frontiers in Physics

ORIGINAL RESEARCH
published: 17 April 2019
doi: 10.3389/fphy.2019.00058

Topical Alignment in Online Social Systems

Felipe Maciel Cardoso[1,2]*, *Sandro Meloni*[2,3], *André Santanchè*[1] and *Yamir Moreno*[2,4,5]

[1] Institute of Computing, University of Campinas, Campinas, Brazil, [2] Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, Zaragoza, Spain, [3] IFISC, Institute for Cross-Disciplinary Physics and Complex Systems (CSIC-UIB), Palma de Mallorca, Spain, [4] Department of Theoretical Physics, University of Zaragoza, Zaragoza, Spain, [5] ISI Foundation, Turin, Italy

Understanding the dynamics of social interactions is crucial to comprehend human behavior. The emergence of online social media has enabled access to data regarding people relationships at a large scale. Twitter, specifically, is an information oriented network, with users sharing and consuming information. In this work, we study whether users tend to be in contact with people interested in similar topics, i.e., if they are topically aligned. To do so, we propose an approach based on the use of hashtags to extract information topics from Twitter messages and model users' interests. Our results show that, on average, users are connected with other users similar to them. Furthermore, we show that topical alignment provides interesting information that can eventually allow inferring users' connectivity. Our work, besides providing a way to assess the topical similarity of users, quantifies topical alignment among individuals, contributing to a better understanding of how complex social systems are structured.

Keywords: social network analysis, topical similarity, data analysis, computational social science, twitter, information networks

OPEN ACCESS

Edited by:
Matjaž Perc,
University of Maribor, Slovenia

Reviewed by:
Marco Alberto Javarone,
Coventry University, United Kingdom
Francesca Lipari,
Libera Universitá Maria SS. Assunta, Italy

***Correspondence:**
Felipe Maciel Cardoso
fmacielcardoso@gmail.com

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *24 January 2019*
Accepted: *27 March 2019*
Published: *17 April 2019*

Citation:
Cardoso FM, Meloni S, Santanchè A and Moreno Y (2019) Topical Alignment in Online Social Systems. Front. Phys. 7:58. doi: 10.3389/fphy.2019.00058

1. INTRODUCTION

Relationships among people determine the structure of complex social systems. As such, the emergence and widespread use of online social networking sites have allowed to address a number of questions related to how humans connect among each other. Research using data from online social media have, in turn, produced new methods and models that are at the core of present (computational) social sciences. In this work, we explore the relationships between users of the microblogging service Twitter and the information shared by them. Information sharing is a very important aspect of Twitter, which is also considered as an information network [1], i.e., it is often a means for the consumption and sharing of contents that are mainly diffused through users' connections. The way Twitter—and other social networks—works leads to an interesting linkage between information and (often adaptive or dynamic) relationships among individuals, which is the focus of our investigation.

In what follows, we inspect how much the information shared by users is related to their connections in the social network. Our goal is to demonstrate that the information spread in Twitter is a crucial component of social dynamics through the verification of topical alignment. Connected users being topically aligned is an indication of how much homogeneity is pervasive across their dimensions of interests and ideas. Accordingly, this requires a proper assessment of the information topics, since focusing only on individual annotations does not capture the latent "context" in which users are engaged when exchanging messages. We infer topics from clusters

of highly associated *hashtags* in messages exchanged by users. This allows us to capture topics exposing latent higher-level semantic entities without the need of an external ontology or manual classification step [2–4].

Figure 1 shows examples of topics we detected in our Twitter data set. Users' affiliation to them indicates individual preferences in the wide range of topics available in the social network and it constitute our pool to assess similarity between users. The engagement in specific topics tells something about a user, and we adopt them as the basis to create a metric based on users' interests in different topics. In short, we want to assess if connected users tend to be topically similar and how much the similarity is relevant to their relationships.

Our results show that, on Twitter, *follow* and *mention* relationships are more likely to have a higher topical alignment than random pairs of users. Furthermore, we verify that both kinds of relationships tend to display a similar alignment pattern, despite the belief that they are relationships of a different kind [5]. Finally, our analysis also shows that connections with strong interactions tend to have higher similarity and that the similarity between connected users indicates a higher probability of interaction.

2. RELATED WORK

In an online social system, the emergence of connections among individuals can be explained by different mechanisms from the preferential attachment [6] to shortcuts for the consumption of information [7]. It is clear that the information shared in an online social network is an important characteristic to be taken into account while analyzing its connections. However, there is no clear definition of information in a social network context. In this work, we consider information as the different kinds of content that flow in a network and may affect people's opinions or ideas. This is analogous to the Bateson's general definition of information as composed of pieces that are supposed to be *"a difference that makes a difference"* [8]. Some recent efforts have been directed to the study of how the information traversing the network is related to its links. Weng et al. [7] recently demonstrated that information flows play an important role in link creation in the Yahoo! Meme network. Around 12% of the new edges were motivated by the information flow, indicating that the network's edges dynamics cannot be explained merely by its topological structure. Furthermore, they showed that, while some users create connections mostly based on friendship, others are more guided by the content that users produce and share. Bogdanov et al. provide a model of pre-specified topics and verified the consistency of their use by Twitter users, they also applied this to predict influencers and to minimize the latency in information dissemination [4]. Meyers et al. [9] were interested in how the rise of abrupt changes in the information flow dynamics influences the creation and removal of links. Their work found that cascade of tweets was likely to cause *follow* or *unfollow* bursts, i.e., people start to follow or unfollow others with the abrupt increase in the retweets of some content.

Also using data from Twitter, Das et al. [10] studied how the difference in users intent affects the content of their messages and their propagation. Suh et al. [11] focused on which features increase the probability of a message to be retweeted finding that the presence of hashtags (i.e., the presence of a context), along with other factors, favors the sharing. Following on how context and interests shape information sharing, Wu et al.[12] categorized influential users in Twitter—i.e., celebrities, media, and bloggers—finding that usually users in the same category show common behaviors that differ from one category to another. Another contribution along this line is the one by Kang and Lerman [13] where they studied how the position in the network and the engagement of users affect the information they receive. The authors found that more engaged users usually occupy bridge positions in the network and are exposed to more diverse and novel information with respect to less engaged ones. Finally, the role of network structure and access to information has also been studied by Aral and Van Alstyne [14] analyzing data from an executive recruitment firm.

2.1. Topical Alignment

We are concerned with the degree to which users are more topically aligned with their connections. This is closely related to the homophily concept [15–19], i.e., the tendency of individuals to form dyads with people similar to them, which have implications for the final network structure [20]. If the similarity between pairs of individuals induces them to form a tie, this tendency is called *choice homophily*, otherwise, if it is just a result of the constraints in the opportunities of connections, *induced homophily*. Both types might be necessary to explain levels of similarity encountered in dyads, as Kossinets et al. showed with dyads of a university community [18]. Nonetheless, choice homophily requires assessing individuals preferences and often this is infeasible. Thus, the concepts of *baseline homophily*, the expected similarity between random pairs of individuals, and *inbreeding homophily*, the similarity of dyads that are above or under the baseline, introduced by McPherson and Smith-Lovin [15] are often used in practical approaches [21].

