Online Social Networks and Media 3–4 (2017) 22–31

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Automatic controversy detection in social media: A content-independent motif-based approach

Mauro Coletto[a,b,*], Kiran Garimella[c], Aristides Gionis[c], Claudio Lucchese[a,b]

[a] Ca' Foscari University, Venice, Italy
[b] ISTI-CNR, Pisa, Italy
[c] Aalto University, Helsinki, Finland

ARTICLE INFO

Article history:
Received 23 August 2017
Revised 12 October 2017
Accepted 13 October 2017
Available online 1 November 2017

Keywords:
Controversy detection
Polarization
Social network analysis
Twitter
Motif
Social media

ABSTRACT

Online social networks are becoming the primary medium by which people get informed, as they provide a forum for expressing ideas, contributing to public debates, and participating in opinion-formation processes. Among the topics discussed in Social Media, some lead to controversy.

Identifying controversial topics is useful for exploring the space of public discourse and understanding the issues of current interest. Thus, a number of recent studies have focused on the problem of identifying controversy in social media mostly based on the analysis of textual content or rely on global network structure. Such approaches have strong limitations due to the difficulty of understanding natural language, especially in short texts, and of investigating the global network structure.

In this work, we show that it is possible to detect controversy in social media by exploiting network motifs, i.e., local patterns of user interaction. The proposed approach allows for a language-independent and fine-grained analysis of user discussions and their evolution over time. Network motifs can be easily extracted both from user interactions and from the underlying social network, and they are conceptually simple to define and very efficient to compute. We assess the predictive power of motifs on a manually labeled twitter dataset. In fact, a supervised model exploiting motif patterns can achieve 85% accuracy, with an improvement of 7% compared to baseline structural, propagation-based and temporal network features. Finally, thanks to the locality of motif patterns, we show that it is possible to monitor the evolution of controversy in a conversation over time thus discovering changes in user opinion.

1. Introduction

The usage of online social networks is becoming an increasing trend through which people around the globe are in contact with others and get informed about topics of interest. Additionally, online social networks provide a forum for expressing ideas, contributing to public debates, and participating in opinion-formation processes. Even though many studies have been devoted to understand different aspects of social network structure and function, such as, community structure [1], information spreading [2], information seeking [3], link prediction [4], etc., much less work is available on analyzing online discussions and public debates.

In this paper, we study the problem of identifying controversies in social media, one of the many different aspects of analyzing online discussions and understanding how people participate in those. The problem of studying controversy in social media has recently drawn some attention [5,6]. However, as this is a difficult problem, involving processing of human language and network dynamics, existing studies have limitations. For example, many papers study controversy in very controlled case studies, or focus on a predefined topic, most typically politics [7], for which they employ auxiliary domain-specific sources and datasets. In other cases, proposed approaches are based on content-based analysis [8], which has several limitations, as well, due to the ambiguity of the language and the fact that models become language-dependent and topic-dependent.

Instead, in this paper we aim to identify controversies on *any* topic, discussed in *any* language. Given this objective, our approach is based on the analysis of the *network structure*. In this sense, our paper is related to the recent work of Garimella et al. [5], who also aim at identifying controversies in the wild, independent of topic or language. In that work, the authors focus on a topic defined by a single hashtag, and then analyze the retweet network after partitioning it into two clusters (the two sides of controversy).

* Corresponding author.
E-mail addresses: mauro.coletto@imtlucca.it, mauro.coletto@unive.it, mauro.coletto@isti.cnr.it (M. Coletto), kiran.garimella@aalto.fi (K. Garimella), aristides.gionis@aalto.fi (A. Gionis), claudio.lucchese@unive.it, claudio.lucchese@isti.cnr.it (C. Lucchese).

https://doi.org/10.1016/j.osnem.2017.10.001
2468-6964/

An obvious limitation in their work is that they assume that a topic partitions the network into two clusters (while none, or more than two clusters, may be present), and that it is computationally feasible to identify those clusters. In our work, we overcome those limitations by analyzing local network patterns (*motifs*), and thus, making no assumption about the global cluster structure of the network, or about our ability to detect network clusters. Moreover, note that the separation of the retweet network in communities does not always reflect controversy; it may also mean that a hashtag is used in two communities with different acceptations. Our model catches antagonism in the conversation and, in fact, we find that some hashtags (#germanwings, #onedirection) that were detected as not controversial by previous studies, contain controversial discussions. Finally, in the work of Garimella et al. [5] the approach of detecting controversy is static and is based on analyzing the retweets of a given hashtag. In our case we focus on the analysis of the discussions generated by those tweets. This allows us to discover potentially controversial sub-topics that may be present within an otherwise non-controversial topic.

We propose the use of motifs extracted from the user reply and friendships graphs to detect controversial threads of discussion in online social networks. The proposed motifs can be easily computed as they encompass interactions among two or three users only. Being graph-based, such motifs are language independent and topic independent: they can be applied to investigate interactions in social networks without any additional domain knowledge. We measure the predictive power of the proposed motifs on a collection of Twitter data. We found that local motifs can improve the accuracy of frequently used graph-based features (e.g., cascade depth, inter-reply time) achieving an accuracy of 85%. We claim that such motifs are able to model both user homophily, through the friendship graph, and user interest in discussing specific topics even beyond their social circles, through the reply graph.

This paper is an extended version of a previous conference paper [9]. The original contributions presented in this paper include: a more detailed description of the proposed method, a dynamic use of the method, an additional experiment on a dataset based on Twitter hashtags (Dataset2: *Twitter hashtags.*). We applied the method to specific accounts (Dataset1: *Twitter pages.*), but also to specific concepts, represented by Twitter hashtags. We used a previously used Twitter hashtags dataset in order to compare our approach to previous ones and we report the analyses. Finally, the proposed motifs, being local to two or three users, allow a fine-grained analysis of the evolution of a discussion over time and of the interactions among its users. We extended the conference paper with the description of a temporal variant of the method, reporting some relevant examples. In fact, we found that non controversial conversations happen to become controversial either limitedly to a sub-tree of the discussion thread, or globally due for instance to external events such as news.

2. Related work

Controversy and polarization. The analysis of controversy on the web and social media has received considerable attention in recent years, with a number of papers studying controversy on general web pages [10], blogs [11], online news [8,12], and social media [5,7,13].

