It's Always April Fools' Day!
On the Difficulty of Social Network Misinformation Classification via Propagation Features

Mauro Conti*, Daniele Lain*, Riccardo Lazzeretti†, Giulio Lovisotto*, Walter Quattrociocchi‡

*University of Padua
{conti,dlain}@math.unipd.it,
giulio.lovisotto@studenti.unipd.it

†Sapienza University of Rome
lazzeretti@dis.uniroma1.it

‡Ca Foscari University of Venice
w.quattrociocchi@unive.it

***Abstract*—Given the huge impact that Online Social Networks (OSN) had in the way people get informed and form their opinion, they became an attractive playground for malicious entities that want to spread misinformation, and leverage their effect. In fact, misinformation easily spreads on OSN, and this is a huge threat for modern society, possibly influencing also the outcome of elections, or even putting people's life at risk (e.g., spreading "anti-vaccines" misinformation). Therefore, it is of paramount importance for our society to have some sort of "validation" on information spreading through OSN. The need for a wide-scale validation would greatly benefit from automatic tools.**

In this paper, we show that it is difficult to carry out an automatic classification of misinformation considering only structural properties of content propagation cascades. We focus on structural properties, because they would be inherently difficult to be manipulated, with the the aim of circumventing classification systems. To support our claim, we carry out an extensive evaluation on Facebook posts belonging to conspiracy theories (representative of misinformation), and scientific news (representative of fact-checked content). Our findings show that conspiracy content reverberates in a way which is hard to distinguish from scientific content: for the classification mechanism we investigated, classification F_1-score never exceeds 0.7.

I. Introduction

An increasing number of people get informed on Online Social Networks (OSN) [26]. However, as OSN allow every user to post content, which propagates among users through viral processes, these platforms became attractive targets for misinformation creators. Moreover, the hyperconnected world and increasing complexity of reality create a scenario in which viral processes on OSN are driven by confirmation bias, eliciting the proliferation of unsubstantiated rumors and hoaxes all the way up to conspiracy theories [7, 6]. News stories undergo the same popularity dynamics as other forms of online contents (such as selfies and cat pictures) [26]. It is not a surprise then that the Oxford Dictionary in 2016 elected *Post-truth* as word of the year. The definition reads:

> *"Relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief".*

Several studies pointed out the effects of social influence online [10, 19, 27, 30]. Results reported in [22] indicate that emotions expressed by others on Facebook influence our own emotions, providing experimental evidence of massive-scale contagion via social networks. As a result of disintermediated access to information and of algorithms used in content promotion, communication has become increasingly personalized, both in the way messages are framed and how they are shared across social networks. Selective exposure has been shown to favor the emergence of echo-chambers — polarized groups of like-minded people where users reinforce their world view with information adhering to their system of beliefs [15]. Confirmation bias, indeed, plays a pivotal role in informational cascades [28, 7, 35, 32]. Recent works [6, 35] showed that attempts to debunk false information are largely ineffective. In particular, discussion degenerates when the two polarized communities interact with one another [36]. OSN users therefore tend to select information consistent with their beliefs (even if containing false claims), propagate it to like-minded friends, and ignore information dissenting with their beliefs. This confirms that misinformation is a huge threat for modern society. Not only it can put people's life at risk, as in the case of "anti-vaccines" misinformation, it is also starting to be used against political opponents, such as in the US, during the 2016 electoral campaign[1] and during Election Day[2].

Rising attention to the spreading of fake news and unsubstantiated rumors led researchers to investigate many of their aspects, from the characterization of conversation threads [4], to the detection of bursty topics on microblogging platforms [16], to the disclosure of the mechanisms behind information diffusion for different kinds of contents [29]. Spreading of misinformation also motivated major corporations like

WIFS'2017, December, 4-7, 2017, Rennes, France.
978-1-5090-6769-5/17/$31.00 ©2017 IEEE.

