagated on Japanese Twitter.

After gathering fake and real news tweets on a keyword basis, we focused on those with more than 200 tweets to create a retweet network. Here we define screen names as nodes, which appeared in the tweet context, and links are mention signs "@" between the author of the tweet and screen names after the sign. This is because many fake retweet users have already deleted their tweet or account itself, and do not appear in the database. Deleting the tweet or account makes the network more segregated and more challenging to capture the real structure of the networks. To avoid network segregation, we use the above-mentioned context-based method to create retweet networks. Furthermore, as of March in 2011, many Japanese Twitter users did not clearly distinguish between mention symbol "@" and clear retweet symbol "RT @". Note that if there are multiple "@" in one tweet, according to the above rules, we extracted multiple screen names as nodes and linked them in order from the beginning of the sentence to create the networks. We compared two types of networks defined by mention symbol and retweet symbol in Fig. S13, and found our major results still hold.

After creating networks, we extract the largest connected component (LCC) without consideration of link directions and analyze only those with LCC size above roughly 200 nodes. A node with the oldest tweet time in LCC was treated as creators. All the fake and real news that we determined are shown in Additional file 1.

Our method of creating a retweet network is different from the way of previous literature [20, 23] that used follower graphs and tweet data simultaneously to create a retweet network. In case that we do not have a follower graph as of 2011, we applied this approximate method of extracting as much information as possible from the tweet context. In principle, because retweet information remains in the tweet context, the topology of the network should be equivalent to the previous literature, but the time information in resolution of seconds is not accurate in our case. Therefore, we only use time information in hours in the Twitter analysis.

Definition of fake news and real news. In a recent paper by Lazer et al. [31], "fake news" is defined as fabricated information that mimics news media content in the form, without news media's editorial norms and processes for ensuring the accuracy and credibility of the information. In our manuscript with Weibo data, the fake news is false information

fact-checked by the platform and verified as having been fabricated. Regarding real news, we collect them directly from reliable Weibo accounts. And all these real news accounts have been officially verified by the platform of Weibo.

For Twitter data, the fake news is also false information which is fact-checked by reliable evidences [32–34]. This is similar as the true/false news defined in paper by Vosoughi et al. [23] that their rumor cascades are checked independently by six fact-checking organizations. However, since there were no official anti-rumor website in Japan as of 2011, we first gathered 57 topics listed on websites [32, 33] and a book [34]. These contents include tweets based on no evidence and malicious tweets, such as starvation of babies and elderly people, someone under the server rack needed help, and the Japan prime minister is taking luxury supper during the disaster. When collecting tweets, we combine a few keywords related to the contents of each fake news. These keywords were proper nouns, such as place names and personal names. After that, we excluded correction tweets whose contents are against fake news including keywords such as "false" and "mistake". Our typical procedure to gather fake news tweets is explained in a previous work [2]. To validate the fake news tweets, three graduate students at the University of Tsukuba checked independently whether these topics are fake and the gathered tweets are properly classified into fake news.

For real news in Twitter, we gathered 71 topics by combining keywords (proper nouns, such as place names and personal names) as with the fake news. We collected most of tweets originated from official accounts with verified Twitter badges such as government agencies, major newspapers and famous people. The contents included tweets about earthquake information, traffic information, donation information and so on. In addition, we also collected five topics originated from civilians without badges, which were widely retweeted. These tweet contents were related to small correct tips during the disaster.

Establishing a network model. Based on the information we analyze above, we establish a directed network as demonstrated in Fig. 1A. The users are the nodes in the network, and the re-postings are the edges in the network. And we the mark network creator using color green. Each edge has a direction that is either from creator to re-poster or from former re-poster to later re-poster. We plot figures of typical networks for both fake and real news of Weibo and Twitter as shown in Fig. 1B to 1E.

Ratio of layer sizes. The layer number is defined as the number of hops from the creator to a given re-poster. The ratio of layer sizes is a measure for each network defined as:

$$\text{ratio of layer sizes} = \frac{n_2}{n_1}, \tag{1}$$

n_1 and n_2 are the sizes (number of nodes) of the first and second layer for a certain network respectively.

Characteristic distances. In order to measure the distances, for each network we first calculate the distances between all pairs of nodes in the network and plot the distribution in a logarithmic scale (y axis). It can be seen from Fig. 4 that the function can be approximated by an exponential function. We consider the linear part of curves where their x value (distance) is above one. We calculate the characteristic distance (a) accordingly:

$$y \sim e^{-\frac{x}{a}+b}. \tag{2}$$

Heterogeneity measure. The heterogeneity [35] is defined as:

$$\text{Heterogeneity} = \frac{\sqrt{\langle k^2 \rangle}}{\langle k \rangle} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} k_i^2}}{\frac{1}{N}\sum_{i=1}^{N} k_i}, \tag{3}$$

N: The number of nodes in the network,

k_i: The degree of node i.

