Large-scale stylometry using network analysis

Maciej Eder

Pedagogical University in Kraków
maciejeder@gmail.com

Abstract

Stylometric methodology, developed to solve authorship problems, can easily be extended and generalized to assess different questions in the field of text analysis. Namely, the underlaying idea of tracing similarities between (anonymous) texts can be extended to map textual relations in large-scale approaches to literature. Explanatory multidimensional methods, relying on distance measures and supported with visualization techniques, are particularly attractive for this purpose. However, they are very sensitive to the number of features (usually: frequent words) analyzed. Even worse, they are either unable to fit dozens of texts on a single scatterplot (e.g. Multidimensional Scaling), or highly dependent on the choice of a linkage algorithm (e.g. Cluster Analysis). The technique introduced in this study combines the concept of network as a way to map large-scale literary similarities (Jockers 2013), the concept of consensus (Lancichinetti and Fortunato 2012), and the assumption that textual relations usually go beyond mere nearest neighborship.

Particular texts can be represented as nodes of a network, and their explicit relations as links between these nodes. The procedure of linking is twofold. One of

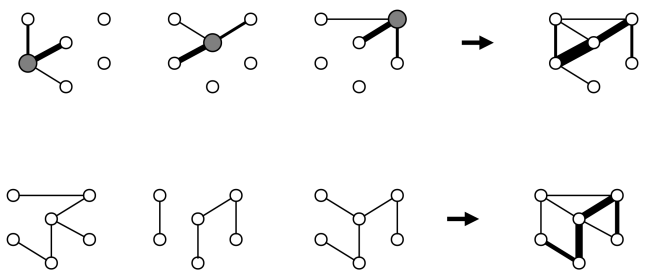

Figure 1: Two algorithms of mapping textual relations

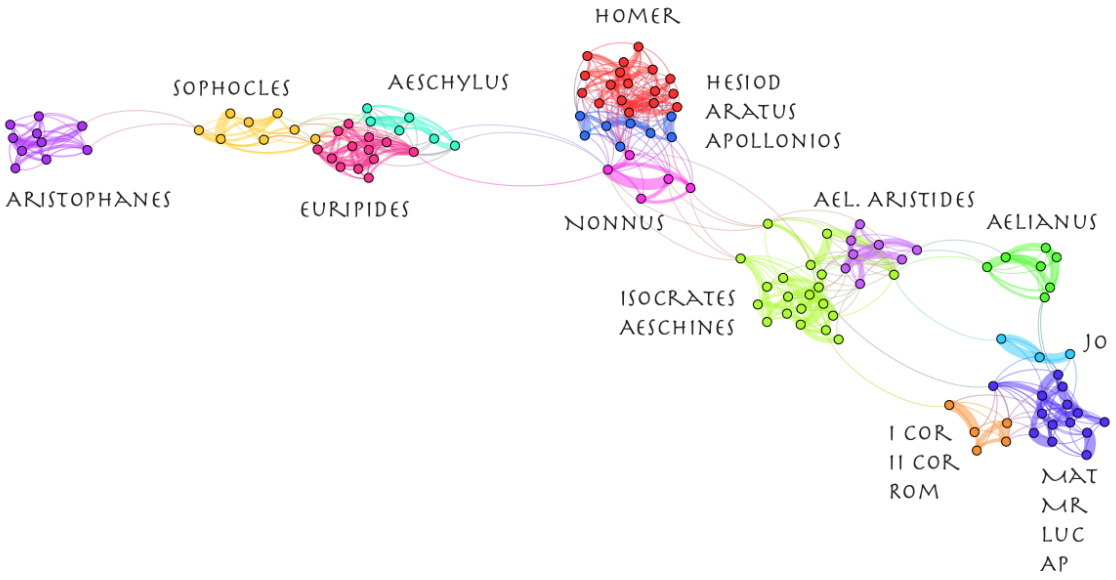

Figure 2: An example: 124 Ancient Greek texts (prose, poetry, drama) represented as nodes of a network

the involved algorithms (Fig. 1, top) computes the distances between analyzed texts, and establishes, for every single node, a strong connection to its nearest neighbor (i.e. the most similar text), and two weaker connections to the 1st and the 2nd runner-up (i.e. two texts that get ranked immediately after the nearest neighbor). The second algorithm (Fig. 1, bottom) performs a large number of tests for similarity with different number of features to be analyzed (e.g. 100, 200, 300, ..., 1,000 MFWs). Finally, all the connections produced in particular "snapshots" are added, resulting in a consensus network. Weights of these final connections tend to differ significantly: the strongest ones mean robust nearest neighbors, while weak links stand for secondary and/or accidental similarities. Validation of the results – or rather self-validation – is provided by the fact that consensus of many single approaches to the same corpus sanitizes robust textual similarities and filters out apparent clusterings.

The idea discussed in this paper can be applied to map large collection of texts, such as the corpus provided by the "Perseus Project" database (1127 texts), but also to represent similarities in smaller corpora. One of the examples include an ad-hoc collection of 124 Ancient Greek texts assessed with the aforementioned technique (Fig. 2). The nodes and edges of the network have been computed using the "stylo" package for R (Eder et al. 2013) and visualized with Gephi. The ForceAtlas2 layout (Bastian et al. 2009) has been used to establish the spacial relations between the nodes,

and the modularity detection algorithm (Blondel et al. 2008) to mark distinctive clusters with different colors. The obtained network reveals a clear genre separation (prose, epic poetry, drama), as well as chronological development of style.

References

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **10**, P1000.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska-Lincoln, pp. 487–89.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Lancichinetti, A. and Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, **2**, 336, 1–7.