[1] buzzfeed.com/craigsilverman/partisan-fb-pages-analysis

[2] nytimes.com/2016/11/09/us/politics/debunk-fake-news-election-day.html

Google and Facebook to provide solutions to the problem[3]. At the time of writing, searching on Google Search for breaking news (for example whether Hillary Clinton really sold uranium to Russia) prompts, as the first search result, the claimed fact and claimant, and the result of manual fact checking by "trusted" sources (e.g., fact-checking news agencies). Also at the time of writing, Facebook advertises on users' *timelines* their own guidelines against misinformation - most of which rely on users' manual feedback [18].

However, given the amount of content posted every day on OSN (e.g., Facebook reports 1.23 billion daily active users on December 2016[4]), effective fact-checking would greatly benefit from automatic classification tools, possibly not requiring human intervention.

Classification of fake news and misinformation should ideally use properties that misinformation creators can not easily manipulate. For example, considering features such as the trustworthiness of news domains, or the topic of content, could lead to an arms race with creators of false news - who are often financially motivated, and therefore do not care about the specific content of fake news, as long as it attracts clicks on it [18]. Moreover, text content classification techniques can also be easily fooled by the similarity between information and misinformation – which often discuss the same topic from different points of view. The same problem applies to sentiment analysis of comments of news stories. Instead, topology of propagation cascades, and patterns of users' interaction with content, are outside of the domain of our "adversaries" and are much more difficult to be manipulated.

In this work, we investigate detection of viral processes by comparing diffusion of posts from scientific and conspiracy pages on the Italian Facebook network. The former diffuse scientific knowledge, where details about the sources (such as authors and funding programs) are easy to access. Such posts can be representative of fact-checked content. The latter aim at diffusing what is neglected by "manipulated" mainstream media. Specifically, conspiracy theories tend to reduce the complexity of reality by explaining significant social or political aspects as plots conceived by powerful individuals or organizations. Since these kinds of arguments can sometimes involve the rejection of science, alternative explanations are invoked to replace the scientific evidence. For instance, people who reject the link between HIV and AIDS generally believe that AIDS was created by the US Government to control the African American population. Such posts can be therefore representative of misinformation.

Contributions of this paper are the following:

- We show that automatic fact-checking with classification techniques employing only structural features of content propagation cascades (robust to attacker's manipulation) does not suggest to bring usable results. Given the grain of the employed dataset, we design a classifier that leverages topological properties of content propagation cascades, and properties of the evolution over time of users' interactions with content. We classify news after a short timespan of their propagation, as being able to issue warnings about possible fake news during their early stages of propagation can be useful, in the fight against OSN misinformation.
- We evaluate our classifiers on a well-known dataset of Facebook posts from Italian pages. We use posts belonging to conspiracy theories and debunked hoaxes (e.g., chem-trails, anti-vaccines) as representative of misinformation, and posts belonging to scientific, peer-reviewed news as representative of fact-checked content. Our findings suggest that in Facebook users interact with different types of content in similar ways, and highlight the complexity of creating automatic solutions to misinformation classification. Indeed, structural features of content propagation do not allow us to obtain notable improvements from a random guess baseline: classification F_1-score is lower than 0.7 after two days of content propagation.

II. Related Work

Several studies moved towards the spreading of rumors and behaviors on OSN, challenging both their structural properties and their effects on social dynamics [25, 17, 31, 8, 14, 9]. In [33], authors find that the probability of contagion is tightly controlled by the number of connected components in an individual's contacts neighborhood, rather than by the actual size of the neighborhood. In [11] researchers show that, although long ties are relevant for spreading information about an innovation or social movement, they are not sufficient to influence social reinforcement mechanisms.

A key factor in identifying true contagion in social network is to distinguish between peer-to-peer influence and homophily: in the first case, a node influences or causes outcomes to its neighbors, whereas in the second one, dyadic similarities between nodes create correlated outcome patterns among neighbors that could mimic viral contagions even without direct causal influence [24]. The study presented in [1] reveals that there is a substantial level of topical similarity among users which are close to each other in the social network, suggesting that users with similar interests are more likely to be friends. In [3] authors develop an estimation framework to distinguish influence and homophily effects in dynamic networks and find that homophily explains more than 50% of the perceived behavioral contagion. In [5] the analysis faces the role of OSN and exposure to friends' activities in information resharing on Facebook. Once having isolated contagion from other confounding effects such as homophily, authors claim that there is a considerably higher chance to share contents when users are exposed to friends' resharing.

