Text networks: foundations and structural analysis

Davide Vega[1] and Matteo Magnani[1]

Department of Information Technology, Uppsala University, Uppsala, Sweden
davide.vega@it.uu.se
matteo.magnani@it.uu.se

1 Introduction

A large amount of human generated information is available online in the form of text exchanged between individuals or groups. Examples include social network sites, online forums and emails. The public accessibility of several of these sources allows us to observe our society at various scales, from conversations among small groups of individuals to the effects of misinformation on large communities.

To cope with the complexity of online information, researchers have typically focused on either the topology of the network, as commonly done in Network Science, or the text exchanged among individuals, using methods from Computational Linguistics. In both cases time has also been taken into consideration, as in Temporal Networks or Temporal Information Retrieval.

In this work, **we introduce an attributed multilayer model for temporal text networks**, enabling the application of a wide range of existing methods to this context. Our model can represent all the information contained in the aforementioned data sources, including different types of text interactions, such as direct messages exchanged between individuals, multicast information targeting specific communities or broadcast news.

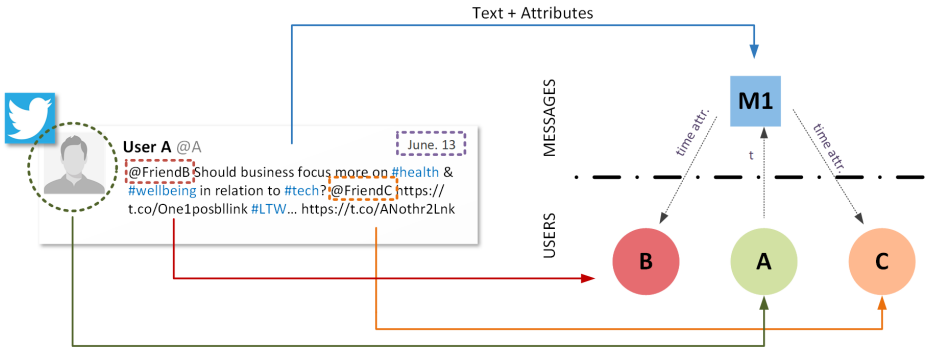

Fig. 1: Codification example of online information from Twitter.

We also introduce two main approaches to analyze text networks, that we call *discrete* and *continuous*. In the discrete model text messages can be classified into a num-

ber of (possibly overlapping) groups. In the continuous approach, a comparison function generates a continuous score indicating the similarity of the text messages (e.g., based on their content).

Finally, we introduce various new types of projections based on the discrete and continuous models. A projection is a generic multilayer network operator that creates edges in one layer based on the information present in another layer. We show that starting from our unified multilayer model of text networks we can project its information into several derived types of networks, such as communication networks, user-annotated networks, topic-based multiplex networks and information propagation networks for which existing analysis methods exist.

2 Results

To exemplify the potential of our model, we will present the results of the analysis of a Twitter dataset containing all the tweets with hashtag *#LTW*, which was the main hashtag used during the London Tech Week — a major technological conference held in London from June 12 to June 16, 2017. By coding each tweet as in Fig. 1, we obtain a multilayer network with 934 users, their follower/followed relations and the 4898 messages exchanged between them (See Fig. 2 *left*).

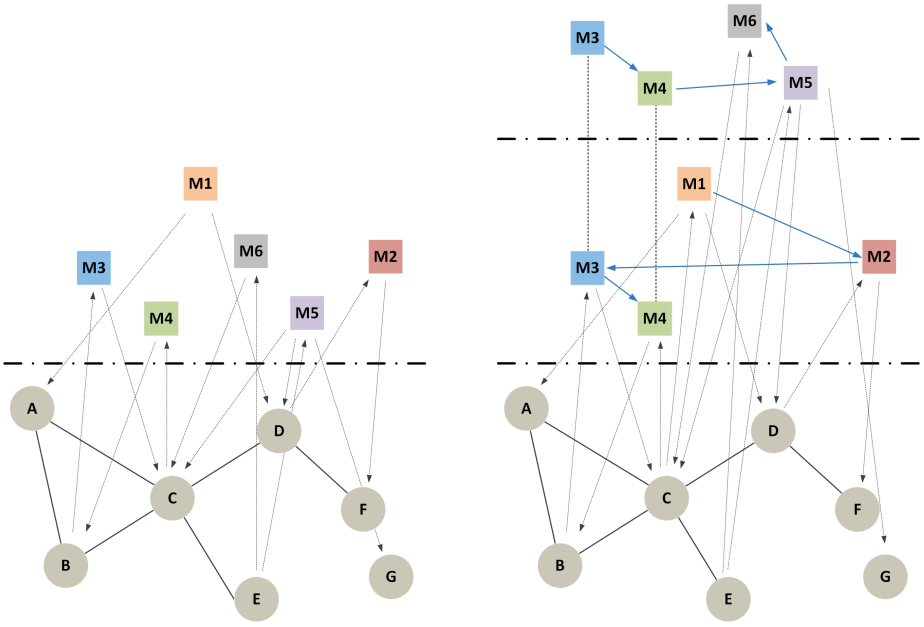

Fig. 2: Attributed multilayer model for temporal text networks (*left*) and one possible projection of the text layer based on topic analysis (*right*).

We can now apply a text discretization by identifying the topics of the tweets, so that each topic is projected into a separate layer (See Fig. 2 *right*). This creates a multiplex network where interactions on different topics are coded into different layers. In our working example we have identified the topics by weighting and clustering the hashtags, obtaining as a result: *education*, *womenintech*, *healthcare*, *hackathon*, *smartcities*, *technology*, *startups*, *ai*, *financial*, *iot*, *menabling17*, *drones*, *industry*, *telecom*, *cloud*. So, our example tweet will be represented as two interconnected nodes present in the *healthcare* and *technology* layers.

At this point we can apply standard multilayer methods [1] [2], obtaining as a result information about the three components: the network structure, the topics and their time evolution. For example, using the abacus [3] community detection algorithm for multilayer networks we are able to identify not only communities of users strongly connected or sets of tweets about similar topics, but also topic-specific discussion groups.

For space reasons, here we have described a specific data analysis workflow, starting from our general data model, applying a text discretization function, a number of projections and finally a multilayer clustering method. Using different types of projections, or working directly with the original model, several types of text and structural analysis can be unified under the same framework. Examples include conversation retrieval [4], where a message ranking method is defined as a function of text, relationships between messages and between senders/recipients, and temporal information, and text message clustering [5] — as opposed to our working example where we cluster people using the information exchanged between them.

References

1. M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer Networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, sep 2014.
2. M. E. Dickison, M. Magnani, and L. Rossi, *Multilayer Social Networks*. Cambridge University Press, 2016.
3. M. Berlingerio, F. Pinelli, and F. Calabrese, "ABACUS: frequent pAttern mining-BAsed Community discovery in mUltidimensional networkS," *Data Mining and Knowledge Discovery*, vol. 27, no. 3, pp. 294–320, 2013.
4. M. Magnani, D. Montesi, and L. Rossi, "Conversation retrieval for microblogging sites," *Information Retrieval*, vol. 15, no. 3-4, pp. 354–372, feb 2012.
5. D. Combe, C. Largeron, E. od Egyed-Zsigmond, and M. Géry, "Combining relations and text in scientific network clustering," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, Turquie, 2012, pp. 1280–1285.

View publication stats