We show a scatter plot (Fig. 5A) for both fake and real news of Weibo. The black line is the theoretical line for star network:

$$\text{Heterogeneous} \sim \sqrt{N}. \tag{4}$$

The h is the difference between the logarithm of a real network heterogeneity value H_r and the logarithm of heterogeneity value of the same-size star network H_s as shown below:

$$h = \log(H_s) - \log(H_r). \tag{5}$$

Probability of being fake news. Here we use the ratio of layer sizes as an example. We divide the ratio of layer sizes into n portions. In the ith portion, the probability of being fake news is:

$$p = \frac{p_i^f}{p_i^f + p_i^r}, \tag{6}$$

p_i^f: The probability of fake news in the ith portion (the number of fake news in this portion divided by the total number of fake news).

p_i^r: The probability of real news in the ith portion.

Significance of difference. When we distinguish fake news from real ones using topological measures such as the ratio of layer sizes or the characteristic distance, it is important to know the significance of the difference. Here we use the ratio of layer sizes as an example. First, we rank the Weibo propagation networks by their ratio of layer sizes ignoring their types (fake or real). Second, we randomly split these propagation networks into n portions that have the same number of networks. Finally, we calculate the difference significance using the following formula:

$$Q = \frac{1}{n}\sum_{1}^{n} \frac{\max(p_i^r, p_i^f)}{p_i^r + p_i^f}, \tag{7}$$

n: The number of portions that we divide.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-020-00224-z.

Additional file 1. Supplementary Information (DOCX 1.3 MB)

Acknowledgements
SH thanks the Israel Science Foundation, ONR, the Israel Ministry of Science and Technology (MOST) with the Italy Ministry of Foreign Affairs, BSF-NSF, MOST with the Japan Science and Technology Agency, the BIU Center for Research in Applied Cryptography and Cyber Security, and DTRA (Grant no. HDTRA-1-10-1-0014) for financial support. JZ was supported by NSFC (No. 71871006) and the National Key Research and Development Program of China (No. 2016QY01W0205). YS was supported by by JSPS KAKENHI Grand Number 17K12783. HT and MT are supported by JST Strategic International Collaborative Research Program (SICORP) on the topic of "ICT for a Resilient Society" by Japan and Israel. JW was partially supported by the National Key R&D Program of China (2019YFB2101804), the National Special Program on Innovation Methodologies (SQ2019IM4910001), and the National Natural Science Foundation of China (71531001, 71725002, U1636210). We also thank Jiali Gao for providing new dataset of real news.

Funding
Not applicable.

Abbreviations
SVM, Support vector machine; RBF, Radial basis function (Gaussian) kernel; LCC, largest connected component.

Availability of data and materials
Our data is provided by author Jichang Zhao and will be available from him based on reasonable request.

Competing interests
The authors declare no competing interests.

Authors' contributions
DL designed the research. ZZ, JZ, YS and OL contributed equally to this paper. ZZ and YS performed data calculation. OL, ZZ, SH and DL wrote the paper. Other authors have analyzed the results and revised the paper. All authors read and approved the final manuscript.

Author details
[1]School of Reliability and Systems Engineering, Beihang University, Beijing, China. [2]National Key Laboratory of Science and Technology on Reliability and Environmental Engineering, Beijing, China. [3]School of Economics and Management, Beihang University, Beijing, China. [4]Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Japan. [5]Department of Physics, Bar-Ilan University, Ramat Gan, Israel. [6]Sony Computer Science Laboratories, Tokyo, Japan. [7]Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan. [8]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 June 2019 Accepted: 25 February 2020 Published online: 03 April 2020