The existence of polarization on social media was first studied by Adamic and Glance [11] who identified a clear separation in the hyperlink structure of political blogs. Conover et al. [7] studied this phenomenon on Twitter, evaluating the polarization on the retweet network. In a more recent work, Garimella et al. [5] showed that the polarized structure in the retweet graph extends beyond politics. They also proposed algorithmic methods to measure the amount of controversy on a topic, by considering the structure of the network formed by retweets and followers. In a similar spirit, Guerra et al. [14] considered a measure based on boundary connectivity patterns in order to identify if a discussion is controversial. Other approaches have also been proposed to identify controversy on social media at a *user* level. For example, BiasWatch is a weakly-supervised approach fusing content and network data to infer user polarity [6].

Controversies are inherently dynamic. Non-controversial topics could become controversial and vice-versa. Morales et al. [15] present an approach based on label propagation in order to quantify the level of controversy in the network. They apply their measure on Twitter data from Venezuela over a long period and showed that they can capture real-life shifts in polarization. Coletto et al. [16] proposed an approach for jointly tracking user polarity and topic evolution. The method proposed in this paper can handle the dynamic nature of a controversial topic.

Conversation graphs (reply graphs) are used to represent the dynamic nature of information and discussion threads in a network. Various studies have proposed methods to analyze conversation graphs on Twitter [17,18]. Those studies analyze various types of conversation graphs, such as *long path-like reply trees, large star-like trees*, and *long irregular trees*. They also show that paths are making up to 60% of the reply graphs. In our work, we observe that reply graphs of Twitter discussions are composed by a majority of star-like trees. For controversial discussions, we additionally detect long trees with multiple branches indicating the different threads of the discussions, e.g., see Fig. 1.

Analysis of conversation graphs in rumor and misinformation spreading has shown that information flow in the network gives rise to certain types of local patterns [19,20]. Smith et al. [21] study the role of social media in the discussion of controversial topics. They try to understand reply and retweet interactions at a user level and conclude that We that users are quicker to spread information that agrees with their position more often.

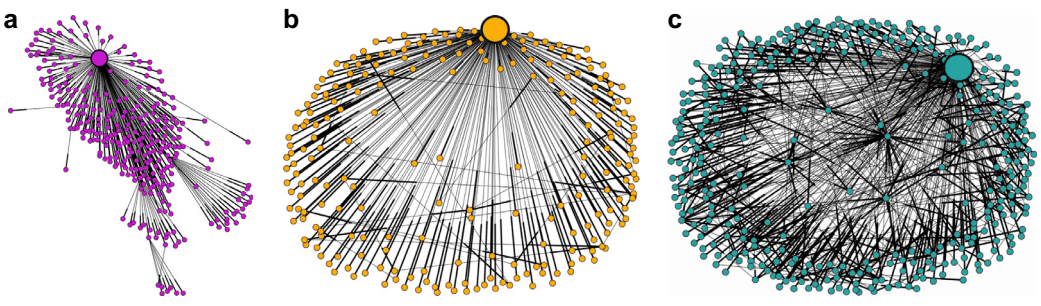

Fig. 1. Examples of different user-interaction networks: (a) content reply tree; (b) user reply graph for a non-controversial conversation; and (c) user reply graph for a controversial conversation.

However, to our knowledge, this is the first work to do an in-depth study of the role of network motifs in the context of identifying controversy in social media.

Motifs indicate patterns of interactions/interconnections in complex networks. The work of Milo et al. [22] was one of the first to analyze the occurrence of different motifs in networks arising in a wide range of fields, from biochemistry to engineering. Their finding that "*motifs may thus define universal classes of networks*" is one of our motivations for exploring simple interaction patterns related to controversy.

In the context of social networks, motifs may indicate a specific function or role of certain nodes. For example, network motifs have been used recently to explain higher-order network organization, and subsequently, use this information to cluster networks [23].

Conversation textual analysis. The problem of detecting disagreement in conversation text was recently studied by Allen et al. [24], who use rhetorical structure features to identify disagreement. They claim that this is a difficult task, even for humans.

Most related to our paper is the work by Chen and Berger [25], who study when, why, and how a conversation is initiated by a controversy. Their main hypothesis is that a controversy generally brings up interest and discomfort in users, and when the former is higher, a controversy causes a conversation, while otherwise, the likelihood of starting a conversation is smaller. Supporting evidence for this hypothesis is obtained by analyzing an online news website.

Furthermore, language-analysis tools have been used widely to determine the emotional tone of a conversation [26], e.g., whether a message is partial or impartial [27], subjective/objective, positive/negative [28], etc.

All the different methods discussed above use only textual information. Even though the use of text features is orthogonal to our method, and they can be added separately, we chose not to do so explicitly, since text-analysis tools are language dependent, and since we are mainly interested in contrasting network motifs with other network-structure features.

3. Data collection

We consider two Twitter datasets.

Dataset1: ***Twitter pages.*** Our main source of data is a carefully-curated set of popular Twitter pages which covers a wide range of domains (news, politics, celebrity, gossip, entertainment) and languages. The way we choose popular pages is generic and can be emulated on other social networks. For each page, we gather the last two hundreds tweets and we manually evaluate them to check if they are controversial or not through multiple annotators. To classify them the content of the tweet and the received user replies were considered. A tweet is labeled controversial if the content is debatable and it expresses an idea or an opinion which generates an argument in the replies, representing opposing opinions in favor or in disagreement with the root tweet. We consider only the pages whose tweets are almost completely controversial or not controversial, and we discard all the tweets from accounts with less than 90% controversial/non-controversial tweets. The final list of the 11 controversial and 7 non-controversial selected pages are shown in Table 1. It is interesting to note that in the controversial class most of the pages are related to politics and breaking news, showing an high controversial nature of the topic, while not-controversial pages are mainly related to celebrities, and entertainment. However in our experiments since we do not use the content of the interactions, the topic of the conversation is not taken into consideration. In the subsequent analysis, we use the page as a label for the collected tweets in that page, i.e., a tweet is deemed controversial (non-controversial) if it originates from a controversial (non-controversial) classified page.

Table 1
List of Twitter pages used in our study (Dataset1).

Controversial	Non controversial
@tedcruz, @mov5stelle, @brexitwatch, @barackobama, @realdonaldtrump, @wikileaks, @berniesanders, @cnnbrk, @bbcworld, @hillaryclinton, @potus	@coldplay, @justinbieber, @cristiano, @adele, @chanel, @xbox, @nba,

Table 2
Datasets statistics.

