Text Network Analysis in Twitter:
Multilayer Networks of Co-Occurrent Hashtags, Mentioning Tweeple
and Topic Modeling Terms

Moses Boudourides
Robert K. Merton Visiting Research Fellow 2019
Institute for Analytical Sociology, Linköping University
Faculty at Northwestern University SPS Master's Program in Data Science Online
(Previously Professor of Mathematics at the University of Patras, Greece)

Nowadays, Twitter data can be easily mined or even be downloaded from publicly available datasets, which have been already retrieved from the Twitter and being distributed in existing open repositories of datasets. Typically, the network analysis of Twitter data proceeds towards the construction of (at least) two important networks: (1) graphs of hashtags co-occurring (in tweets) and (2) graphs of mentions or, better said, graphs among mentioning/mentioned tweeple, i.e., among mentioning-tweeple, who are senders of tweets, and mentioned-tweeple, who are simply mentioned in the contents of tweets sent by the former; notice that a mentioned-tweeple is not necessarily a sender-tweeple, although for the graph of mentions to be a nontrivial directed graph it is necessary that some of the mentioned-tweeple are themselves mentioning-tweeple. Furthermore, there exist various techniques of text analysis, which aim towards the extraction and classification of keywords from a corpus of textual data. Here, we are going to focus on a single one of these techniques, Topic Modeling (TM), and we are going to apply it in the corpus of the tweet contents of the analyzed Twitter dataset in order to be able to extract certain salient words from this corpus. Apparently, these salient words are called “TM-terms” or simply “terms,” since each one of them belongs to at least one of the “topics” that TM yields.

Thus, given a Twitter dataset, we are going to extract from it three kinds of keywords as three distinct layers of nodes: (a) the layer of hashtags, (b) the layer of tweeple and (c) the layer of TM-terms. Subsequently, we are going to construct the multilayer network of co-occurring three types of keywords inside the tweet contents. Notice that all edges among any pair of layers are simply symmetrical relations of co-occurrences of the corresponding keywords. Moreover, since any pair of keywords may co-occur inside multiple tweets, these edges (i.e., links connecting nodes between any two layers) are weighted according to the

number of tweets, inside of which a pair of keywords may co-occur. Similarly, edges inside the layer of hashtags and inside the layer of terms are also relations of co-occurrences weighted in the same way as before. The only layer of a different graph type is the layer of mentioning tweeple, constituting a multi-digraph, i.e., a weighted directed network.

After constructing the three-layer Twitter network of hashtags, tweeple and TM-terms, our aim is to examine how an existing partition in the subgraph of one layer may induce a corresponding grouping in the subgraphs of the remaining layers. There are three cases to be considered:

1. The community partition of hashtags in the layer of co-occurrent hashtags: Then each community of hashtags may induce two groupings: a grouping of tweeple consisting of those tweeple who co-occur with (at least) one hashtag of the given community (of hashtags) and a grouping of TM-terms each one of which similarly co-occurs with hashtags of that community. Therefore, the community partition of hashtags induces a grouping of the set of tweeple and a grouping of the set of terms. Apparently, these groupings do not constitute partitions of the sets of tweeple/terms (respectively), since they might be overlapping, due to the fact that the co-occurrence relation is not necessarily one-to-one (evidently, in general, it is one-to-many).
2. The decomposition of the layer of terms into topics (this is certainly not a partition because topics are overlapping): Again each topic may induce two groupings: a grouping of tweeple consisting of those tweeple who co-occur (inside tweet contents) with (at least) one term of the given topic and a grouping of hashtags each one of which similarly co-occurs with terms in that topic. Thus, the structure of topics (obtained by TM) induces a corresponding decomposition of tweeple and hashtags (respectively).
3. Similarly, the label propagation algorithm for community detection in the layer of mentioning tweeple may induce two corresponding decompositions on hashtags and terms (respectively).

In this seminar, I am going to present a framework of computations (implemented in Python) which uphold the above described construction of a 3-layer Twitter network and the generation of induced decompositions in each layer from existing partitions or decompositions in another layer.