References

1. Schmidt AL, Zollo F, Del VM et al (2017) Anatomy of news consumption on Facebook. Proc Natl Acad Sci USA 114(12):3035
2. Takayasu M, Sato K, Sano Y, Yamada K, Miura W, Takayasu H (2015) Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study. PLoS ONE 10(4):e0121443
3. A BuzzFeed news of hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis?utm_term=.glr1n5VYr#.kaJBYd4a8
4. Fact-checking fake news on Facebook works—just too slowly. https://phys.org/news/2017-10-fact-checking-fake-news-facebook-.html#jCp (2018.1.23 accessed)
5. Daley DJ, Kendall DG (1964) Epidemics and rumours. Nature 204(4963):1118
6. Pastorsatorras R, Vespignani A (2000) Epidemic spreading in scale-free networks. Phys Rev Lett 86(14):3200
7. Eguíluz VM, Klemm K (2002) Epidemic threshold in structured scale-free networks. Phys Rev Lett 89(10):108701
8. Newman MEJ (2002) Spread of epidemic disease on networks. Phys Rev E, Stat Nonlinear Soft Matter Phys 66(1 Pt 2):016128
9. Moreno Y, Pastor-Satorras R, Vespignani A (2002) Epidemic outbreaks in complex heterogeneous networks. Eur Phys J B 26(4):521–529
10. Barthélemy M, Barrat A, Pastor-Satorras R et al (2004) Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. Phys Rev Lett 92(17):178701
11. Zhou T, Liu JG, Bai WJ et al (2006) Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity. Phys Rev E, Stat Nonlinear Soft Matter Phys 74(5 Pt 2):056109
12. Kuperman M, Abramson G (2000) Small world effect in an epidemiological model. Phys Rev Lett 86(13):2909–2912
13. Yang F, Liu Y, Yu X et al (2012) Automatic detection of rumor on Sina Weibo. In: ACM, pp 1–7
14. Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: can we trust what we RT? In: Social media analytics, SOMA, KDD workshop, pp 71–79
15. Ma J, Gao W, Wei Z et al (2015) Detect rumors using time series of social context information on microblogging websites. In: ACM international on conference on information and knowledge management. ACM, New York, pp 1751–1754
16. Zheng H, Xue M, Lu H et al (2017) Smoke screener or straight shooter: detecting elite sybil attacks in user-review social networks. arXiv preprint. arXiv:1709.06916

17. Castillo C, Mendoza M, Poblete B (2011) Information credibility on Twitter. In: International conference on World Wide Web, WWW 2011, Hyderabad, India, March 28–April, DBLP, pp 675–684
18. Qazvinian V, Rosengren E, Radev DR et al (2011) Rumor has it: identifying misinformation in microblogs. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 1589–1599
19. Zollo F, Bessi A, Del VM et al (2017) Debunking in a world of tribes. PLoS ONE 12(7):e0181821
20. Kwon S, Cha M, Jung K et al (2014) Prominent features of rumor propagation in online social media. In: IEEE, international conference on data mining. IEEE, pp 1103–1108
21. Wu K, Yang S, Zhu KQ (2015) False rumors detection on Sina Weibo by propagation structures. In: IEEE, international conference on data engineering. IEEE, pp 651–662
22. Vosoughi S (2015) Automatic detection and verification of rumors on Twitter. Ph.D. thesis, Massachusetts Institute of Technology
23. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151
24. Rhoades SA (1993) The Herfindahl–Hirschman index. Fed Reserve Bull 79:188
25. Ma R (2008) Spread of SARS and war-related rumors through new media in China. Commun Q 56(4):376–391
26. Chua AYK, Banerjee S (2017) A study of tweet veracity to separate rumors from counter-rumors. In: Proceedings of the 8th international conference on social media & society, pp 1–8
27. Varol O, Ferrara E, Menczer F et al (2017) Early detection of promoted campaigns on social media. EPJ Data Sci 6(1):13
28. Del Vicario M, Quattrociocchi W, Scala A et al (2018) Polarization and fake news: early warning of potential misinformation targets
29. Baumeister RF, Bratslavsky E, Finkenauer C et al (2001) Bad is stronger than good. Rev Gen Psychol 5(4):477–509
30. Weibo official web page for fake news reporting. http://service.account.weibo.com (2018.1.23 accessed)
31. Lazer D, Baum MA, Benkler Y et al (2018) The science of fake news. Science 359(6380):1094
32. Matsunaga H Social psychology at the time of panic which classified and organized 80 hoaxes after the earthquake. http://blogos.com/article/2530/ (in Japanese) April 8th 2011 (2018.1.20 accessed)
33. Ogiue C (2011) Validation: rumor and hoax during the Great East Japan Earthquake, Kobunsha, Japan (in Japanese)
34. Ishizawa Y, Akamine T Time series analysis of "hoax information" diffused on Twitter during earthquake. https://sites.google.com/site/prj311/event/presentation-session/presentation-session4#TOC-Twitter-2 (in Japanese) (2018.11.27 accessed)
35. Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Syst Biol 1(1):24

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com