Topically aligned dyads are not necessarily a result of homophily, as connected individuals also tend to become more similar to each other over time, what is know as *social influence*, or social contagion [22–24]. Social influence is an important ingredient for synthetic models such as the one proposed by Robert Axelrod [25] and is also verified in social networks [26]. However, real data also show that some effects attributed to social contagion may be a result of homophily [23]. Furthermore, the creation of dyads may be motivated by latent or by unknown characteristics as pointed by Shalizi et al., thus, it might be impossible to verify whether ties similarity is really the result of homophily or social influence [24]. This infeasibility in disentangling both processes does not affect our work, as we are not interested in verifying which one is driving the similarity of the dyads. Our goal is to assess to which degree connected users are topically similar, independently of the generating mechanism.

FIGURE 1 | Topics. Word clouds for hashtags of six selected topics we detected in our Twitter data set. Hashtags' size in each figure is proportional to their degree in the co-occurrence subgraph of their community. See the text for further details.

Nonetheless, we consider that the works more related to our own were strictly interested in homophily in online networks. Laniado et al. [27] inspected gender homophily—i.e., the prevalence of same-gender relationships—in the Tuenti Spanish social network. They based their analysis on self-reported gender data and their results showed the presence of gender homophily in dyadic and triadic relationships. Aiello et al. [28] explored homophily in the context of tagging social networks (Flickr, Last.fm, and aNobii). In these networks, tags are used to classify resources—a different usage than hashtags on Twitter. In their approach, tags employed by the users are used to compute their similarity, which quantifies their proximity in tags usage. They found that users topical similarity is related to their shortest path distance on the social graph and that it could predict some links on the graph. Crandall [26] explored homophily using datasets extracted from Wikipedia and LiveJournal—article and blogging based networks—and modeled users according to their articles editing history. Choudhury explored homophily over a set of demographic users characteristics and its relation to the structure of their ego-network, most importantly they showed that the presence of homophily concerning topical interests is independent of the ego network structure [29].

Some of these works are more related to networks centered in some kind of digital artifact, e.g., image, article, etc. Twitter, however, is more centered on the information posted by its users. Furthermore, hashtags or other features, by themselves, are not sufficient to assess similarity among users as they do not fully capture the context of users' messages. Thus, despite their findings, these works leave aside the latent semantics in the information sharing. Others had to rely on an external tool or specific classification to measure users similarity [29–31]. It is necessary to look at a higher granularity to capture the different kinds of content that users are engaged with, which we achieve using topics of information. To the best of our knowledge, no study has explored the topic in this way. Thereby, our work contributes to the understanding of the nature of relationships in a social network exploring a component still hard to be

manipulated: the different kinds of information that traverse the network.

3. METHODS AND DATA

3.1. Twitter Dataset

There is no clear definition of social media or online social network, however, there is a general consensus that services like Twitter are instances of social media services [32]. Due to its microblogging nature, some consider Twitter also as a news media or an information network [1, 33]. This is an important feature as we are interested in the content shared by the users and their relationships. We explore both **mentions**, mentioning a user in a tweet, and **follow**, subscribing to receive other user tweets, relationships in this work. In our analyses, we explicitly decided not to include **retweets** as we are more interested in information created by the users than shared information. Explicitly creating a new tweet supposes a larger effort than retweeting one, thus we believe that this is a more reliable proxy of users real interests with respect to retweets. Moreover, not considering retweets has also the side effect of limiting the number of bots in our datasets. As suggested in Ferrara et al. [34] the majority of contents produced by bots are retweets. So excluding them and users with only retweets should reduce the number of bots in our analyses. Finally, the last interaction form present nowadays in Twitter—**quoted** tweets—was not present in 2013 when we started our data collection. Thus, we do not consider quoted tweets in this work.

Our dataset is composed by all the geo-localized tweets—tweets with valid GPS coordinates—located in the United Kingdom and Ireland in a 7 months period from January to September 2013 through the Twitter's Streaming API [1]. Further, more tweets of the users with geo-localized tweets and their follow/friend connections were obtained through the REST public API [2]. The decision of focusing only on geo-localized tweets, although could partially limit our results, has several advantages. Firstly, it guarantees that the vast majority of the tweets are in the same language. Secondly, it allows us to focus more on discussions between users and local events rather than global events that are more likely to be influenced by other media sources. The final dataset, excluding retweets has 98 million tweets from January 18th to September 2nd, 2013.

3.2. Topics of Information

Information in Twitter flows through tweets, which are short messages with a highly dynamic vocabulary, encumbering traditional text clustering techniques. We decided to build topics of information considering the tweets with hashtags, as they are indicators of the tweet content. Hashtags are users generated annotations containing a shared meaning, similar to acronyms generated organically by a population [35] Furthermore, it is common for users to insert more than one hashtag in a tweet, and we exploit this aspect to build a semantic mapping of information in Twitter. We assume the existence of a semantic association between hashtags that co-occur in the same tweet. This is analogous to the assumption that words are semantically associated if they are likely to co-occur frequently [36]. Thus, our method focuses only on the implicit semantics given by Twitter messages, i.e., it does not consider explicit semantics given by other sources. This semantic mapping is captured by a weighted co-occurrence graph of hashtags, which we built by extracting all pairs of hashtags that co-occurred in each tweet in our dataset. Therefore, in this graph, an edge (h_i, h_j) indicates that the hashtags h_i and h_j co-occurred and, as the graph is weighted, $w(h_i, h_j)$ gives the number of different tweets in which they are both present.

We built a hashtag weighted co-occurrence graph using the 16,935,625 tweets with hashtags belonging to our dataset. As we removed hashtags that did not co-occur with any other, the co-occurrence graph resulted in 2,090,971 from the total of 4,320,429 distinct hashtags. As noted before, the edges of this graph represent a semantic association between hashtags. In order to further restrict our analysis to cases in which the statistics is not very scarce, and to reduce possible noise coming from low co-occurrences which might not have a clear significant association, we additionally removed all the edges between pairs of hashtags that co-occurred in less than 3 tweets. This process produces our final co-occurrence graph, which includes 104,308 hashtags and 526,522 edges.

We consider that topics of information are sets of hashtags clustered together in the graph. Thus, we expect that they will reflect the higher level structures that emerge from the latent semantic association of hashtags, providing the different contexts to which messages refer to. It is natural to see that these clusters could be captured by a community detection method and we decided to use the OSLOM tool [37]. OSLOM is able to capture overlapping communities, a desirable feature considering that one hashtag may be used in different contexts.The application of OSLOM resulted in 2,074 communities and 14,118 homeless nodes, i.e., hashtags that did not belong to any community. We considered the communities and the homeless nodes as topics. Despite the latter possibly not significantly benefiting our future procedures, we believe that a hashtag alone can also carry information. Furthermore, our method to assess topical similarity should not be affected by this increase of topics as it does not take into consideration the topics that are not shared by two users (see the **Supplementary Information** for more details). Summing up both communities and homeless nodes in our analysis we consider a total of 16,192 topics with an average of 622 users per topic.