Dataset1: *Twitter pages*			
Filtering	Root posts	Avg. users	Tot. tweets
> 2 users	1202	108	192.7*K*
> 3 users	1175 (97%)	110	192.5*K*
> 10 users	1046 (87%)	123	191.3*K*
Dataset2: *Twitter hashtags*			
Filtering	Root posts	Avg. users	Tot. tweets
> 2 users	1302	32	61.4*K*
> 3 users	1211 (93%)	34	60.5*K*
> 10 users	699 (54%)	54	54.4*K*

For each collected tweet in each page (*root post*), we reconstructed the generated discussion thread by recursively crawling the tweet's replies. The task requires a complex crawling procedure to obtain the full tree. Moreover, since we are interested in analyzing the discussion generated by each post, we restrict to the tweets that generate a conversation involving more than k users, with k=2,3 and 10. (including the author of the original post). The reply tweets are often in a different language than the language of the original tweet, including Arabic, Russian, and others. Table 2 reports the number of root posts and total reply tweets that we collect with the above procedure, with $k = 2, 3, 10$. The final dataset contains more than $190\,K$ tweets in total. Moreover, the table reports the average number of users who take part in the conversation for each root post. Each collected root post generates a network of replies that involves on average about 100 users.

Dataset2: ***Twitter hashtags.*** In order to be consistent with the recent literature, we also collect tweets based on controversial and non-controversial hashtags, in particular, the ones used by Garimella et al. [5]. We use four controversial (#beefban, #baltimore, #netanyahuspeech and #russia_march) and four non-controversial hashtags (#germanwings, #onedirection, #sxsw, #ultralive). For each hashtag we collect the recent posts. For each post we collect all the reply tweets and build the dataset in the same way that was described before. Statistics on this dataset are reported in Table 2. Dataset2 contains more than $60\,K$ tweets in total.

We note that, upon manual inspection, for many hashtags in the above-mentioned dataset, there is a mix of different behaviors depending on the context in which the hashtag is used in the tweets. Some are predominantly controversial or non-controversial, while others are mixed. Dataset2 is used as an additional test set for our model trained on Dataset1 to assess the controversial nature of popular hashtags.

4. Controversy analysis and detection

Given a social network we are interested in modeling the interactions among users and the dynamics incurring due to generated content. Users in social networks establish *friendship* or *subscription* relationships with each other, and when users interact with or publish new content their *friends* are informed. We model these

relationships with a *user graph* $\mathcal{G} = (U, E)$, where U is the set of users of the network and an edge $e = (u_i, u_j) \in E$ indicates that users u_i and u_j are friends (undirected case) or that user u_i follows user u_j (directed).

Moreover, a user may publish some new content item c_i, possibly *in response to* another content item c_j authored by another user, thus generating complex threads of discussion. Interactions within a single thread are modeled with a content *reply tree* $\mathcal{T} = (C, R)$, where C is the set of content items in the thread, and an arc $r = (c_i, c_j) \in R$ indicates that c_i is a reply to c_j. Note that $\mathcal{T}$ is indeed a tree as each content item, except the first one (the root), is a response to exactly one other item (its parent). Additionally, the nodes of $\mathcal{T}$ are enriched with information about publishing time and authoring user.

The tree $\mathcal{T}$ can be projected onto the users to model reply interactions among users. The resulting structure is a user *reply graph* $\mathcal{R} = (U, I)$, where an edge $e = (u_i, u_j) \in I$ indicates that the user u_i has replied to some content item posted by user u_j. We refer to the user who authored the first content item as *origin*.

Fig. 1a shows a content *reply tree* (also referred to as just *reply tree*) present in our data, while Fig. 1b and c shows the user *reply graph* (or just *reply graph*) of two other discussion threads. Note that a social network may have several disconnected reply trees and reply graphs. Fig. 1b and c, even though are just examples, show how the network in the case of low controversy and high controversy might be really different from a structural point of view. The density of the graph in Fig. 1c for instance is higher than in Fig. 1b.

Our main hypothesis is that the structure of the user graph $\mathcal{G}$, the reply tree $\mathcal{T}$, and the reply graph $\mathcal{R}$ can be characterized by simple *motifs* of local user interactions that can be effectively exploited to distinguish between *controversial* and *non-controversial* content.

In addition to local motifs, we also explore whether baseline features (including network structure, content propagation, and temporal features) are predictors of controversy. This standard graph-based analysis is discussed in the next section while the motif-based analysis is presented in the section "Motifs."

4.1. Baseline graph-based analysis

Structural features. The simplest structural features to extract from the user-interaction networks are the *size* in terms of *number of nodes* and *number of edges*, and the *degree distribution*.

Fig. 2a shows the distribution of the sizes of the reply tree $\mathcal{T}$ and the reply graph $\mathcal{R}$ in terms of number of nodes and number of edges for Dataset1 about Twitter pages with all the reply networks with at least 3 users involved in the conversation. To some extent, these measures are related to the popularity of the content taken into consideration. Note that in our data the sizes of $\mathcal{T}$ and $\mathcal{R}$ are very similar for both controversial and non-controversial content. This finding is in line with Smith et al. [21] that controversial content does not necessarily generate larger threads of conversation. From this, we can conclude that for distinguishing controversy *among popular topics*, just the graph sizes do not suffice.

Fig. 2b reports the average degree for the reply tree $\mathcal{T}$ and the reply graph $\mathcal{R}$. In this case, the distributions are quite different for controversial and non-controversial content. A larger average degree is observed for controversial content, suggesting that such conversations generate more engagement among users.

Propagation-based features. In order to understand how information propagates among controversial and non-controversial conversations, we investigate a number of different properties of the reply trees $\mathcal{T}$ related to information propagation. Fig. 2c shows the distribution of average and maximum cascade depths, where a cascade is defined as a path from the root to a leaf of a reply tree. The figure also shows the distribution of the maximum-size subtree among all subtrees rooted in a child of the root node. We observe that for controversial content the reply trees generally have larger depth.

Fig. 2d reports the distribution of the degree for the root, as well as the node with the larger degree excluding the root in $\mathcal{T}$. We see that in this case the controversial and non-controversial discussions have similar distributions. Nevertheless, reply trees of controversial discussions have higher probability of having a smaller root degree than non-controversial, suggesting that controversial discussions go beyond the first level of interaction.

Given the above analysis, to summarize content propagation, we decided to use the two most significant features in the content reply trees. The other features, e.g. max cascade depth, are discarded because they are strongly related to popularity. In particular:

- *average cascade depth*: the average length of root-to-leaf paths;
- *maximum relative degree*: the largest node degree excluding the root node, divided by the degree of the root.

