N20 Diesner, J., & Pak, S. (2015). Evaluation and Contextualization of Networks Extracted from Text Data.

Relation extraction is a set of techniques that allows for identifying representations of social agents - among other types of entities - and their relationships from unstructured, natural language text data. This is particularly useful for identifying the structure of covert and hard-to-access networks, networks on which only archival records exist, e.g. bankrupts companies or groups that have ceased to exist, and networks that primarily interact through communication of which digitized accounts exist. One methodological key issue with relation extraction is the validation of the extracted relational data. The classic assessment question to ask here is: How do we know if the extracted network data represent the true network structure? Prior research provides partial answers to this question. Another take on validation is to ask: "Are the observed structures different from connections among these actors in other contexts"? A baseline strategy to answer this question is to develop an ERGM (Exponential Random Graph Model) to test if the extracted network differs in a significant and network theoretically meaningful way from a network of the same size and properties. We propose an alternative approach, which consists of keeping the set of identified social agents fixed and identifying their connections – including to other sets of agents - in other contexts or domains. With classic network data collection approaches, e.g. surveys, questionnaires and observations, this is hard to do as this procedure represents the difference between collecting one-mode versus multi-mode network data. However, text mining enables a more efficient solution to this problem since automated relation extraction techniques can be applied to multiple text corpora. We believe that this strategy has benefits beyond enabling the comparison and contextualization of networks extracted from text data as it allows for contrasting relationships between actors in one domain to their potential connections in other domains, e.g. for the case of interlocking directorates. There, one interesting question that typically doesn't get asked is: Through what other ties are a given set of board members connected in the real world as based on evidence from open source text data archives? Practically speaking, we solve this task by conducting syntactically disambiguated entity extraction on a set of text documents and identifying which social agents are connected through what type of interactions. We detect typed and categorized relationships; overcoming the arbitrariness of co-occurrence based networks. We then query public data, such as collections of news documents and legal documents, for new documents on the identifying agents in another content domain, applying the same network construction techniques, and comparing the resulting. We use the ConText toolkit for this work. We provide two case studies for the feasibility and results of this process: First, the networks of institutions associated with the causing and mandated prevention of the savings and loan crisis. Second, networks of social agents who take a position on the question of a relationship between economic inequality and climate issues.