This approach of building a co-occurrence graph and using a community detection method to find topics was also used by Weng and Menczer [38] through the Louvain method [39], although they were not concerned with topical alignment. They assumed, based on the topical locality assumption, that semantically similar hashtags would appear in tweets together. Notwithstanding the resemblance to our premises, we do not presume that hashtags are similar, only semantically associated. Even though there is not an easy way to ground the accuracy of this approach, we believe that it is a sound method for assessing information topics. Its premises and

[1] https://dev.twitter.com/streaming/overview, accessed in September 2016
[2] https://dev.twitter.com/rest/public, accessed in September 2016

procedures are well-defined over the semantic associations of hashtags.

3.3. Users Dataset

Users considered in our analysis had, at least, one tweet with a hashtag in order to assess which topics of information they were affiliated with. Thus, we selected the 774,596 users from the 1 million of users with tweets. Before starting the analysis, we also took particular care in reducing the number of bots in our dataset. Along with excluding retweets we also decided to remove users that have been active for less than one day and those who showed an unusual activity. Specifically, we excluded users that had, on average, more than 400 tweets per day, as we consider that it is normally unfeasible for a real person to produce this quantity of tweets (for more information on bot filtering and their possible impact see the **Supplementary Information**). Finally, users had to have, at least, one hashtag belonging to the topics detected (described in the previous section), leading to a final set of 608,899 users. We name this set *Population* as it includes all the users in our experiment.

After that, we extracted from the entire population another set of 9, 490 users, which we define as *central users*. Those users are the core of our analyses as we calculate the topical similarity between them and their direct connections and compare it against random users selected from the entire population. Central users have been extracted randomly from the users in our dataset that have been active for the entire 7 months data collection period and produced at least 10^2 tweets to guarantee a large corpus of tweets about their interests. Details for the two sets are shown in **Table 1**.

3.4. Users Representation

Each user is represented by a feature vector **u**, which comprises her affiliation to all topics of information. The process of building a user vector is illustrated in **Figure 2**. Feature u_i corresponds to her affiliation in topic i and its value represents the number of hashtags belonging to t_i(the set of hashtags belonging to the topic i) that were used by the user in her tweets. As the communities obtained by OSLOM may overlap, the same hashtag may be computed in more than one feature. In this case, each hashtag adds a proportional value to each feature it belongs . The value of a feature u_i is given by

$$u_i \leftarrow \sum_{\{h \in H\,:\,h \in t_i\}} \frac{m_U(h)}{|\{t \in T : h \in t\}|} \tag{1}$$

All the hashtags used by a user are contained in a multiset $U = (H, m_U)$, wherein H is the set of used hashtags and m_U gives the number of occurrences of each hashtag. T is the set of topics, i.e., communities of hashtags. Strictly speaking, each element $t \in T$ stands for a topic and it is a set containing the hashtags inside one cluster built by the community detection method. **Figure 2** illustrates a user multiset and its transformation in the user feature vector via Equation (1). As *#love* appears in the topics t_1 and t_2, it adds 1 to their respective features.

3.4.1. Weighting Users' Vectors

The previous definition of users' features vector considers that all topics have the same weight, i.e., the values of the respective features are directly derived from the number of hashtags used. This may be not suitable for our task as some popular topics or of general use could be over-represented and thus should have a smaller weight. To overcome this distortion, we consider that topics shared by a large percentage of the users ought to have a small weight, likewise, topics possessed by only a small percentage of users ought to weight more. The intuition behind this is that features corresponding to rare topics should be more discriminative of the topical proximity of users than features corresponding to frequent topics.

Strictly speaking, we would like to take into account the information content of each topic [40]. To do so, we rely on TF-IDF [36] to weight users affiliation to each topic u_i following:

$$u_i \leftarrow u_i \times log\frac{|I|}{|\{v \in I : v_i > 0\}|} \tag{2}$$

where I is the set of all individuals, i.e., Twitter users. For each feature i in the user vector, this method will weigh its value according to the number of users that also used it—e.g., a feature that is shared by all users will have its value set to 0 as it does not provide information to discriminate users.

3.5. Computing Similarity Between Users

With the representation of users as feature vectors, we are able to compute topical similarity between two users using as metric the cosine similarity of their vectors [36]. The cosine similarity fits well to this task as it only focuses on the angle between vectors—i.e., it does not consider their length. Cosine similarity ranges from 0 to 1; identical users would have similarity 1; users that do not share anything in common 0. It is evaluated using Equation (3) below. In preliminary analyses, we also tested Kendall's tau, Spearman's rho and Jaccard similarity measures. We did not adopt them as they did not present significant differences or improvements with respect to cosine similarity.

$$sim_{cos}(u, v) = \frac{u \cdot v}{||u||||v||} \tag{3}$$

TABLE 1 | Summary of crawled data.

Data	Raw
Tweets	98,506,315
Tweets with Hashtags	16,935,625
Distinct Hashtags	4,320,429
Users with Tweets	1,286,816
Users with Hashtags	774,596
Population set	608,899
Central Users set	9,490

*For a further description of users activity, see the **Supplementary Information**.*