All these contributions strive to understand the inner mechanism of rumor spreading and to eventually predict massive diffusion processes, i.e. cascades. Cascades recurrence and prediction has been shaped in [12] and [13].

[3] firstdraftnews.com/about/

[4] newsroom.fb.com/company-info/

III. Methods

In this section, we first report and describe the employed dataset, and then we present our reference model and definition of propagation mechanism in Facebook.

A. Dataset

We employ a well-known dataset of posts shared by Italian Facebook users [6]. This dataset contains posts published by 73 public Facebook pages: 34 pages that publish scientific content (e.g., press releases of peer-reviewed articles), and 39 pages that publish conspiracy theories-related content (e.g. new world order, chem-trails). The dataset contains information about the interaction of users with these posts, and users' ego-networks (i.e., the list of users that are their friends, when such list is public). For a set of posts, the dataset provides information about the *propagation cascade* of such content, generated by users' reshares, and subsequent reshares from their friends. This propagation from one user to other users can happen multiple times, forming a cascade of resharing. Using information from the dataset, we extracted 112141 non-empty propagation cascades, 89491 for conspiracy and 22650 for science, respectively. We underline that the dataset is obtained by using the Facebook Graph API, and contains only public information. Hence, timestamps of *reshares* and *comments* are available, but timestamps of *like* interactions are not.

B. Background and Definitions

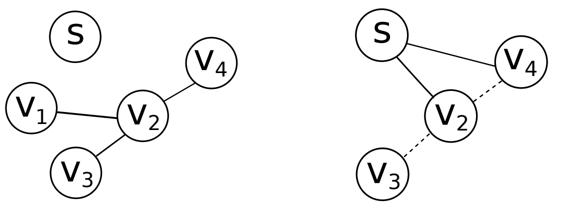
We now present our reference formal representation of Facebook's *friendship graph*, and the *potential propagation graph* generated by content posted on the social network.

Facebook Friendship Graph. We model Facebook relationships as a graph $\mathcal{G}\langle V, E\rangle$, that we call *Facebook friendship graph*, where V is the set of nodes that represent *entities*, namely user accounts and page accounts. We assume two main differences among these two types of entity: (i) pages can post new content on the OSN, while users can only interact with such content by liking, commenting, and resharing it; and (ii) users can establish *friendship* relationships with other users, while pages cannot. Indeed, two users $v_1, v_2 \in V$ are connected by an edge $e(v_1, v_2) \in E$ if they are *friends* on Facebook. Pages are not connected by any edge, as they do not have proper friendship relationships. This model is a simplification of how Facebook actually works, because users can post new content, and pages and users are linked by *like* relationships. Similarly, users can *follow* other users, without having any friend relationship with them. However, for this work, we do not focus on new content generated by users. Moreover, the dataset we use lacks information about the *like* and *follow* relationships, that we therefore can not consider.

Potential Propagation Graph. Before formally modeling the spreading of content on Facebook, we describe fundamental concepts of the OSN. We recall that, in our model, only pages can post new content on the social network. Henceforth, we refer to the page that originally posted some content as the *seed page*. Instead, users find new content by looking at their *timeline*, where they see recent content posted by the pages they like, and content that their friends recently interacted with. They can then interact with such content through *resharing*, *liking*, and *commenting* it. However, all of these interactions can happen directly on the original content, or on some types of interactions of users' friends. For example, a user observing a comment or reshare of a post by one of his friends may interact directly with the original post, or with his friend's interaction itself. In the first case, user's interaction looks exactly the same and it is impossible to understand whether the content was found thanks to his friend's interaction, or directly on the seed page. In fact, differently from related work [20], the dataset we employ does not provide this type of information. We therefore take a conservative approach, saying that the content *potentially propagated* to the user both from the seed page and from any of his friends who interacted with the content, without distinction. On the other hand, in the second case, it becomes clear that the user found the content thanks to his friend. We therefore say that the content *propagated* to the user from his friend.

