Aaron Gerow
Computation Institute
University of Chicago
Chicago, IL USA
gerow@uchicago.edu

Bowen Lou
Computation Institute
University of Chicago
Chicago, IL USA
bowenlou@uchicago.edu

James A. Evans
Dept. of Sociology & Computation Institute
University of Chicago
Chicago, IL USA
jevans@uchicago.edu

Title

Networks in Natural Language: Named Entity Co-occurrences in Diachronic Corpora

Abstract

Systematic analysis of textual data often requires representational decisions that carry theoretical, computational and interpretive burdens and constrain research. Analyzing network-structures in language allows researchers to capitalize on robust mathematical foundations in statistical physics and computer science as well as techniques from social network analysis. Graph-based representations have been used in semantic and concept networks to model vocabulary growth, lexical retrieval, word associations, social and cultural fields. Perhaps most commonly, word co-occurrence networks over collections of documents are used in text clustering, document retrieval, information extraction, lexical semantic modeling, and analysis of inter-language variation. After briefly introducing common methods for construction and use of networks in natural language, we present an analysis of named entities (NEs) networks in a corpus of biomedical abstracts. Extracting NE co-occurrences in documents, we build an ordered series of networks representing relations among people, places and organizations at yearly intervals. Incorporating a degree of link-decay over time (analogous to fading relations), we explore the dynamics of the network's community structure. We find that while people make up a slight majority of nodes, person-to-person links are more than twice as likely as any other type. Results also suggest that over time, even with high rates of link decay, the network's community structure is increasingly modular and organizations are more (and increasingly) central to this structure than people. While our analysis highlights the utility of network-based representations, we hold that NE-graphs are only one of a number of lexical, semantic and discursive structures that can be mined and robustly analyzed in large collections of text.