Data Challenges in Disinformation Diffusion Analysis

(Abstract)

Paolo Papotti
papotti@eurecom.fr
EURECOM

1 THE NEED FOR BETTER DIFFUSION NETWORKS

Social media enable fast and widespread dissemination of information that can be exploited to effectively spread disinformation by bad actors [1]. We refer to disinformation as the malicious and coordinated spread of inaccurate content for manipulation of narratives[1]. It has been showed that social media disinformation has effectively reached millions of people in state-sponsored campaigns[2].

Several computational solutions have been proposed for the identification of coordinated campaign on a single platform [12]. They study how content is disseminated across a network of inter-connected users. However, two main practical challenges limit the impact of such approaches.

First, existing approaches focus on a *single sources*, such as Twitter or Reddit. Unfortunately, misinformation campaigns span multiple platforms and there is a recognized need to jointly analyze the diffusion of content across different sources, such as social networks, online forums (e.g., Reddit), and traditional news outlets (e.g., comments in reputable sites).

Second, the content diffusion graphs that are currently generated from social network APIs are *limited in quality*. For example, only the information about content re-posting (e.g., re-tweets) of a user is directly provided. But information is disseminated also by manipulating the original content to add bias, "evidence", or propaganda material. Moreover, fine granularity of the re-posting is not available, with the recognized problem of star-effect for re-tweets that can heavily degrade the quality of the network model [13].

Consider the example in Figure 1 that shows the coordinated sharing of the same initial piece of content (say, a textual news) by three users over different platforms. With the current infrastructure and APIs, a journalist or a fact-checker willing to study the diffusion network would look at each network in isolation. S/he would be able to follow content across users (nodes) only when they re-post explicitly (full edges across nodes).

In this example, the information would not be enough for the early identification of the coordinated campaign started by the three users. Looking at only one source with limited information does not enable the analytics, neither in terms of scope nor evidence, that we need to identify and understand how false and biased content is used in online campaigns [12].

To overcome this limitation, recent approaches explore evidence across users and platforms, such as coordinated link sharing [7]. While this signal has proven to be useful, we believe this

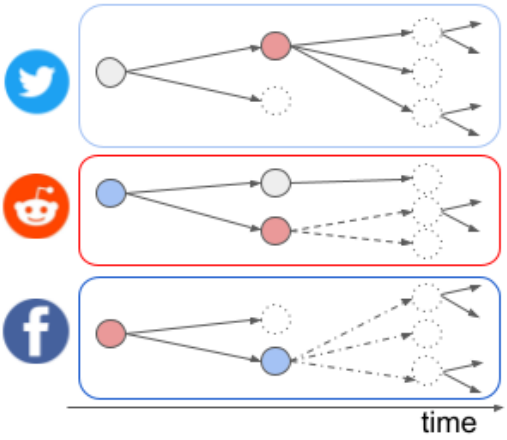

Figure 1: Coordinated campaign for the same content across three social platforms. A node with the same color denote the same user, dashed arrows denote manipulated content.

is just one example of the richer kind of *metadata* that is needed for better diffusion networks.

In fact, to tackle the first challenge, diffusion networks should be *heterogeneous*, covering multiple platforms, with the ability to recognize the same content and the same users across services. In Figure 1, users that refer to the same real world person are annotated with nodes having the same color. Also, to handle the second challenge, the edges should be *typed* with fine-grained metadata that model different actions in the spread of the content.

We believe that data here plays a role as important as the algorithms used for the analysis and therefore more attention is deserved to the problem of creating such richer diffusion networks. Their creation can lead to better identification of coordinated efforts [11] and ultimately allow the analysis of disinformation campaigns in terms of actors, space, and time.

The goal is therefor to develop methods for modeling and creating the rich diffusion networks from the existing platform APIs. The resulting networks can be exploited to assist users, such as fact-checkers and journalists, in

(1) monitoring sources at scale and recognize misleading information (in terms of false or biased textual content) on social networks and forum websites;
(2) tracking the spread and diffusion of the content in terms of time and actors;
(3) generating visualizations that support the fight against misinformation and related literacy efforts.

This network generation is indeed challenging, as the desired metadata is not available and hard to profile automatically in an accurate way. We discuss next two research directions that we identify as critical to tackle these challenges.

[1]The observations in this work apply also for *mis*information, where actors spread incorrect content unintentionally.
[2]E.g., https://transparency.twitter.com/en/reports/information-operations.html

2 RESEARCH DIRECTIONS

The goal is to develop methods for the automatic modeling of content manipulation and diffusion across time and different media sources, such as social networks, forums, and news outlets.

Not only we want the diffusion graph for a given content to be across sources and very well described in terms of information, we also want it (i) to preserve precisely the *provenance* of the data (who created and shared, how and when) and (ii) to be as much as possible *automatic* in its creation, both to handle the Web scale and to not put additional burden on the users. There are therefore several challenges that we need to overcome

- different sources do not contain any readily available information to connect users/content across networks and the automatic matching is a difficult task in both cases;
- labeling the content in terms of being false, manipulated, and biased requires deep understanding of the language and of the reference and background information;
- Web scale implies massive ingestion from heterogeneous data sources, but we would prefer tools that can be used by end users on their machines for confidentiality;
- support for different languages as we aim at helping users across different countries.