Temporal features. Considering the simple assumption that controversial topics may generate "dense" discussions in time, we analyze the time elapsed between a content item and its reply. Fig. 2e shows the distributions of minimum, maximum and average inter-reply time. Additionally, we measure the ratio of nodes in a reply tree occurring within one hour from the root. For all the measures above, there is no significant difference between controversial and non-controversial reply trees. For prediction purposes, we chose to use as features only the average inter-reply time and the ratio of replies in the first hour. Maximum and minimum inter-reply time are influenced by a single reply and for this reason they were not considered further.

4.2. Motifs

Our main hypothesis in this paper is that *local patterns* of user interaction can be used to discriminate between controversial and non-controversial discussions. This hypothesis is consistent with previous studies, where it was shown that local patterns can be used to characterize different types of networks [22,29]. As with previous work, we consider local patterns to be 2- and 3-node connected subgraphs. We refer to such patterns as *motifs*.

We consider motifs in the user graph $\mathcal{G}$ and the reply graph $\mathcal{R}$. These two graphs encompass two different kinds of information. An edge in the user graph $\mathcal{G}$ indicates that a user follows another user. These two users are likely to have similar interests and/or opinions. On the other hand, the reply graph $\mathcal{R}$ models the activity among users who may not know each other but they are willing to discuss or comment on a specific topic. In this sense, the reply graph $\mathcal{R}$ is much more dynamic and content-dependent. Antagonism between users, which can not be captured by the user graph $\mathcal{G}$ can be captured by the reply graph $\mathcal{R}$. Our basic assumption is that a combined analysis of the two graphs, $\mathcal{G}$ and $\mathcal{R}$, can lead to an improved model for controversy detection.

Dyadic motifs. We consider all possible patterns between two users in graphs $\mathcal{G}$ and $\mathcal{R}$, such that that there is at least one reply (i.e., one edge in graph $\mathcal{R}$) — otherwise the two users do not interact with each other in the discussion thread. There are seven possible configurations, which are shown in Fig. 3a. Fig. 3b shows the frequency distribution of dyadic motifs in our data. Note that patterns are mutually exclusive, therefore, pattern *A* where u_i replies to u_j also implies than u_j does not reply u_i and that the two users do not follow each other.

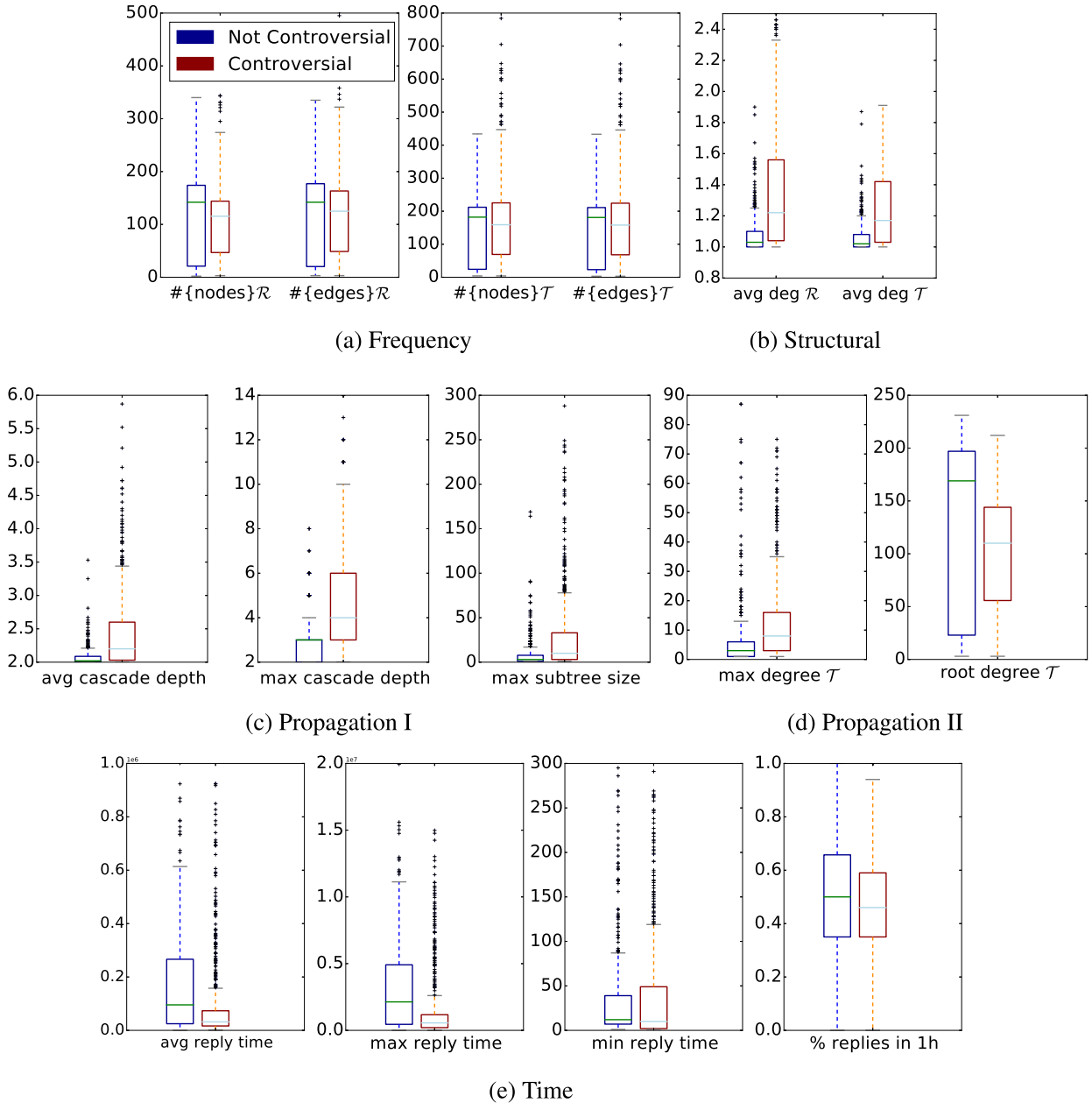

Fig. 2. (a) Distribution of the number of nodes and edges in $\mathcal{T}$ and $\mathcal{R}$. (b) Distribution of average node degree in $\mathcal{T}$ and $\mathcal{R}$. (c) Distribution of avg./max. cascade depth and max. subtree size. (d) Distribution of origin degree and max. degree in $\mathcal{T}$ and $\mathcal{R}$. (e) Distribution of average, max., min. inter-reply time, and percentage of replies within one hour from the root. Non-controversial in blue (left side) vs. controversial in red (right side). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The most frequent dyadic motifs are *A* and *C*. According to Fig. 3b, it is more likely to observe a reply to a followed user in non-controversial cases. Conversely, in controversial cases it is likely to reply to a user not being followed. This confirms the intuition that controversial discussions thread interactions also among users not directly connected in the user graph $\mathcal{G}$. The features used for detecting controversial content are the frequencies of all dyadic motifs.

