

Thesis Defense

Institute for Software Research
Computation, Organizations and Society

Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts



Jana Diesner

Wed Feb 29, 1.30 pm, GHC
(Gates Hillman Center) 4405

Socio-technical networks are ubiquitous and impact society on many dimensions. Frequently, the functioning and dynamics of networks involve the flow and processing of information, which is often available as unstructured, natural language text data. This thesis is motivated by the need for scalable and reliable methods and technologies for constructing network data from text data.

When extracting socio-technical network data from texts, the relevant node classes might include categories that named entity extraction tools do not typically consider. I address this lack of technology by developing entity extractors that combine a network model from social science with supervised machine learning. However, for the resulting technology as well as alternative approaches to relation extraction, we do not know how the retrieved networks compare with respect to their structure and properties. I tackle this issue by contrasting network data constructed with common relation extraction methods from various corpora, and outline how these methods can be combined to capture different facets of networks.

One main limitation with constructing network data from texts is validation, which is hard when no ground truth data are available, e.g. for covert networks. I address this problem by identifying the impact of choices that engineers must make when building relation extraction tools and end-users must make when applying these tools on analysis results. I recommend strategies for mitigating the identified issues in practical applications.

When both, text data and network data, are available as a source of information, these data have been previously integrated by enhancing social networks with content nodes. I present and evaluate a methodological advancement to this approach, which allows for taking multiple types of behavioral data into account, namely interactions between social agents and language use.

The contributions made with this thesis help people to collect, manage, analyze and interpret rich network data. This is a precondition for asking substantive and graph-theoretical questions about networks and advancing network theories. I use a computationally rigorous and interdisciplinary approach for working towards this goal; thereby advancing the intersection of network analysis, natural language processing and machine learning.

Committee: Dr. Kathleen M. Carley (Chair), Dr. William W. Cohen,
Dr. Carolyn Penstein Rosé, Dr. Jeffrey Johnson (ECU)