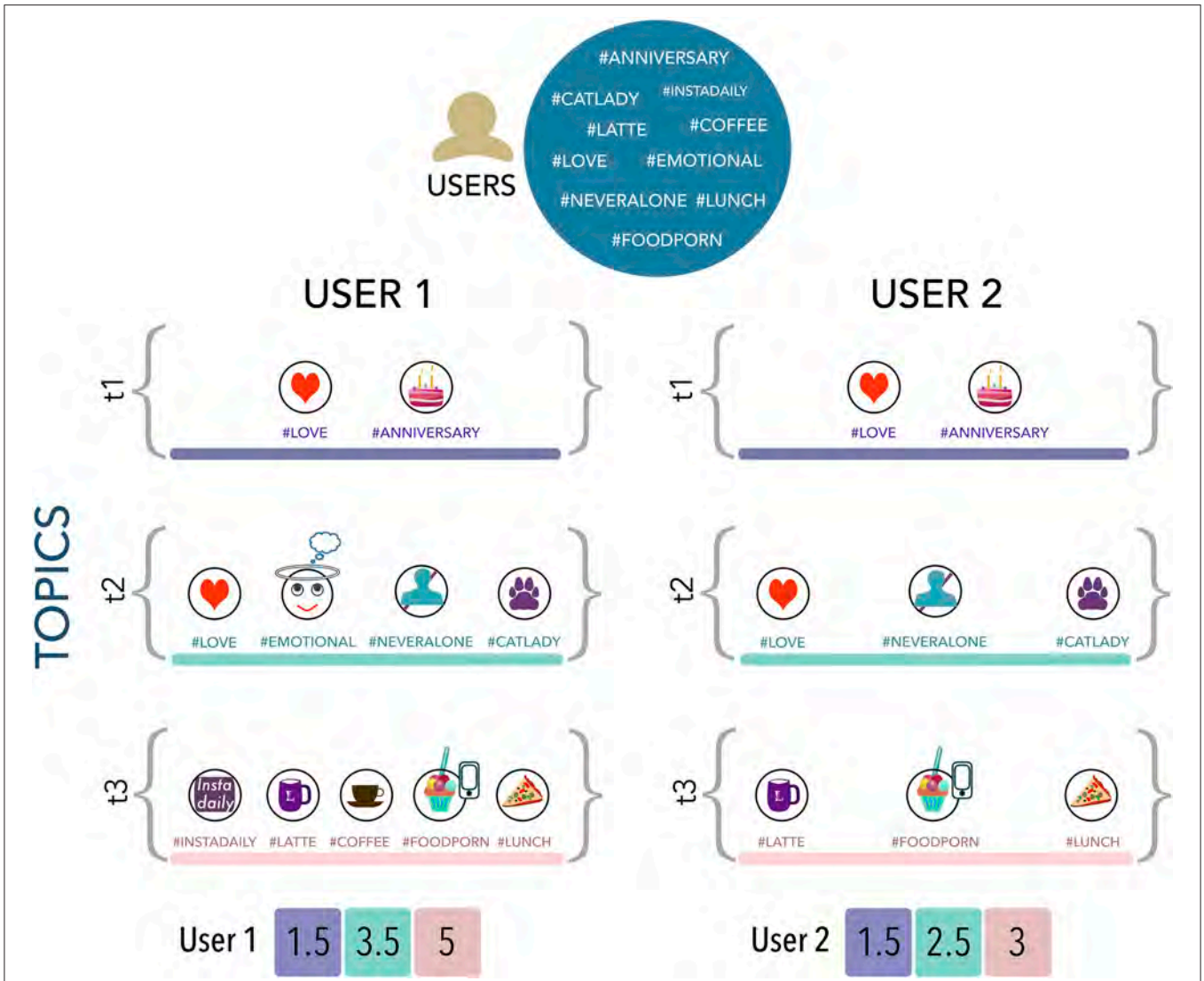

FIGURE 2 | Extraction of a user feature vector from hashtags in different topics. The figure shows an example of how the feature vector is computed for two users that are part of the datasets. Hashtags shown are a small sample of those contained in the whole dataset and have been selected only to illustrate the example depicted.

4. TOPICAL ALIGNMENT

The hypothesis that users are more topically aligned to their neighbors than to random users will be addressed here in terms of *baseline alignment* and *inbreeding alignment* similar to the classification introduced by McPherson and Smith-Lovin [15]. Here, we consider **baseline alignment** as the expected average similarity between users and a random group of the population. Inbreeding alignment is defined as the difference between the baseline distribution and the distribution of average similarity between the users and those with whom they form a dyad, which is formed by a *follow* or *mention* relationship. In other words, baseline alignment is our null model and inbreeding alignment a measure of how much real values deviate from the null model. This deviation is captured by the Kolmogorov-Smirnov test [41] and the likelihood of the distribution of dyads yielding higher (or lower) values of average similarity is captured by a Mann-Whitney *U*-test [42, 43]. We believe this approach has significant benefits than just looking at the hashtags shared by users as we comment on the **Supplementary Information** material.

4.1. Topically Aligned Follow Relationships

We initially explore inbreeding alignment with respect to *follow* connections. Our hypothesis is that users are, on average, more similar with their followees, i.e., we expect their topical alignment to be significant. This means that the distribution of similarity averages of the individuals with their followees is expected to yield higher values than the distribution of averages with randomly chosen individuals from the population. We tested this hypothesis using the central users and their followees.

Figure 3 shows the histograms of the two distributions: *Followees*, the distribution of averages computed for each central

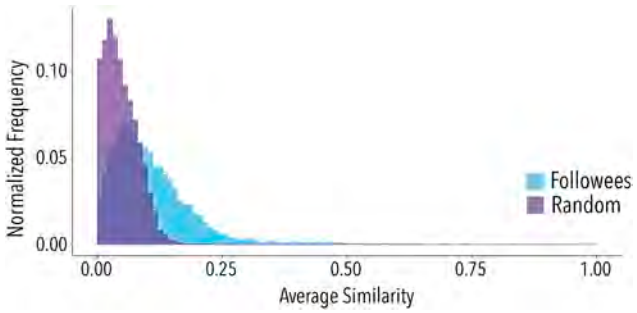

FIGURE 3 | Distribution of average similarity between central users and their followees (blue) and between central users and randomly selected groups of the same size (purple). KS(Kolmogorov-Smirnov statistic) = 0.37, $p < 0.001$, MW(Mann-Whitney *U* effect size) = 0.75, $p < 0.001$ The medians of the distributions of *Followees* and the *Random* are 0.087 and 0.041, respectively. Distributions have been calculated considering the whole set of 9,490 central users.

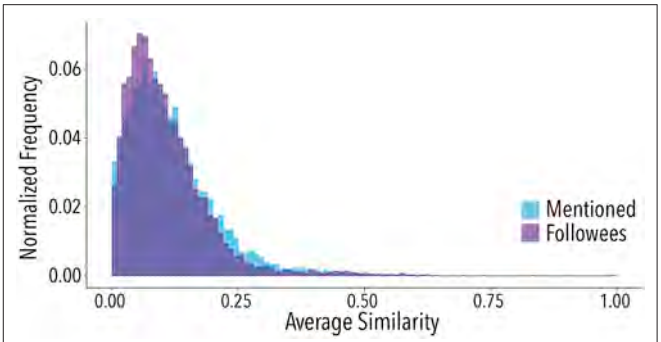

FIGURE 4 | Distributions of average similarity between users followed (purple) and mentioned (blue) by central users. KS = 0.06, $p < 0.001$. Distributions have been calculated considering the whole set of 9,490 central users.

user with her followees; and *Random*, the distribution wherein, for each central user, averages have been computed with a group composed of randomly chosen users from the *Population* set, with the same size as the set of central user followees. As it can be seen, all the distributions are centered around low values of the cosine similarity spectrum. We consider that this effect is a result of the large number of topics and does not impact our results.

There is an overlap among the distributions, mostly concentrated in lower similarities. However, it is clear that there is a difference between the random distribution and the followees distribution. The Kolmogorov-Smirnov statistics between the distributions is 0.37, $p < 0.001$. We also used the Mann-Whitney U test to verify if the distribution with followees was likely to have a higher average similarity than the other. Results were positive with an effect size of 0.75, $p < 0.001$. Overall, the analysis shows that, on average, users tend to be connected to whom they are more similar with, that is, the similarity between followees is higher than the baseline similarity, thus showing the presence of inbreeding alignment. This implies that a user tends to have a stronger topical similarity with followees than with randomly chosen users.

4.2. Users Interactions

Users on Twitter can use the convention *@username* to mention another user in a tweet. The interactions that happen through mentions are often seen as a relationship stronger than the *follow* connections [44]. One hypothesis that emerges from such affirmation is that the topical similarity between mentioned users tends to be higher than between followed users. To test this hypothesis, we verified if the distribution of similarity averages with the mentioned users tended to be concentrated in higher values of similarity than the same distribution for followees. As shown by **Figure 4**, the distributions are roughly the same. Thus, in this context, we cannot say that the *mention* relations are more topically aligned than the connections with followed users.