Triadic motifs. We also consider 3-node motifs, in particular closed triangles. As in the case of dyadic motifs, we combine structural information from the user graph $\mathcal{R}$ and the reply graph $\mathcal{G}$. Fig. 4a shows some motifs we considered. We detect a triadic motif only if there is a reply interaction among the three users. Due to the high number of possible motifs and since most motifs are relatively rare in the data, we coalesce motifs in groups. Overall, we form our set of triadic motifs by considering (*i*) the number of follow edges among the three users (Fig. 4a), (*ii*) the number of reciprocal follow edges, and (*iii*) the number of non reciprocal follow edges with opposite direction with respect to the reply edge. In total we have 20 different triadic motifs. The frequency of each motif is considered as a feature for predicting controversy.

For the lack of space we do not report the distribution for all the motifs, but generally most of the patterns we considered for closed triangles were quite rare in the dataset. Only a few of them are frequent and mostly in controversial threads, confirming the intuition that controversial discussions exhibit a more complex structure. The reason for the scarcity of complex structures is that in microblogging platforms the interactions are brief and generally involve few users. Because of the infrequency of the appearance of patterns that include more than three nodes we limited the

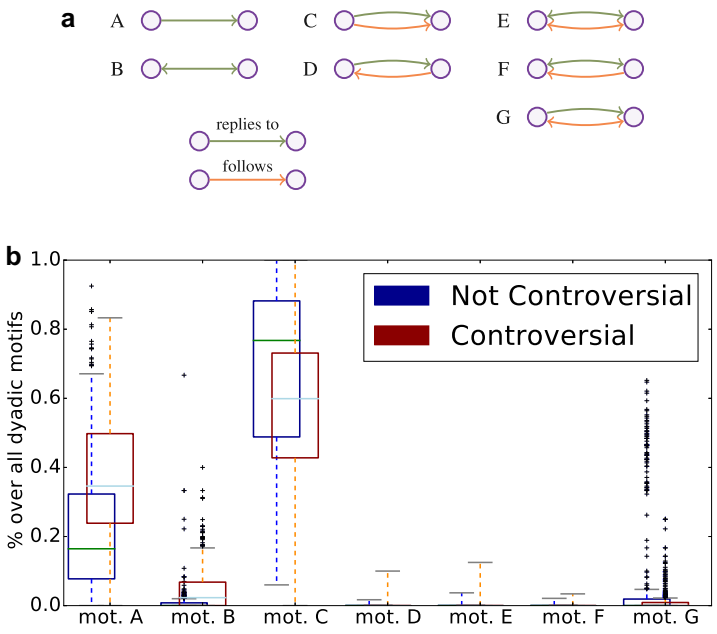

Fig. 3. (a) Dyadic motifs and (b) their frequency distribution.

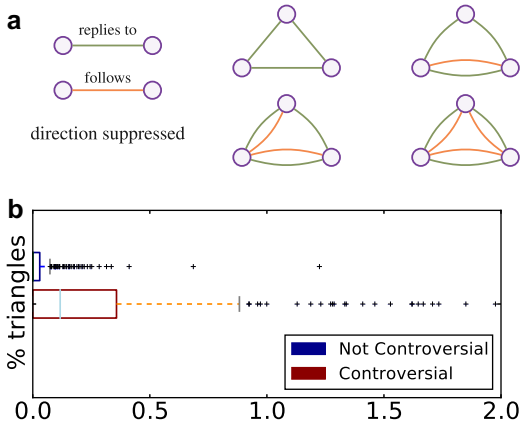

Fig. 4. (a) Triadic motifs and (b) distribution of undirected reply triangles ratio.

study to dyadic and triadic structures (the triads already showed a marginal value in our identification task). Due to this choice the network motifs used can be easily extracted. Network motifs can be easily extracted both from user interactions and from the underlying social network, and they are conceptually simple to define and very efficient to compute.

To provide additional insights on user interactions, we consider as additional feature the ratio of triangles in the reply graph $\mathcal{R}$ over the number of all possible triangles $\binom{|U|}{3}$. Again, a larger triangle ratio indicates that controversial content generates more complex discussion threads with more interactions among users and not only dyadic relations between the author of the post and the replying user, as it in the case of non-controversial situations.

We also considered "open" triadic motifs, i.e., 3-user subgraphs connected by only two replies. Such patterns did not seem to help much in predicting controversial discussions and therefore they are not considered further. The features considered in this work are shown in Table 3.

Table 3
Summary of all features.

Baseline: Structural	Avg. degree in $\mathcal{T}$ Avg. degree in $\mathcal{R}$
Baseline: Propagation	Avg. cascade depth in $\mathcal{T}$ Max. relative degree
Baseline: Temporal	Avg. inter-reply time % replies in 1h
Dyadic motifs	7 2-node motifs (shown in Fig. 3a)
Triadic motifs	20 3-node motifs Triangles ratio

5. Experiments

5.1. Detection of controversy in Twitter pages

We used the Twitter datasets presented in the data collection section. As already discussed, the Twitter pages of Dataset1 can be entirely labeled controversial or non-controversial, therefore we classify tweets according to the page it belongs. The dataset is quite balanced, with about 60% instances belonging to the controversial class and 40% to the non-controversial. Reported experiments are performed using 5-fold cross-validation and averaged over 100 trials.

We evaluated different classifiers, including AdaBoost, Logistic Regression, SVM and Random Forest, and chose AdaBoost as it resulted in the best performance. We want to detect the controversial nature of a post by analyzing user graph and reply trees. To show the relevance of detecting motifs to quantify controversy we compare the results with baseline graph-based features. We analyzed the performance by the baseline graph-based features and by using motif-based features (in addition and alone). We report the accuracy of the classifier on both controversial and non-controversial classes, and the precision, recall and F-measure with respect to the controversial class.

As shown in Table 4 the baseline approach accuracy (with structural, propagation-based and temporal features) is above 75% and increases only slightly when restricting to reply trees with

Table 4
Performance of the motif based classifier.