We now formalize the above observations. Let P be the set of contents posted on Facebook by seed pages. For each post $p \in P$, created at some time t by a seed page, at a generic subsequent point in time $t + \delta$ we define a *potential propagation graph* $\mathcal{G}^p_{t+\delta}\langle V^p_{t+\delta}, E^p_{t+\delta}\rangle$, where $V^p_{t+\delta}$ is the set composed by the seed page, and by the users that interacted with p during the time interval $[t, t+\delta]$. *Final potential propagation graph* $\mathcal{G}^p\langle V^p, E^p\rangle$ is the graph formed considering all interactions with p on the timespan of the analysis (as new interactions with old content are always possible on Facebook). Two nodes $v_1, v_2 \in V^p_{t+\delta}$ are connected by an undirected *potential propagation edge* $e^p(v_1, v_2) \in E^p_{t+\delta}$ if either (i) v_1 or v_2 already interacted with p and $\exists e \,|\, e(v_1, v_2) \in E$ (that is, v_1 and v_2 are friends on Facebook), or (ii) v_1 or v_2 is the seed page. Therefore, an edge $e(v_1, v_2) \in E^p_{t+\delta}$ indicates that the content p *potentially propagated* either from v_1 to v_2, or from v_2 to v_1. Edges $e \in E^p_{t+\delta}$ have two different associated properties: (i) *time* when the interaction between v_1 and v_2 took place, and (ii) *type* of the interaction between v_1 and v_2, that is either "like", "comment", "reshare", or "friendship".

Figure 1 depicts an example of this propagation model. We represent a simple Facebook friendship graph in Figure 1a,



(a) Facebook Friendship Graph

(b) Potential Propagation Graph

Fig. 1: Sample Facebook friendship graph, and potential propagation graph of some content. A dashed line represents potential propagation edges that also corresponds to a friendship edge of the friendship graph. As node v_1 does not interact, it is not present in the potential propagation graph.

where edges represent friendship relations. Nodes $v_1, \dots, v_4$ are OSN users, while node s represents a public page posting content on the platform. We suppose v_2 and v_4 interacted with a given content posted by s, and v_3 reshared the post from v_2. Node v_1 did not interact with the content, hence is not present on the potential propagation graph of the post, that we represent in Figure 1b. There, with edges $(s, v_2), (s, v_4), (v_2, v_4)$ we represent possible propagation paths of the content: either v_2 or v_4 could have seen the content from the seed page, or from previous interaction of their friend. Node v_3 only has an edge (v_2, v_3), as we know that the interaction happened thanks to v_2, and therefore the content propagated from this node. Additionally, we highlight that some possible propagation edges, represented with a dashed line, correspond to friendship edges of the friendship graph.

IV. Experiments

In this section, we describe the experimental design and we motivate the features we extracted from propagation cascades.

A. Experimental Design

The aim of our experiments is to show that it is difficult to discriminate between conspiracy theories, and fact-checked scientific news, by using only the propagation graph of the post, as it would be difficult to manipulate by misinformation creators. We evaluate this at the final stage of the propagation, meaning that the classification happens when the post already stopped propagating, and its cascade is complete. Being able to perform such a classification would serve as a warning in order to prevent its further diffusion. Moreover, retroactively flagging old posts as potential misinformation would help users to discriminate between fact-checked and dubious information, a major direction in the fight against misinformation.

To investigate this scenario, we set up binary classification experiment, where the two classes are *conspiracy* and *science*. We describe the final propagation graph $\mathcal{G}^p$ of post p with a set of high-level, topological, and evolution features, that we describe in more detail in Section IV-B. Since we observed that most of the interactions with a post happen within the first two days after its publication (>98%), we suppose that a post mostly propagates within this period. Analyzing longer periods would introduce only a few nodes and edges in the propagation graph. We compare the performance of two well-known classifiers, namely Random Forests (RF) and Support Vector Machine (SVM) in a cross-validation scheme.

B. Propagation Features

To extract information from the different propagation graphs $\mathcal{G}^p$ and $\mathcal{G}^p_t$, we identify three possible categories of features: (i) high-level properties of the content propagation, (ii) topological properties of the propagation graphs, (iii) evolution properties of the content propagation.