Both *mentions* and *followees* histograms show that most of the averages fall into low values of similarity and there is a positive skewness of the two distributions, that is not evident in the distributions with random users (**Figure 3**). Given the proximity between the two distributions presented in **Figure 4**, users on average might follow and mention others in a close similarity pattern. This hypothesis is verified in **Figure 5**, which indicates that users that tend to follow similar users, also tend to mention similar users.

4.3. Reciprocity of Relationships

Relationships in Twitter are not reciprocal, a user following another does not imply that the other will choose to follow back. Thus, the existence of reciprocity indicates a stronger relationship between two users as both decided to establish this bond. In the scope of this work, the relationship strength is also viewed in terms of the topical similarity, thus, we expect that reciprocal dyads have a higher similarity than non-reciprocal dyads. This was verified for both *mention* and *follow* relationships, i.e., relationships wherein the two users mentioned each other and relationships in which the two follow each other. We first present the result regarding the reciprocity of the *follow* connections in **Figure 6A**. The two distributions differ, the distribution of similarity for the reciprocal followees is concentrated around higher values of similarity. The comparison for the reciprocal mentions distribution is shown in **Figure 6B**. The distribution of reciprocal mentions also has a higher similarity. This indicates that reciprocal relations are more prone to have a higher topical similarity, i.e., users have a more similar topic affiliation if they have a reciprocal relationship.

The tests conducted in this subsection reinforce what was seen in the previous section: there is no significant difference between the nature of *mention* and *follow* relationships with respect to topical similarity. The distributions of both relationships are very alike when considering the dyads similarity, even with reciprocal relationships. Furthermore, we could verify that, in the case of reciprocal relationships, there is a higher topical alignment than with non-reciprocal relationships. This indicates that users with a reciprocal relationship tend to become more similar by social influence or, conversely, that users similarity can be a factor which influences both to establish the relationship. Our method is unable to discriminate between either of the two mechanisms,

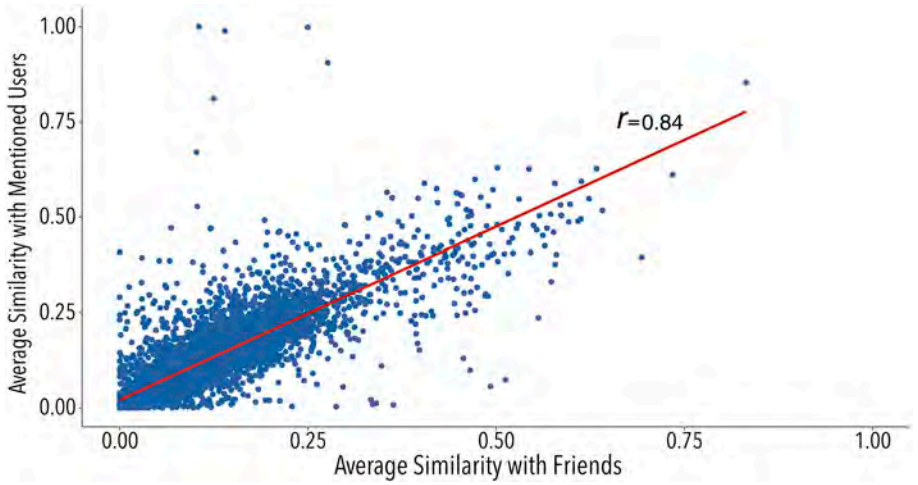

FIGURE 5 | Correlation between average similarity with followees and mentioned users. Each point corresponds to the average similarity between a central user and the users she follows and the average similarity between the central user and the users mentioned by her. The Pearson correlation between the two variables is 0.84.

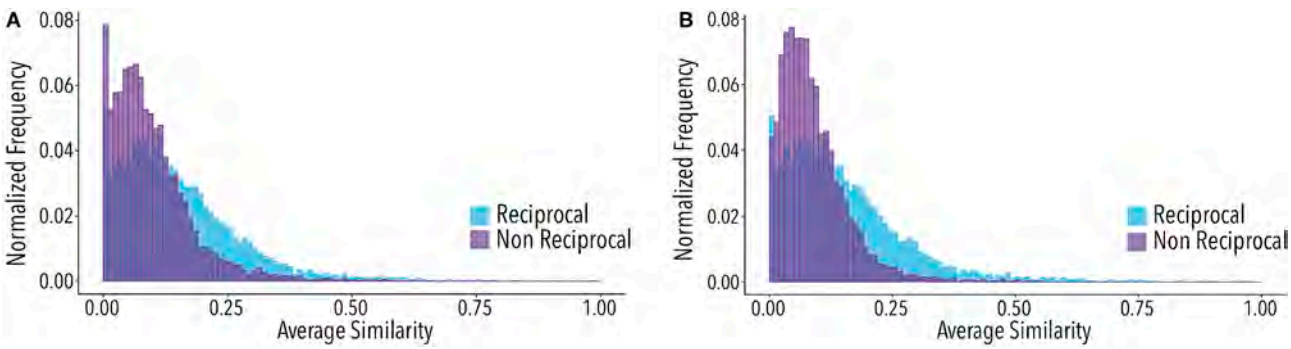

FIGURE 6 | (A) Distribution of average similarity between central users and reciprocal (blue) and non reciprocal (purple) followees. KS = 0.27, $p < 0.001$, MW = 0.66, $p < 0.001$. Medians of the distributions of reciprocal and of nonreciprocal relationships are 0.12 and 0.07, respectively. Distributions have been calculated considering only the 5,872 central users that have both reciprocal and non reciprocal followees in the dataset. **(B)** Distribution of average similarity between central users and reciprocal (blue) and non reciprocal (purple) mentions. KS = 0.22, $p < 0.001$, MW = 0.64, $p < 0.001$. The median similarity of the distribution of nonreciprocal mentions is 0.08 while for reciprocal mentions is 0.12. Distributions have been calculated considering only the 8,663 central users that have both reciprocal and non reciprocal mentions in the dataset.

as we would need to add a temporal dimension to the evolution of similarity and the network structure.

4.4. Mention Probability

All the analyses shown until now indicate that the similarity of most of the dyads is concentrated around low values. Therefore, it is natural to presume that most of the mentions made by central users involve users with low similarity with them. However, this contrasts with common sense as we expect that users in dyads with high similarity are more likely to be mentioned.