Filtering	Accuracy	Precision	Recall	F-measure
Baseline				
> 2 users	0.76	0.79	0.81	0.80
> 3 users	0.77	0.80	0.82	0.81
> 10 users	0.78	0.81	0.83	0.82
Baseline + dyadic motifs				
> 2 users	0.82	0.84	0.86	0.85
> 3 users	0.83	0.85	0.86	0.85
> 10 users	**0.84**	**0.86**	**0.88**	**0.87**
Baseline + dyadic and triadic motifs				
> 2 users	0.83	0.85	0.86	0.85
> 3 users	0.84	0.86	0.85	0.86
> 10 users	**0.85**	**0.87**	**0.88**	**0.87**
Dyadic motifs only				
> 2 users	0.75	0.77	0.82	0.80
> 3 users	0.75	0.77	0.82	0.80
> 10 users	0.77	0.79	0.84	0.82
Triadic motifs only				
> 2 users	0.73	0.89	0.62	0.73
> 3 users	0.74	0.88	0.64	0.74
> 10 users	0.77	0.89	0.71	0.79
Dyadic + Triadic motifs only				
> 2 users	0.77	0.81	0.80	0.81
> 3 users	0.77	0.81	0.80	0.80
> 10 users	0.80	0.84	0.82	0.83

Table 5
Feature importance (filtering > 10 users).

Feature	Error reduction
(1) Avg. inter-reply time	0.18
(2) Max. relative degree	0.16
(3) Motif *A*	0.14
(4) % Replies within 1 h	0.08
(5) Motif *B*	0.08
(6) Motif *G*	0.06
(7) Triangles ratio	0.04
(8) Triadic motif	0.04

more than 10 users. With the addition of dyadic motifs, all the performance figures are significantly improved. Note that the precision of the algorithm improves in both controversial and non-controversial classes. The addition of triadic motifs leads to the best results, but the improvement is only marginal. This is because, as discussed in the previous section, triads are infrequent: even if conveying relevant information, they may help in improving the classification of a limited number of instances. The best results highlighted in boldface in Table 4 are statistically significant (t-test with p-values $\ll 0.01$) w.r.t. baseline features. Using dyadic motifs alone, moreover, the accuracy of the model is comparable with the baseline, with a limited improvement if we add the triadic patterns.

In Table 5 we report the 8 most relevant features exploited by the AdaBoost model according to their contribution in the error reduction. Temporal features are important to detect controversy. The first feature is the average inter-reply time, and the fourth is the ratio of replies posted within one hour of the original tweet: when the discussion is polarized people tend to reply in a shorter time. This result is in line with other contexts. For example, it is known that temporal features play the main role to predict popularity [30]. The second most important feature is the maximum relative degree, i.e., the maximum degree normalized by the root node degree. In non-controversial reply trees, the root is the only node with a large degree, i.e., the node attracting most of the reply activity.

The other features among the top-6 are dyadic motifs. The most relevant being motif *A*, which corresponds to a user u_i replying to u_j without any following relationship among the two. We deduce that controversial threads create engagement among users not being directly connected in the social network. On the other hand, the fact that motif *C* is not relevant (where a user replies to a follower), suggests that it is less likely to have controversial discussions among friends. Interestingly, dyadic patterns seem to be more relevant than propagation-based features. For instance, the depth of the cascades, which was expected to model the complexity of the interactions, is not among the top-8 features. Presumably, complex propagation features are superseded by the simple motif patterns.

Finally, the last two important features are based on triangles. In particular the relevance of the triangle-ratio feature suggests that triadic patterns are able to grasp interactions occurring in controversial discussions. However it is harder to draw any conclusion on the role of specific triads patterns, due to their low frequency. The most significant specific triadic pattern included in the list in Table 5 is a close reply triangle with two follow edges: one reciprocal and one not reciprocal with the same direction of the underlying reply edge. Since triadic patterns provide a limited contribution to the classifier, we conclude that dyadic motifs are already effective, and there is not much information that can be extracted based on specific triadic motifs.

5.2. Dynamic tracking of controversy

We found it is not always appropriate to classify a reply tree as controversial or not. This is because each reply may generate unexpected reaction. For instance, there may be sub-threads of controversy, within a non-controversial discussion. To test this intuition, we analyzed the direct replies of the *origin* tweets that were classified as non-controversial. This can be achieved easily as the proposed approach can be applied to any tweet given its reply tree, or in this case, its reply sub-tree. By applying the model discussed in the previous section, we found that about 7% of the direct-reply sub-trees of a non-controversial tweet are controversial.

One such example is shown in Fig. 5, illustrating the reply tree of a post by Justin Bieber. A majority of the replies are not controversial and are written by his fans with compliments and expressions of affection and love. However, the proposed algorithm detected as controversial one sub-tree (highlighted in red) generated by a reply in support to another singer: “Zayn is better.” This post generated a subtree with animated discussion among fans. A similar case was found for Cristiano Ronaldo’s profile, where a number of users started discussion about his rivalry with Messi.

Both of the previous examples are typical cases in which the controversial portion of the discussion is limited to a few branches, and its detection might be challenging. We claim that the proposed approach, based on local motifs can successfully detect small controversial sub-threads.

5.3. Hashtags evaluation

Since on Twitter, topics are often identified through hashtags, we tested the proposed method on tweets mentioning a given hashtag (Dataset2), obtained from the previous work [5]. Table 6 shows the fraction of controversial posts per hashtag, as detected by our model. The smallest fraction of controversial discussions is found with #sxsw and #ultralive hashtags (related to music events), where most conversations are expected to happen among supporters of the same music band. The most controversial discussions are found with the #beefban, #onedirection, #netanyahu,

