A Motif-based Approach for Identifying Controversy

Mauro Coletto
Ca' Foscari Uni.- Venice
mauro.coletto@unive.it

Kiran Garimella
Aalto University - Helsinki
kiran.garimella@aalto.fi

Aristides Gionis
Aalto University - Helsinki
aristides.gionis@aalto.fi

Claudio Lucchese
CNR Pisa
claudio.lucchese@isti.cnr.it

Abstract

Among the topics discussed in Social Media, some lead to controversy. 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, 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 and efficient-to-compute analysis of user discussions and their evolution over time. The supervised model exploiting motif patterns can achieve 85% accuracy, with an improvement of 7% compared to baseline structural, propagation-based and temporal network features.

Introduction

In this paper we study the problem of identifying controversies in social media, which has recently drawn some attention (Garimella et al. 2016; Coletto et al. 2016). 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 (Conover et al. 2011), for which they employ auxiliary domain-specific sources and datasets. In other cases, proposed approaches are based on content-based analysis (Mejova et al. 2014), which has several limitations, as well, due to the ambiguity of the language and the fact that models become language-dependent and topic-dependent. We aim to identify controversies on *any* topic, discussed in *any* language. In this sense, our paper is related to the recent work of Garimella et al. (Garimella et al. 2016), who also aim at identifying controversies based on the analysis of the *network structure*. An obvious limitation in their work is that they assume that a topic partitions the network always

into two clusters 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 approach catches antagonism in the conversation and it allows to dynamically discover potential controversial sub-discussions that may be present within an otherwise non-controversial topic.

Data collection

Dataset: *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. 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 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 resulting in 11 controversial and 7 non-controversial pages: a tweet is deemed controversial (non-controversial) if it originates from a controversial (non-controversial) classified page. For each collected tweet in each page (*root post*), we reconstructed the generated discussion thread by recursively crawling the tweet's replies. 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). Table 1 reports the number of root posts and total reply tweets that we collect. The final dataset contains more than $190K$ tweets in total. Each collected root post generates a network of replies that involves on average about 100 users.

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. We consider 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 user u_i follows user u_j. 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. 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. Our hypothesis is that the structure of $\mathcal{G}$, $\mathcal{T}$, and $\mathcal{R}$ can be characterized by simple *motifs* of local user interactions useful 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.

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*.

Figure 1a 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 in the dataset (at least 3 users involved in the conversation). 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. (Smith et al. 2013) that controversial content does not necessarily generate larger threads of conversation.

Figure 1b reports the average degree for the reply tree $\mathcal{T}$ and the reply graph $\mathcal{R}$. In this case, the distributions are quite different: 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, we investigate a number of different properties of the reply trees $\mathcal{T}$ related to information propagation.

Figure 1c 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.

Figure 1d reports the distribution of the degree for the root, as well as the node with the larger degree excluding the root in $\mathcal{T}$. 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. We decided to use the two most significant features in the content reply trees: (*average cascade depth*) *the average length of root-to-leaf paths and (*maximum relative degree*) the largest node degree excluding the root node, divided by the degree of the root. The other features, e.g. max cascade depth, are discarded because they are strongly related to popularity.

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 (Figure 1e). Additionally, we measure the ratio of nodes in a reply tree occurring within one hour from the root. For prediction purposes, we chose to use as features only the average inter-reply time and the ratio of replies in the first hour, since maximum and minimum inter-reply time are influenced by a single reply.

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 (Milo et al. 2002). We consider motifs in the user graph $\mathcal{G}$ and the reply graph $\mathcal{R}$. 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.

We consider all possible patterns between two users in graphs $\mathcal{G}$ and $\mathcal{R}$, such that that there is at least one reply. There are seven possible configurations (Figure 2a). Figure 2b shows the frequency distribution of dyadic motifs in our data. Note that patterns are mutually exclusive. The most frequent dyadic motifs are A and C. According to Figure 2b, it is more likely to observe a reply to a followed user in non-controversial

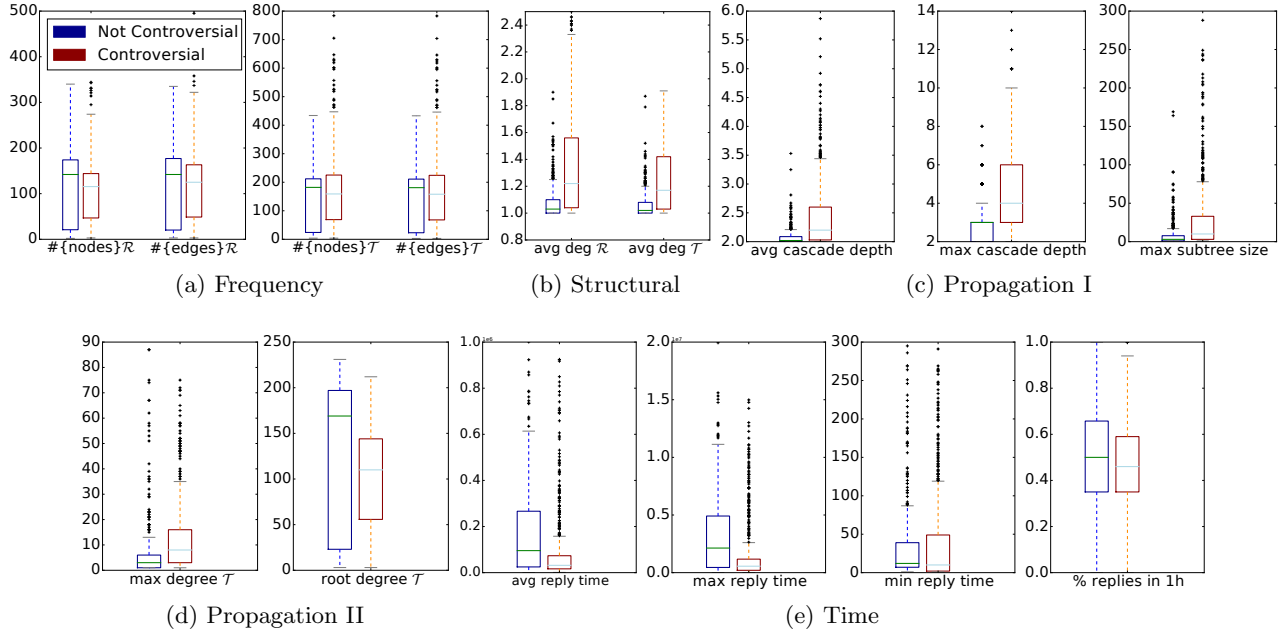(a) Frequency (b) Structural (c) Propagation I