We explored this question, i.e., if the probability of being mentioned is higher for users with a high similarity, by looking at all dyads of followees. We also took into account the number of times that each followee was mentioned by a central user. To do so, we first defined $m_{u,v}$ as the number of mentions made by central user u to followee v and $s_{u,v}$ as their similarity. Then we calculated $P(m_{u,v} > M | s_{u,v} \leq S)$ as the conditional probability of a user being mentioned more than M times, given that her similarity with the mentioning user is smaller than S:

$$P(m_{u,v} > M | s_{u,v} \leq S) = \frac{P(m_{u,v} > M \cap s_{u,v} \leq S)}{P(s_{u,v} \leq S)} \quad (4)$$

Figure 7 shows the cumulative conditional probabilities of followees being mentioned by central users more than $M = 0$, 2, 5, and 10 times, given their similarity with the central user. As expected, the probability decreases when the minimum number of mentions increases. **Figure 7** also shows that followees which have low similarity with central users do not have a higher probability of being mentioned. Actually, it is observed a stable growth until 0.4 and, after that, all the curves reach a plateau. Overall, the pattern of conditional probabilities appears to be the

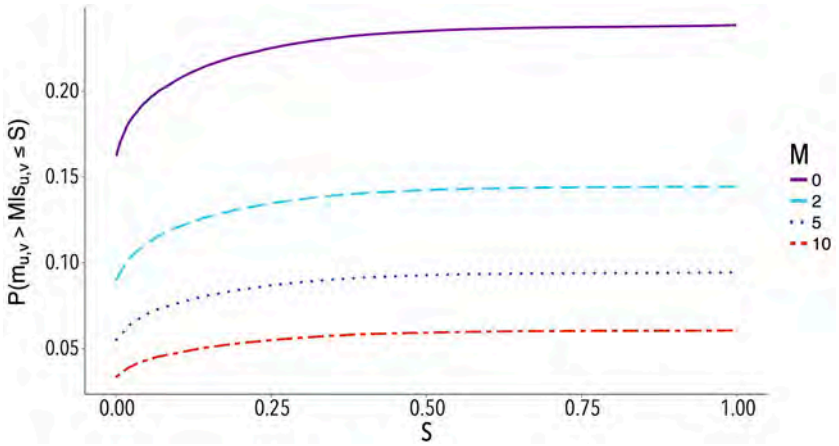

FIGURE 7 | Conditional probability of followees being mentioned more than M times by central users, given their similarity is smaller than S. The probability has been calculated using 547, 346 dyads involving connected users.

same for larger values of M, there is only a shift in the probability, as being mentioned more times is more challenging.

This analysis shows how the similarity gives an indication of the interactions inside connections, at least for some values of similarity. As more similar is the connected users, the higher is their probability to have interacted.

4.5. Inference by Similarity

There is a correlation between users average similarity with followees and mentioned users. It indicates that users, on average, follow and mention other users in a similar fashion with respect to topical similarity. However, until now, we did not provide a way to verify to which degree similarity among users is an indicator of their connections. In other words, we would like to know if topical similarity might be an effective way to predict relationships between users. Our question here is the following: is it possible to infer a user's followees or mentions from a group of randomly selected users looking only at the similarity between them?

Using pools of users of different sizes, we try to extract from them all the connections of a central user considering their relative similarities. In this case, a pool always contains all the user's followees user mixed with other randomly selected users from the entire population. To create pools of different sizes we use a multiplicative factor k. The size of a pool is given by $k \times |fr(u)|$ where $fr(u)$ is the set of followees of user u and $|fr(u)|$ is its cardinality. Thus, with $k = 1$ the pool only contains u's followees; for $k = 2$ the pool will be constituted by all of u's followees and the same number of random users. With $k = 3$ we have all u's followees and twice random users and so on until we reached pools of 60 or 80 times the size of the original set.

Once we created the pools, the similarity between central user u and all the users in the pool is computed. After that, a set with the same size of $fr(u)$ and containing the users that were most similar to u is returned. Finally, this set is compared with the original set of followees of user u. To quantify the effectiveness of this method we calculated the average PPV(positive predictive values) for all the central users, i.e., the average of the fraction of followees that were correctly predicted. As previously mentioned, there are differences between users' average similarity, that suggest the presence of different following patterns. Thus, we repeated our analysis for users with different values of similarity, e.g., 0, 0.2, 0.4, and 0.6, along with considering all the central users together.

Results are shown in **Figure 8**. Each line shows the averages for each group of users. The blue line shows the average PPV considering all users together. To quantify the performance of this method against random selection, the red curve represents the average PPV if users were selected randomly instead of resting on similarity. For all the groups, our method outperforms random selection, indicating that similarity is an important feature in users connection process.

Results for an average similarity of 0.4 and 0.6 are worthy of a deeper analysis as the PPV remains roughly constant (or declines very slowly) for a wide range of values of k. This plateau in PPV means that even with an increasing set of users to choose from, the method keeps returning a significant fraction of their followees. This happens because they continue to be the most similar available in the whole pool. We believe that this is due to the fact that topics' affiliation patterns are almost unique for some dyads, hence, the majority of other users in the pool does not have a larger similarity than the actual followees of the user.

Even if the results for an average similarity of 0.4 and 0.6 are quite remarkable in terms of the match between inferred and real followees, the results obtained considering all the users together is not too good. Nonetheless, it is important to notice that the method applied here does not take into consideration the whole social network structure, which is likely the main factor responsible for determining connections. Our focus is to explore the relation between information and users' relationships, not

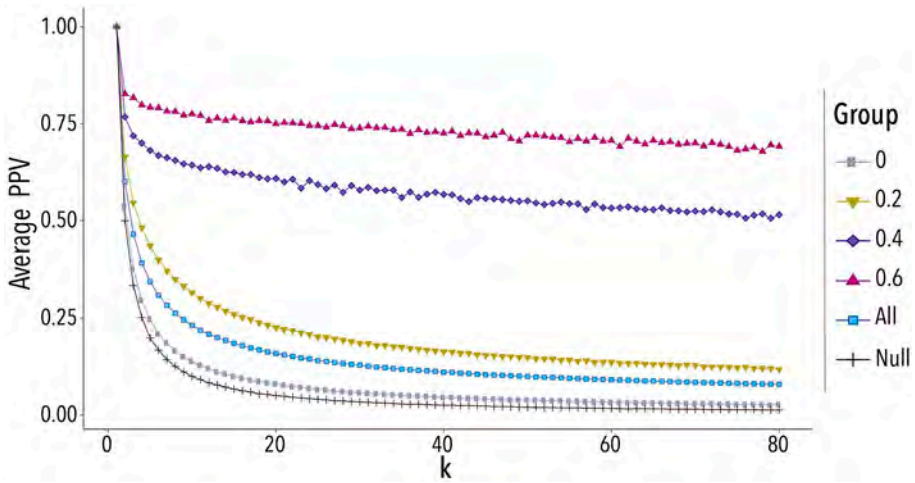

FIGURE 8 | Average PPV(positive predictive values) of the inference mechanism for following connections with a pool of size $fr(u) \times k$ for different values of average similarity. The *All* curve presents all the central users irrespectively of their average similarity while the *Null* represents sets of randomly selected users.