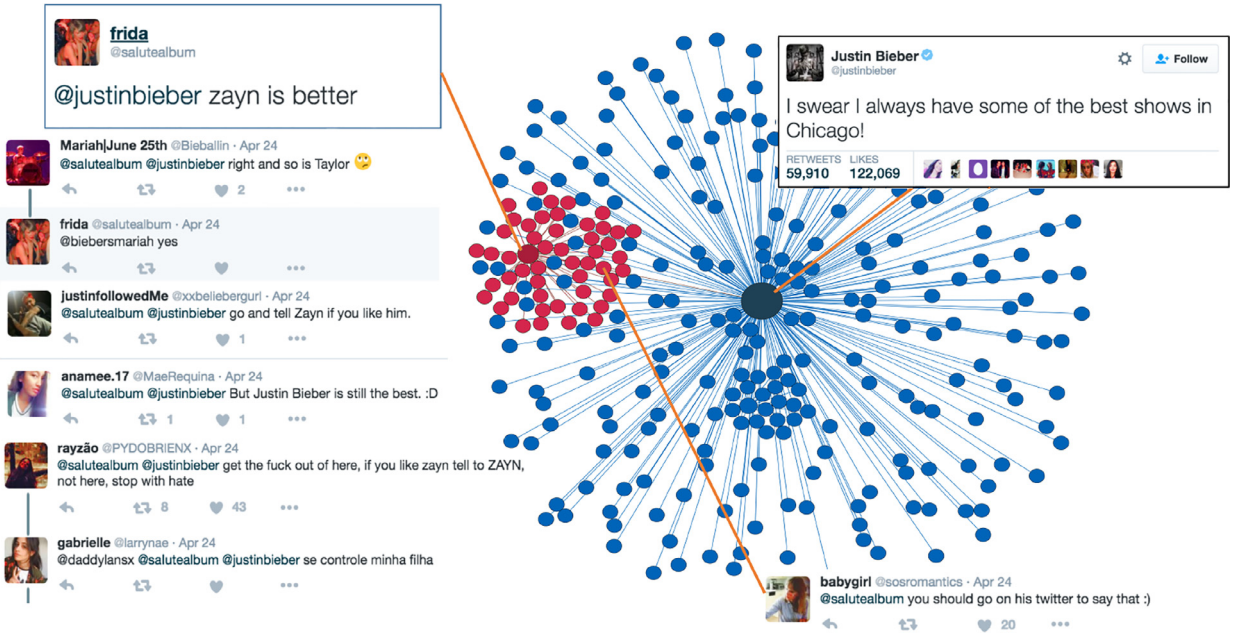

Fig. 5. A controversial reply sub-tree (in red) originated by a non-controversial post (in blue) by Justin Bieber. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

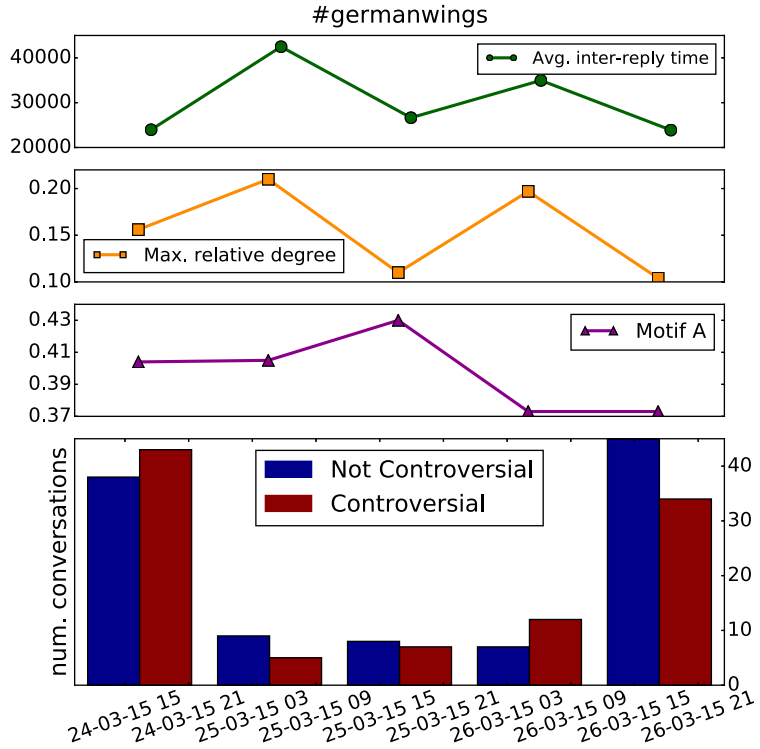

Fig. 6. Distribution of controversial (red) vs. non-controversial (blue) posts and top-3 features values over time for the #germanwings hashtag. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

#baltimore hashtags. The classification of these hashtags as controversial is in line with the previous results [5], with the exception of #onedirection for which we detected antagonist replies, upon manual inspection. Most of the hashtags exhibit a mixed behavior as far as controversy is concerned.[1] Indeed, simply counting the number of tweets classified as controversial is a quite naïve approach, strongly dependent on different factors, such as the daily volume of tweets, on external events, and many others. For these reasons, we believe that it is more interesting to study how the controversy related to a given hashtag evolves over time.

Fig. 6 shows the evolution of the controversy for the #germanwings hashtag. Note that some hours after the accident happened on March 24 the majority of threads are controversial. In the evening the discussions become less controversial and mainly about sorrow and condolences. An interesting increase of the

[1] E.g.: A controversial tweet id=580330769912061953 and a non-controversial tweet id=580361863449444352 about #germanwings.

Table 6
Hashtag controversy classification.

Hashtag	Ratio of controversial posts
sxsw	0.32
Germanwings	0.49
Beefban	0.70
Netanyahu	0.55
Ultralive	0.29
Onedirection	0.61
Baltimore	0.58
Russia-march	0.46

controversy level is registered the next day, until details about the accident were released. Then the discussion becomes predominantly non-controversial showing that the audience has digested the news. We highlight that the level of controversy is anti-correlated with the frequency for motif *A*, thus confirming the prediction power of the proposed motifs. Moreover, we showed in the figure also the trend for other significant features used by the classification model. The average inter-reply time and the maximum relative degree are trends are very similar but the correlation with the classification results is not easily evident, thus explaining the importance of combining many different features to get a useful classification.

6. Conclusion

We proposed a novel approach based on local graph motifs for identifying controversy on online social networks. The proposed method is language independent and exploits local patterns of user interactions to detect controversial threads of discussion. Given a content item, users reply to each other generating different configurations of the reply graph. We investigated local motifs extracted from this graph and from the user friendship graph. Such motifs correspond to different interaction patterns among two users, which may be linked by a possibly reciprocal reply action and by a possibly reciprocal friendship relationship. Similar motifs regarding the interaction of three users were considered.

We proved on a benchmark Twitter dataset that such motifs are more powerful in predicting controversy than other baseline frequently used graph properties such as cascade depth. Specifically dyadic patterns seem to be more relevant than structural features to detect controversy. We observed that in most cases controversy arise when users participate to discussions beyond their social circles. This means that it is less likely to have controversial discussions among friends. Finally, as the proposed motifs can be easily extracted from any reply tree or sub-tree, we experimented with the use of such patterns in monitoring the evolution of discussions and sub-discussions over time. Indeed, we found that a topic of discussion develops over time changing its level of controversy depending on different sub-topics or on external events (e.g., news). About 7% of the direct-reply sub-trees of a non-controversial tweet are detected as controversial.

Therefore, a fine-grained analysis, as provided by the proposed local motifs, is necessary for a better understanding of controversy in online social networks.

Acknowledgments

This work was partially supported by the EC H2020 Program INFRAIA-1-2014-2015 *SoBigData: Social Mining & Big Data Ecosystem* (654024).