High Level Properties. These features represent high-level properties of the complete propagation cascade $\mathcal{G}^p\langle V^p, E^p\rangle$. Some represent very general properties related to the virality of the content: *lifetime* of the cascade, measured as the distance in minutes from the first to the last captured interaction with the content; *size* of the cascade in terms of number of nodes (users who interacted with the content); *number of total interactions*; and time required for the cascade to reach its 90% total interactions (referred to as *90% interactions time*).

Other such features capture different types of interaction with the content. *Friendships ratio* is defined as the proportion of edges whose type is "friendship" over the total number of edges – representing the number of times, in proportion, that the post potentially propagated among friends, rather than directly from the seed node. Indeed, if no friends of spreaders interact with some content, its potential propagation graph only contains edges with the seed page. *Interactions ratio*, instead, represents the average exposure to interactions from friends of users with the content. Since vertices are interacting users, and edges are potential interactions, higher values of this metric mean lower exposure (little interaction with the content by one's friends). These features are motivated by the observation that it is possible that the users' fruition of different types of content is different, with some types of content being interacted directly from the source, and other types of content relying more on word-of-mouth propagation [29].

Topological Properties. We also select as features some well-known topological properties of graphs, that have been successfully applied in link analysis and prediction [2], cascade and virality prediction [21, 12], and especially rumor propagation in OSN [23]. *Average vertex degree* represents the average number of possible propagation edges for the content at any given hop. Higher values of this metric indicate the presence of interacting users greatly exposed to the content, or able to influence many of their social friends. The *global clustering coefficient*, a measure of the density of connections of graphs, is another indication of whether the possible propagation paths are generated by interactions between friends, or directly with the seed page. *Assortativity coefficient*, defined as the degree correlation between pairs of linked nodes, can measure how friends influence each others in interacting with content on the social network. *Average path length*, also known as Wiener index, gives us indications of the virality of the content, in terms of distance of propagation from the seed page. Long cascading news, reshared many times from interacting friends, will exhibit a longer average path length than news whose interactions happened mostly from the seed node. Finally, *diameter*, defined as the longer shortest path between any pair of nodes of the graph, indicates the spreading distance of posts.

Evolution Properties. These features represent evolution properties of the propagation of a post p over time, from its creation time t to a subsequent point $t + \delta$. To compute these features, we construct propagation graphs at different time steps $\mathcal{G}_{t+30}, \dots, \mathcal{G}_{t+\delta}$. We select 30 minutes steps, as this proved to be a good trade-off between the granularity of the analysis and the number of intervals to consider. We then calculate the value of three of our high-level features for each graph at each time step: (i) Friendships Ratio, (ii) Size, and (iii) Interactions Ratio. For each of these high-level feature, we obtain a time series $v_1, v_2, \dots, v_{\delta/30}$, on which we compute *mean*, *standard deviation*, *linear weighted mean* and *quadratic weighted mean*, *average absolute change*, and *maximum*, to represent its evolution over time [34]. We derive time-series only for the three high-level features listed above: the other features either have no temporal properties (e.g., lifetime, time to reach 90% interactions), evolve in similar or predictable ways (e.g., diameter), or describe behaviors that are already captured by the selected time series.

V. Results

In this section we motivate the chosen evaluation metrics, and we report the results of our experiments.

A. Evaluation Metrics

As usual in binary classification, the classification baseline is the performance of a random classifier on the data: without any information regarding the propagation graph, it guesses either *science* or *conspiracy* with equal probability. The goal of our experiments is to show that structural features do not help sophisticated models in improving the baseline performance.

Unfortunately, our dataset is highly imbalanced (composed by 89491 news for *conspiracy*, and only 22650 news for *science*). With such imbalance, standard evaluation metrics (such as *precision*, *accuracy*, and *recall*) can be misleading, because they do not account for the uneven class frequencies. Even computing averages of these metrics using weights based on the class frequencies does not fit our intentions of consistently comparing our results with the fixed baseline. To deal with this imbalance, we performed two distinct experiments: (i) we consider only metrics that take imbalance into account and use the full dataset, and (ii) we consider meaningful metrics with balanced dataset, obtained *undersampling* the original dataset.