Given the challenges above, a natural first line of work is to conduct data integration research to generate a unified representation from heterogeneous, non-aligned sources. A second line of work is to deploy natural language processing (NLP) techniques to profile the content and enrich the graph with typed nodes and edges.

Data integration. In the first line of research, the aim is the online creation of a dissemination network for a given textual content. Given a textual article, for example, the first task is the identification of its citations and appearance across sources (online articles, boards in forums, social posts) and time. This is not trivial as one requirement is to go beyond the identification of content by links, which act as unique identifiers. For this goal, one promising direction is to exploit text-matching literature [14] to identify also manipulated texts that express the original input content. The goal is to have one diffusion network, as in Figure 1 for every given content to analyze, such as a web page, a social message post, or a generic textual claim. The linking and merging of actors across sources, *nodes* in the graphs, is also important. This can be modelled as an entity resolution problem, from a data integration perspective, for example by using deep learning techniques [4, 15]. However, the task is especially challenging in real settings where we drop assumptions about trusted information about the user accounts.

Example 1. Consider a textual article A about a new vaccine that circulates on social platform 1. Existing APIs allow the tracing of the diffusion of the specific content A on platform 1 across users $u_1^1, \ldots, u_n^1$, but the same content may be circulating in a different form (A') and across a different platform 2 by users $u_1^2, \ldots, u_m^2$. We aim at identifying that the two posts refer to the same content ($A = A'$) and at matching the subset of users that are sharing the article across the two networks (e.g., $u_3^1 = u_6^2$).

Metadata from text. Existing NLP tools should be extended and integrated to characterize the nature of the interactions across actors w.r.t. the specific content. This can lead to labelling the *edges* in the graph with information (metadata) about the interaction between the nodes (actors). Possible metadata for such edges include the *nature* of the connection between two users, if it is based on friendship or topic affinity [3], if an node is likely to be a *bot* [6], if the content has been *manipulated* by inserting false claims or bias in the language [5, 10]. In this line of research, it seems promising to exploit both linguistic analysis of the text and external knowledge. The latter could be modelled as reference information in relational datasets [10], knowledge-graphs [2], or check corpora [8, 14]. Recent results show that transformer-based language models and query generation techniques can automatically detect text containing false claims[3] and therefore provide valuable metadata to enrich the network.

Example 2. Consider again the article example from *Example 1.* When it is shared across users, some of them introduce incorrect statistics about its impact ("it works only for young people"), or facts that are not supported by any source ("it will cost 100$ per dose"). We aim to enrich the network edges by recognizing how the content goes from its form A to a new form A^* when it is shared by a certain user u.

We believe that an effective solution to the problem of creating diffusion networks for textual content across heterogeneous sources would enable better disinformation campaigns detection. The resulting graph with typed nodes (persons, organizations) and typed relationships (copy or manipulation in terms of content or form) can be then analyzed with existing methods such as clustering and geometric deep learning [9, 12], or with novel methods that take full advantage of the new information and better identify emerging coordinated campaigns.

REFERENCES

[1] 2020. Cross-platform disinformation campaigns: lessons learned and next steps. *The Harvard Kennedy School (HKS) Misinformation Review* (2020). https://doi.org/10.37016/mr-2020-002

[2] Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable Fact Checking with Probabilistic Answer Set Programming. In *TTO*.

[3] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2014. Who to follow and why: link prediction with explanations. In *SIGKDD*. ACM, 1266–1275.

[4] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. 2020. Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. In *SIGMOD*. ACM, 1335–1349.

[5] Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten Social Dimensions of Conversations and Relationships. In *WWW*. ACM / IW3C2, 1514–1525.

[6] Emilio Ferrara, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[7] Fabio Giglietto, Nicola Righetti, Luca Rossi, and Giada Marino. 2020. Coordinated Link Sharing Behavior as a Signal to Surface Sources of Problematic Information on Facebook. ACM.

[8] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *KDD*.

[9] Sameera Horawalavithana, Kin Wai Ng, and Adriana Iamnitchi. 2020. Twitter Is the Megaphone of Cross-platform Messaging on the White Helmets. In *Social, Cultural, and Behavioral Modeling*. 235–244.

[10] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification. *Proc. VLDB Endow.* 13, 11 (2020), 2508–2521.

[11] Franziska B. Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication* 37, 2 (2020), 256–280. https://doi.org/10.1080/10584609.2019.1661888

[12] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake News Detection on Social Media using Geometric Deep Learning. *CoRR* abs/1902.06673 (2019).

[13] Francesco Pierri, Carlo Piccardi, and Stefano Ceri. 2020. Topology comparison of Twitter diffusion networks reliably reveals disinformation news. *Sci. Rep.* 10 (2020). https://doi.org/10.1038/s41598-020-58166-5

[14] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *ACL*. 3607–3618.

[15] Saravanan Thirumuruganathan, Nan Tang, Mourad Ouzzani, and AnHai Doan. 2020. Data Curation with Deep Learning. In *EDBT*. 277–286.

[3] E.g., https://coronacheck.eurecom.fr