References

[1] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[2] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the WWW, ACM, 2012, pp. 519–528.
[3] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media? in: Proceedings of the WWW, ACM, 2010, pp. 591–600.
[4] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.
[5] K. Garimella, G. De Francisci Morales, A. Gionis, M. Mathioudakis, Quantifying controversy in social media, in: Proceedings of the WSDM, ACM, 2016, pp. 33–42.
[6] H. Lu, J. Caverlee, W. Niu, Biaswatch: a lightweight system for discovering and tracking topic-sensitive opinion bias in social media, in: Proceedings of the CIKM, ACM, 2015, pp. 213–222.
[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political Polarization on Twitter, in: Proceedings of the ICWSM, 2011.
[8] Y. Mejova, A.X. Zhang, N. Diakopoulos, C. Castillo, Controversy and sentiment in online news, Symp. Comput. Journal. (2014).
[9] M. Coletto, K. Garimella, A. Gionis, C. Lucchese, A motif-based approach for identifying controversy, in: Proceedings of the ICWSM, 2017.
[10] S. Dori-Hacohen, J. Allan, Detecting controversy on the web, in: Proceedings of the CIKM, ACM, 2013, pp. 1845–1848.
[11] L.A. Adamic, N. Glance, The political blogosphere and the 2004 us election: divided they blog, in: Proceedings of the LinkKDD, 2005, pp. 36–43.
[12] Y. Choi, Y. Jung, S.-H. Myaeng, Identifying controversial issues and their sub–topics in news articles, in: Intelligence and Security Informatics, Springer, 2010, pp. 140–153.
[13] J. An, D. Quercia, J. Crowcroft, Partisan sharing: Facebook evidence and societal consequences, in: COSN, 2014, pp. 13–24.
[14] P.H.C. Guerra, W. Meira Jr, C. Cardie, R. Kleinberg, A measure of polarization on social media networks based on community boundaries, in: Proceedings of the ICWSM, 2013.
[15] A. Morales, J. Borondo, J. Losada, R. Benito, Measuring political polarization: Twitter shows the two sides of Venezuela, Chaos 25 (3) (2015).
[16] M. Coletto, C. Lucchese, S. Orlando, R. Perego, Polarized user and topic tracking in twitter, in: Proceedings of the SIGIR, Pisa, Italy, 2016.
[17] P. Cogan, M. Andrews, M. Bradonjic, W.S. Kennedy, A. Sala, G. Tucci, Reconstruction and analysis of Twitter conversation graphs, in: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research, ACM, 2012, pp. 25–31.
[18] R. Nishi, T. Takaguchi, K. Oka, T. Maehara, M. Toyoda, K.-i. Kawarabayashi, N. Masuda, Reply trees in twitter: data analysis and branching process models, Soc. Netw. Anal. Min. 6 (1) (2016) 1–13.
[19] M. De Domenico, A. Lima, P. Mougel, M. Musolesi, The anatomy of a scientific rumor, Sci. Rep. 3 (2013) 2980.
[20] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the WWW, ACM, 2011, pp. 675–684.
[21] L.M. Smith, L. Zhu, K. Lerman, Z. Kozareva, The role of social media in the discussion of controversial topics, in: Proceedings of the SocialCom, IEEE, 2013, pp. 236–243.
[22] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.
[23] A.R. Benson, D.F. Gleich, J. Leskovec, Higher-order organization of complex networks, Science 353 (6295) (2016) 163–166.
[24] K. Allen, G. Carenini, R.T. Ng, Detecting disagreement in conversations using pseudo-monologic rhetorical structure., in: Proceedings of the EMNLP, 2014, pp. 1169–1180.
[25] Z. Chen, J. Berger, When, why, and how controversy causes conversation, J. Consum. Res. 40 (3) (2013) 580–593.
[26] A. Kanavos, I. Perikos, P. Vikatos, I. Hatzilygeroudis, C. Makris, A. Tsakalidis, Conversation emotional modeling in social networks, in: Proceedings of the IEEE 26th International Conference on Tools with Artificial Intelligence, IEEE, 2014, pp. 478–484.
[27] M.B. Zafar, K.P. Gummadi, C. Danescu-Niculescu-Mizil, Message impartiality in social media discussions, in: Proceedings of the ICWSM, 2016.
[28] L. Barbosa, J. Feng, Robust sentiment detection on twitter from biased and noisy data, in: ICCL: Posters, Association for Computational Linguistics, 2010, pp. 36–44.
[29] I. Bordino, D. Donato, A. Gionis, S. Leonardi, Mining large networks with subgraph counting, in: Proceedings of the Eighth IEEE ICDM, IEEE, 2008, pp. 737–742.
[30] B. Shulman, A. Sharma, D. Cosley, Predictability of popularity: Gaps between prediction and understanding, in: Proceedings of the ICWSM, 2016.

Mauro Coletto is a research fellow at DAIS, Ca' Foscari University, and he received his Ph.D. degree from IMT - Institute for Advanced Studies (Lucca) in the track "Computer, Decision, and Systems Science". He graduated in Information Management Engineering at the University of Udine in 2012. During his doctoral studies at IMT, he has worked in collaboration with CNR (ISTI-HPC) on the following research topics: Web Mining, Online Social Networks, and Social Media Analysis.

Kiran Garimella is a Ph.D. student at Aalto University. His research focuses on identifying and combating filter bubbles on social media. Previously he worked as a Research Engineer at Yahoo Research, QCRI and as an Research Intern at LinkedIn and Amazon. His research on polarization on social media received the best student paper awards at WSDM 2017 and Webscience 2017.

Aristides Gionis is Professor in the Department of Computer Science of Aalto University, leading the Data Mining Group. His research focuses on data mining and algorithmic data analysis. He is particular interested in algorithms for graphs, social-network analysis, and algorithms for web-scale data. He was a senior research scientist in Yahoo! Research, and previously an Academy of Finland postdoctoral scientist in the University of Helsinki. He obtained his Ph.D. from Stanford University in 2003. He had the pleasure to work as a summer intern in Microsoft Research, AT&T Labs, and Bell Labs Lucent Technologies.

Claudio Lucchese is Associate Professor with the Dipartimento di Scienze Ambientali, Informatica e Statistica at the Universit Ca' Foscari di Venezia. Between 2007 and 2017 he was researcher with the I.S.T.I. "A. Faedo" C.N.R.. Prior to joining C.N.R., he received his MS.c. and Ph.D. in Computer Science from the Universit Ca' Foscari di Venezia in 2003 and 2008, respectively. His main research activities are in the areas of data mining techniques for information retrieval and large-scale data processing.