To perform (i), as metrics we use Area Under Receiver Operating Curve (AUC), and Cohen's Kappa (scaled into the interval [0, 1]). Value of these metrics for a random classifier is 0.5, which we use as baseline. This way, we can use the full dataset and be able to compare our results with the baseline.

To perform (ii), we undersampled the most frequent class (i.e., *conspiracy*), with no replacement. We therefore extracted exactly 22650 *conspiracy* samples from the full dataset, and created a subsample with perfectly balanced classes. To account for possible biases caused by the undersampling, we repeated the process several times, and averaged the outcomes. Using perfectly balanced datasets, we can evaluate *precision*, *recall*, *accuracy*, and F_1-*score* values of our classifiers. Indeed, the value of these metrics for a random classifier on balanced data is exactly 0.5, which we use as baseline.

Hereafter, the metrics that require a positive class (such as *precision*, *recall*, and F_1-*score*) use *conspiracy* as the positive class, and *science* as the negative one.

B. Classification Performance

Figure 2 reports average AUC and Cohen's Kappa on a 5-fold cross-validation scheme, on the full dataset, for the two classifiers. The dotted horizontal line shows the baseline performance of random classification. We observe that our propagation classifiers do not significantly improve the baseline, with the metrics remaining below 0.75. Although 0.75 might seem an acceptable result, we observe that it originates from a good true positive rate (i.e., most of the *conspiracy* cascades are correctly identified), and from a very high false positive rate (i.e., lots of *science* cascades are mistakenly identified as *conspiracy*). These rates cause a significant number of false positives, which would perform poorly in real-world applications: having most science-related news classified as misinformation would not be acceptable for such a system.

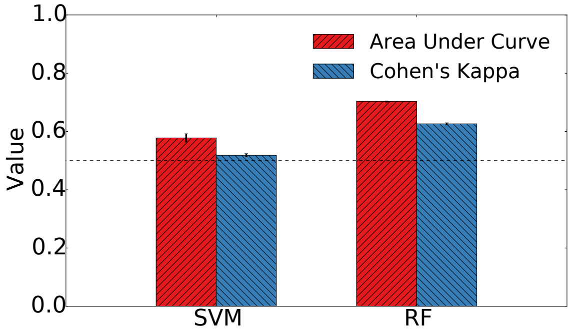


Fig. 2: Classification AUC and Cohen's Kappa.

In Figure 3 we report the ROC curve, on a single fold of the cross-validation scheme, on the full dataset. From Figure 3 we can observe that no classifier can reach a good tradeoff between true positive rate and false positive rate. Indeed, the curves are relatively close to the baseline (diagonal dotted line), meaning that, as the decision threshold changes, lots of samples are misclassified.

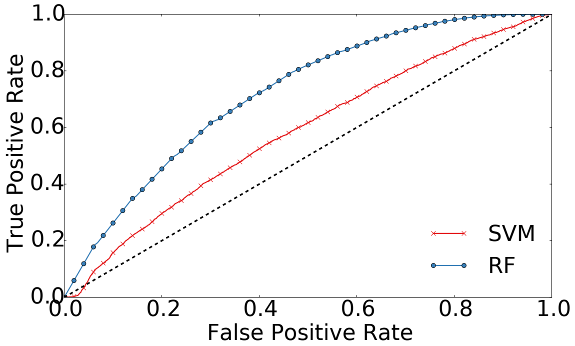


Fig. 3: Classification ROC curves.

Table I reports several metrics computed on the undersampled balanced dataset. Results are averaged on a 5-fold cross-validation scheme, and then over ten repetitions of the undersampling procedure. These metrics show that even if the classifiers are able to improve the baseline slightly, they can not reach good performance.

TABLE I: Classification performance on balanced dataset.