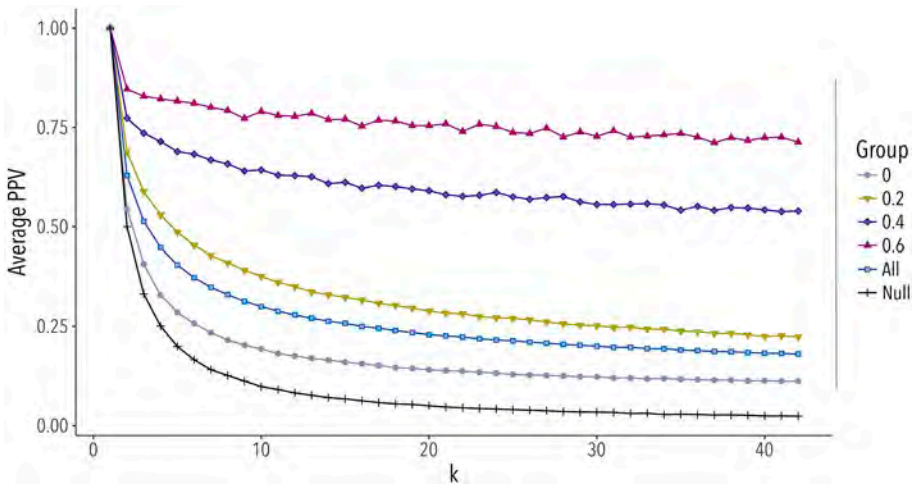

FIGURE 9 | Average PPV(positive predictive values) of the inference mechanism for mentioning relations with a pool of size $fr(u) \times k$ for different values of average similarity. The *All* curve presents all the central users irrespectively of their average similarity while the *Null* represents sets of randomly selected users.

to provide a complete algorithm for link prediction or recommendation. Having said that, we, however, believe that our results show that users affiliation in topics can be an important feature to be taken into account in link prediction or recommendation algorithms.

We repeated the process done for following relations considering, in this case, the probability of mentioning another user. In this case, we verified whether we could infer if a central user mentioned another user only looking at the similarity between them. Results are shown in **Figure 9** and are quite similar to the ones for the following probability with, in some cases, a better performance. This once again reinforces the idea that, in the case of topical alignment, following and mentioning interactions show a similar behavior and highlights the importance that topical similarity might have for some users.

5. CONCLUSIONS

In today's world, online social networks as Twitter provide a laboratory where information and users connections are available for study. In this work, we analyzed how the pair-to-pair structure of a social network is related to the information shared on it. Connections in a social network are the substrate over which information flows, which makes their flow partially dictated by the network structure. However, information flow cannot be seen as an independent phenomenon; its contents can affect how individuals behave. For instance, people might be inclined to bond with others following the affinity in the information they share. On the other hand, information shared by an individual can make other users less prone to establish a bond with her. We have explored this relation using Twitter's information and connection data demonstrating that individuals

which have a relationship tend to be more similar than expected regarding the information they share, i.e., connected users tend to be topically aligned.

On the other hand, in order to investigate how information is coupled with social connections, a key point is to design a model which captures its desired characteristics. We achieve this by modeling information as semantic topics of hashtags as Weng et al. [38]. These topics encompass contents of information shared among users. We computed users affiliation in topics to characterize individuals' interests and preferences on Twitter. This characterization served as a basis for the exploration of topical similarity between individuals and we found that, on average, individuals are more likely to have a relationship with more similar users. For some users this effect is so profound that they are essentially connected to the users most similar to them in all our dataset, which suggests an effective way to predict new connections at least for a subset of individuals in the network.

We have also verified if the influence of topical similarity between individuals differed in *mentions* and *follows* relations. Our results show a consistency across the two types of relationships, showing no significant difference between them. This was also verified when considering reciprocal relationships, which, in both cases, showed a higher level of similarity than non-reciprocal ones.

The approach presented in this work uses hashtags to build information topics. This limited our results to users that used hashtags, which significantly reduced our sample. Moreover, as we did not have the whole Twitter network structure, our hypothesis was restricted to exploring dyads and could not explore questions involving network measures, such as distance and centrality. Additionally, considering only geo-localized tweets further reduced the size of our datasets. Nonetheless, we believe that our sample provides a significant support to understand some relationships among users. There is also the possibility to improve our method to build topics, which currently ignores the temporal behavior of hashtags. The moment in which hashtags co-occur might contain specificities that we were not able to capture. However, even with these limitations, we could verify that the topics detected have a semantic sense and our datasets were sufficiently large as to achieve statistically relevance.

Our work demonstrates the importance of topical similarity between users regarding their connections and interactions. Our contribution also provides a feasible computational way to compute the similarity between users and can be used to further explore homophily and social influence in a social network. This can be further enhanced to improve our understanding of the mechanisms by which users connect, analyzing the whole social network structure, which was not available to us. Furthermore, it is necessary to further investigate how the flow of information is related to network dynamics. Our results also leave open opportunities to explore how topics' semantics affect the behavior of users who adopt them. Other possibilities include using our method in applications for link recommendation or finding missing links in social networks.

ETHICS STATEMENT

All data used in this work have been obtained using the Twitter public API. We adhered to the Spanish Law for personal data protection, which does not require obtaining permission from an Ethical Committee to use public and anonymized Twitter data. We also confirmed that we followed Twitter's terms and conditions when conducting this study.

AUTHOR CONTRIBUTIONS

FC, SM, AS, and YM designed research. FC and SM performed research. FC analyzed data. FC, SM, and YM discussed the results. FC, SM, AS, and YM wrote the paper.

FUNDING

FC and AS acknowledges support from Microsoft, Santander, CAPES, CNPq, and FAPESP Project 2015/01587-0. SM acknowledges support from the Ramón y Cajal Program by MINECO, Spain. YM and SM acknowledge support from the Government of Aragón, Spain through a grant to the group FENOL, by MINECO and FEDER funds (grant FIS2017-87519-P) and by the European Commission FET-Proactive Project Multiplex (grant 317532). SM also acknowledge the Spanish State Research Agency, through the María de Maeztu Program for Units of Excellence in R&D (MDM-2017-0711).

ACKNOWLEDGMENTS

A pre-print version of this work is available at http://arxiv.org/abs/1707.06525 [45].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2019.00058/full#supplementary-material

REFERENCES

1. Myers S, Sharma A, Gupta P, Lin J. Information Network or Social Network? The Structure of the Twitter Follow Graph. In: *WWW'14 Companion*. Seoul (2014). p. 493–8. doi: 10.1145/2567948.2576939
2. Lehmann J, Gonçalves B, Ramasco JJ, Cattuto C. Dynamical classes of collective attention in twitter. In: *Proceedings of the 21st International Conference on World Wide Web - WWW '12*. New York, NY: ACM Press (2012). p. 251.
3. Lee K, Palsetia D, Narayanan R, Patwary MMA, Agrawal A, Choudhary A. Twitter Trending Topic Classification. In: *2011 IEEE 11th International Conference on Data Mining Workshops*. Washington, DC: IEEE (2011). p. 251–8.