(d) Propagation II (e) Time

Figure 1: (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).

cases. Conversely, in controversial cases it is likely to reply to a user not being followed, confirming our intuitions. The features used for detecting controversial content are the frequencies of all dyadic 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 $\mathcal{G}$. Due to the high number of possible motifs and since most motifs are relatively rare in the data, we coalesce motifs in groups (20 types). The frequency of each group 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. To provide additional insights on user interactions, we consider also the ratio of triangles in the reply graph $\mathcal{R}$ over the number of all possible triangles.

Experiments

Controversy Detection

We evaluated different classifiers, including AdaBoost, Logistic Regression, SVM and Random Forest, and chose AdaBoost as it resulted in the best performance. 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).

The baseline approach accuracy (with structural, propagation-based and temporal features) is above 75%. With the addition of dyadic motifs, all the perfor-

Table 2: 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
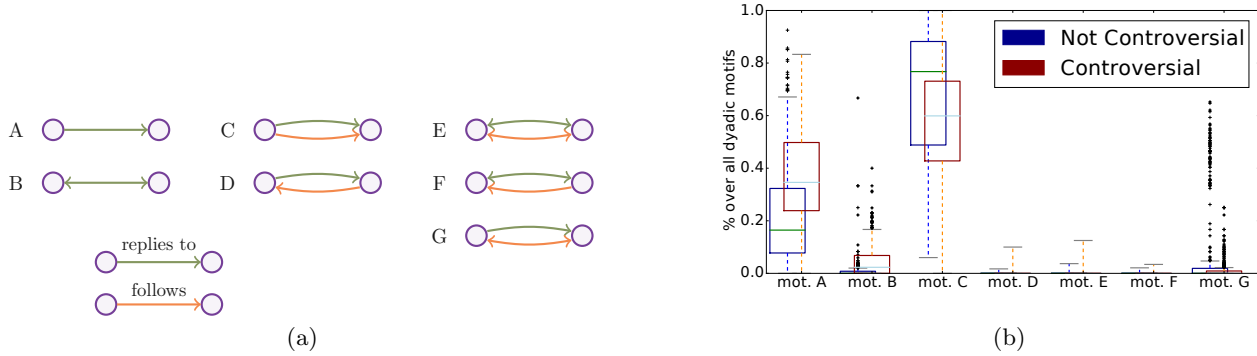
Figure 2: (a) Dyadic motifs and (b) their frequency distribution.

mance figures are significantly improved. The addition of triadic motifs leads to the best results, but the improvement is only marginal because they are infrequent. The best results, highlighted in boldface, are statistically significant w.r.t. baseline features. Using dyadic motifs alone, moreover, the accuracy of the model is comparable with the baseline. We evaluated the importance of the features in the model: the first feature is the average inter-reply time: when the discussion is polarized people tend to reply in a shorter time. The second most important feature is the maximum relative degree. The other features among the top-6 are dyadic motifs. The most relevant being motif A: 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 suggests that it is less likely to have controversial discussions among friends. Interestingly, dyadic patterns seem to be more relevant than propagation-based features. We found also that 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. Studying how the controversy related to a given hashtag evolves over time is an interesting task: for the sake of space we do not include further performed analyses on Twitter hashtags, but they confirm the efficacy of our approach in monitoring controversy over time.

Conclusion

We proposed a novel language-independent approach based on local graph motifs 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. We proved on a benchmark Twitter dataset that such motifs are more powerful in predicting controversy than other baseline frequently used graph properties. We observed that in most cases controversy arise when users participate to discussions beyond their social circles. 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.

Acknowledgments

References

[Coletto et al. 2016] Coletto, M.; Lucchese, C.; Orlando, S.; and Perego, R. 2016. Polarized user and topic tracking in twitter. In *SIGIR 2016, Pisa, Italy*.

[Conover et al. 2011] Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political Polarization on Twitter. In *ICWSM*.

[Garimella et al. 2016] Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2016. Quantifying controversy in social media. In *WSDM*, 33–42. ACM.

[Mejova et al. 2014] Mejova, Y.; Zhang, A. X.; Diakopoulos, N.; and Castillo, C. 2014. Controversy and sentiment in online news. *Symposium on Computation + Journalism*.

[Milo et al. 2002] Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827.

[Smith et al. 2013] Smith, L. M.; Zhu, L.; Lerman, K.; and Kozareva, Z. 2013. The role of social media in the discussion of controversial topics. In *SocialCom*, 236–243. IEEE.