Classifier	**Precision**	**Recall**	**Accuracy**	F_1 **score**
SVM	0.583	0.549	0.578	0.565
RF	0.669	0.734	0.685	0.700

VI. Conclusions

Detection of misinformation plays a crucial role in social networks. In this paper, we analyzed the difficulty of discerning conspiracy posts from scientific posts on Facebook. We focused on using only structural features of content propagation, because they cannot be easily manipulated by misinformation creators. Our results show that misinformation classification considering the cascade at the end of content propagation does not help: the improvement provided by a classifier over random coin flips is negligible.

Our findings suggest that in Facebook users interact with different types of content in similar ways, reinforcing the hypothesis of echo chambers [15]. Inside chambers, strongly polarized by topic [7], content propagation exhibits similar structural properties, that are therefore less useful in content classification. These results highlight the necessity of including content-related features, or polarization metrics, in future analysis (i.e., whether users and their echo chambers are more polarized towards one type of content). Unfortunately, misinformation creators can easily control content-related features, to avoid algorithmic detection. Moreover, it takes time to understand polarization of new OSN users. Hence, automatic detection of fake news remains an open challenge.

Acknowledgments

Mauro Conti is supported by a Marie Curie Fellowship funded by the European Commission (agreement PCIG11-GA-2012-321980). This work is also partially supported by the EU TagItSmart! Project (agreement H2020-ICT30-2015-688061), the EU-India REACH Project, (agreement ICI+/2014/342-896), by the project CNR-MOST/Taiwan 2016-17 "Verifiable Data Structure Streaming", the grant n. 2017-166478 (3696) from Cisco University Research Program Fund and Silicon Valley Community Foundation, and by Intel grant "Scalable IoT Management and Key security aspects in 5G systems".

References

[1] L.M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. "Friendship Prediction and Homophily in Social Media". In: *ACM TWEB* 6.2 (2012), pp. 1–33.

[2] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. "Link Prediction using Supervised Learning". In: *IEEE ICDM Workshop on Link Analysis, Counter-Terrorism and Security*. 2006.

[3] S. Aral, L. Muchnik, and A. Sundararajan. "Distinguishing Influence-based Contagion from Homophily-driven Diffusion in Dynamic Networks". In: *PNAS* 106.51 (2009), pp. 21544–21549.

[4] L. Backstrom, J.M. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil. "Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-Entry". In: *ACM WSDM*. 2013.

[5] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. "The Role of Social Networks in Information Diffusion". In: *ACM WWW*. 2012, pp. 519–528.

[6] A. Bessi, M. Coletto, G.A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. "Science vs Conspiracy: Collective Narratives in the Age of Misinformation". In: *PLOS ONE* 10.2 (2015), pp. 1–17.

[7] A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi. "Viral Misinformation: The Role of Homophily and Polarization". In: *ACM WWW*. 2015.

[8] J. Borge-Holthoefer, S. Meloni, B. Gonçalves, and Y. Moreno. "Emergence of Influential Spreaders in Modified Rumor Models". In: *Journal of Statistical Physics* 151.1-2 (2013), pp. 383–393.

[9] J. Borge-Holthoefer, A. Rivero, and Y. Moreno. "Locating Privileged Spreaders on an Online Social Network". In: *APS Physical Review E* 85.6 (2012), p. 066123.

[10] D. Centola. "The Spread of Behavior in an Online Social Network Experiment". In: *Science* 329.5996 (2010), pp. 1194–1197.

[11] D. Centola and M. Macy. "Complex Contagions and the Weakness of Long Ties". In: *American Journal of Sociology* 113.3 (2007), pp. 702–734.

[12] J. Cheng, L. Adamic, P.A. Dow, J.M. Kleinberg, and J. Leskovec. "Can Cascades Be Predicted?" In: *ACM WWW*. 2014.

[13] J. Cheng, L.A. Adamic, J.M. Kleinberg, and J. Leskovec. "Do Cascades Recur?" In: *ACM WWW*. 2016, pp. 671–681.

[14] E. Cozzo, R.A. Banos, S. Meloni, and Y. Moreno. "Contact-based Social Contagion in Multiplex Networks". In: *APS Physical Review E* 88.5 (2013), p. 050801.