4. Bogdanov P, Busch M, Moehlis J, Singh AK, Szymanski BK. The social media genome. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*. Niagara, ON (2013). p. 236–42. doi: 10.1145/2492517.2492621
5. González-Bailón S, Wang N, Rivero A, Borge-Holthoefer J, Moreno Y. Assessing the bias in samples of large online networks. *Soc Netw*. (2014) **38**:16–27. doi: 10.1016/j.socnet.2014.01.004
6. Barabási AL, Albert R. Emergence of scaling in random networks. *Science*. (1999) **286**:509–12. doi: 10.1126/science.286.5439.509
7. Weng L, Ratkiewicz J, Perra N, Gonçalves B, Castillo C, Bonchi F, et al. The role of information diffusion in the evolution of social networks. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13*. New York, NY: ACM (2013). p. 356–64.
8. Foxman D, Bateson G. Steps to an ecology of mind. *Western Polit Q*. (1973) **26**:345. doi: 10.2307/446833
9. Myers SA, Leskovec J. The bursty dynamics of the Twitter information network. In: *Proceedings of the 23rd International Conference on World Wide Web - WWW '14*. New York, NY: ACM Press (2014). p. 913–24.
10. Das A, Gollapudi S, Kiciman E, Varol O. Information dissemination in heterogeneous-intent networks. In: *Proceedings of the 8th ACM Conference on Web Science. WebSci '16*. New York, NY: ACM (2016). p. 259–68.
11. Suh B, Hong L, Pirolli P, Chi EH. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter Network. In: *2010 IEEE Second International Conference on Social Computing*. Washington, DC (2010). p. 177–84.
12. Wu S, Hofman JM, Mason WA, Watts DJ. Who says what to whom on Twitter. In: *Proceedings of the 20th International Conference on World Wide Web. WWW '11*. New York, NY: ACM (2011). p. 705–14.
13. Kang JH, Lerman K. Effort mediates access to information in online social networks. *ACM Trans Web*. (2017) **11**:1–19. doi: 10.1145/2990506
14. Aral S, Alstyne MV. The diversity-bandwidth trade-off. *Am J Sociol*. (2011) **117**:90–171. doi: 10.1086/661238
15. Mcpherson M, Smith-Lovin L, Cook J. Birds of a feather: homophily in social networks. *Annu Rev Sociol*. (2001) **27**:415–44. doi: 10.1146/annurev.soc.27.1.415
16. McPherson JM, Smith-Lovin L. Homophily in voluntary organizations: status distance and the composition of face-to-face groups. *Am Sociol Rev*. (1987) **52**:370. doi: 10.2307/2095356
17. Newman MEJ. The structure and function of complex networks. *SIAM Rev*. (2003) **45**:167–256. doi: 10.1137/S003614450342480
18. Kossinets G, Watts DJ. Origins of homophily in an evolving social network. *Am J Sociol*. (2009) **115**:405–50. doi: 10.1086/599247
19. Lazarsfeld PF, Merton RK. Friendship as a social process: a substantive and methodological analysis. In: Berger M, Abel T, Page C, editors. *Freedom and Control in Modern Society*. New York, NY: Van Nostrand (1954). p. 18–66.
20. Javarone MA, Armano G. Perception of similarity: a model for social network dynamics. *J Phys A Math Theor*. (2013) **46**:455102. doi: 10.1088/1751-8113/46/45/455102/meta
21. Robinson DT, Aikens L, Aikens L. Homophily. In: Levine JM, Hogg MA, editors. *Encyclopedia of Group Processes {&} Intergroup Relations*. Thousand Oaks, CA: SAGE Publications, Inc. (2009). p. 404–7.
22. Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. *Stat Med*. (2013) **32**:556–77. doi: 10.1002/sim.5408
23. Aral S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc Natl Acad Sci USA*. (2009) **106**:21544–9. doi: 10.1073/pnas.0908800106
24. Shalizi CR, Thomas AC. Homophily and contagion are generically confounded in observational social network studies. *Sociol Methods Res*. (2011) **40**:211–39. doi: 10.1177/0049124111404820
25. Axelrod R. The dissemination of culture: a model with local convergence and global polarization. *J Conflict Resol*. (1997) **41**:203–26. doi: 10.1177/0022002797041002001
26. Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S. Feedback effects between similarity and social influence in online communities. In: *Proc. KDD'08*. New York, NY: ACM Press (2008). p. 160.
27. Laniado D, Volkovich Y, Kappler K, Kaltenbrunner A. Gender homophily in online dyadic and triadic relationships. *EPJ Data Sci*. (2016) **5**:19. doi: 10.1140/epjds/s13688-016-0080-6
28. Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F. Friendship prediction and homophily in social media. *ACM Trans Web*. (2012) **6**:1–33. doi: 10.1145/2180861.2180866
29. Choudhury MD. Tie formation on Twitter: homophily and structure of egocentric networks. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. Boston, MA (2011). p. 465–70.
30. Halberstam Y, Knight B. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *J Public Econ*. (2016) **143**:73–88. doi: 10.1016/j.jpubeco.2016.08.011
31. Kang J, Lerman K. Using lists to measure homophily on twitter. In: *AAAI Workshop on Intelligent Techniques for Web*. Toronto, ON (2012). p. 26–32.
32. Obar JA, Wildman S. Social media definition and the governance challenge: an introduction to the special issue. *Telecommun Policy*. (2015) **39**:745–50. doi: 10.1016/j.telpol.2015.07.014
33. Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In: *Proceedings of the 19th International Conference on World Wide Web. WWW '10*. New York, NY: ACM (2010). p. 591–600.
34. Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The rise of social bots. *Commun ACM*. (2016) **59**:96–104. doi: 10.1145/2818717
35. Javarone MA, Armano G. Emergence of acronyms in a community of language users. *Eur Phys J B*. (2013) **86**:474. doi: 10.1140/epjb/e2013-40662-5
36. Turney PD, Pantel P. From frequency to meaning: vector space models of semantics. *J Artif Int Res*. (2010) **37**:141–88. doi: 10.1613/jair.2934
37. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S. Finding statistically significant communities in networks. *PLoS ONE*. (2011) **6**:e18961. doi: 10.1371/journal.pone.0018961
38. Weng L, Menczer F. Topicality and impact in social media: diverse messages, focused messengers. *PLoS ONE*. (2015) **10**:e0118410. doi: 10.1371/journal.pone.0118410
39. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech Theor Exp*. (2008) **2008**:P10008. doi: 10.1088/1742-5468/2008/10/P10008/meta
40. Cover TM, Thomas JA. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Hoboken, NJ: Wiley-Interscience (2006).
41. Conover WJ. *Practical Nonparametric Statistics. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Hoboken, NJ: Wiley (1980).
42. Rice JA. *Mathematical Statistics and Data Analysis*. No. p. 3 in Advanced series. Boston, MA: Cengage Learning (2006).
43. McGraw KO, Wong SP. A common language effect size statistic. *Psychol Bull*. (1992) **111**:361–5. doi: 10.1037/0033-2909.111.2.361
44. Romero DM, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics. In: *Proceedings of the 20th International Conference on World Wide Web - WWW '11*. New York, NY: ACM Press (2011). p. 695.
45. Cardoso FM, Meloni S, Santanche A, Moreno Y. Topical alignment in online social systems. *arXiv*:1707.06525.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.