[15] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H.E. Stanley, and W. Quattrociocchi. "The Spreading of Misinformation Online". In: *PNAS* 113.3 (2016), pp. 554–559.

[16] Q. Diao, J. Jiang, F. Zhu, and E.P. Lim. "Finding Bursty Topics from Microblogs". In: *ACL*. 2012, pp. 536–544.

[17] B. Doerr, M. Fouz, and T. Friedrich. "Why Rumors Spread so Quickly in Social Networks". In: *Communications of the ACM* 55.6 (2012), pp. 70–75.

[18] Facebook. *Working To Stop Misinformation and False News*. https://newsroom.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/. [Online; accessed 23-June-2017]. 2017.

[19] J.H. Fowler and N.A. Christakis. "Cooperative Behavior Cascades in Human Social Networks". In: *PNAS* 107.12 (2010), pp. 5334–5338.

[20] A. Friggeri, L.A. Adamic, D. Eckles, and J. Cheng. "Rumor Cascades". In: *AAAI ICWSM*. 2014.

[21] L. Hong, O. Dan, and B.D. Davison. "Predicting Popular Messages in Twitter". In: *ACM WWW*. 2011.

[22] A.D.I. Kramer, J.E. Guillory, and J.T. Hancock. "Experimental Evidence of Massive-scale Emotional Contagion through Social Networks". In: *PNAS* 111.24 (2014), pp. 8788–8790.

[23] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. "Prominent features of rumor propagation in online social media". In: *IEEE ICDM*. 2013, pp. 1103–1108.

[24] M. McPherson, L. Smith-Lovin, and F.M. Cook. "Birds of a Feather: Homophily in Social Networks". In: *Annual Review of Sociology* (2001), pp. 415–444.

[25] Y. Moreno, M. Nekovee, and A.F. Pacheco. "Dynamics of Rumor Spreading in Complex Networks". In: *APS Physical Review E* 69.6 (2004), p. 066130.

[26] Newman, N. and Levy, D.A.L. and Nielsen, R.K. *Reuters Institute digital news report*. http://www.digitalnewsreport.org/survey/2015/. [Online; accessed 23-June-2017]. 2015.

[27] W. Quattrociocchi, G. Caldarelli, and A. Scala. "Opinion Dynamics on Interacting Networks: Media Competition and Social Influence". In: *Scientific Reports* 4 (2014).

[28] W. Quattrociocchi, A. Scala, and C.R. Sunstein. "Echo Chambers on Facebook". In: *Available at SSRN: https://ssrn.com/abstract=2795110* (2016).

[29] D.M. Romero, B. Meeder, and J.M. Kleinberg. "Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter". In: *ACM WWW*. 2011, pp. 695–704.

[30] M.J. Salganik, P.S. Dodds, and D.J. Watts. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market". In: *Science* 311.5762 (2006), pp. 854–856.

[31] E. Seo, P. Mohapatra, and T. Abdelzaher. "Identifying Rumors and their Sources in Social Networks". In: *SPIE Defense, Security, and Sensing*. 2012, pp. 83891I–83891I.

[32] C.R. Sunstein. "The Law of Group Polarization". In: *Journal of Political Philosophy* 10.2 (2002), pp. 175–195.

[33] J. Ugander, L. Backstrom, C. Marlow, and J.M. Kleinberg. "Structural Diversity in Social Contagion". In: *PNAS* 109.16 (2012), pp. 5962–5966.

[34] J. Wiens, E. Horvitz, and J.V. Guttag. "Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task". In: *NIPS* (2012), pp. 467–475.

[35] F. Zollo, A. Bessi, M. Del Vicario, A. Scala, G. Caldarelli, L. Shekhtman, S. Havlin, and W. Quattrociocchi. "Debunking in a World of Tribes". In: *arXiv* (2015). preprint, http://arxiv.org/abs/1510.04267.

[36] F. Zollo, Petra K. Novak, M. Del Vicario, A. Bessi, I. Mozetič, A. Scala, G. Caldarelli, and W. Quattrociocchi. "Emotional Dynamics in the Age of Misinformation". In: *PLOS ONE* 10 (Sept. 2015), pp. 1–22.