Computational integration of text data and relational data to support actionable meaning in socio-technical networks

Thesis Proposal by Jana Diesner
Carnegie Mellon University (CMU), School of Computer Science (SCS), Institute for Software Research (ISR), PhD Program in Computation, Organizations and Society (COS)

Committee:
Kathleen M. Carley (chair, CMU, SCS, ISR)
William W. Cohen (CMU, SCS, Machine Learning Department and Language Technologies Institute)
Carolyn P. Rosé (CMU, SCS, Language Technologies and Human Computer Interaction Institute)
Jeffrey Johnson (East Carolina University, Institute for Interdisciplinary Coastal Science and Policy and Department of Sociology)

Abstract
Socio-technical networks are ubiquitous and impact society on many dimensions. As network participants become socialized into those networks, they alternately internalize network behavior or transform network behavior through their participation. Frequently the functioning of networks involves communication within the network or processing of communication and information originating from outside the network. Such communication and information data are often available as unstructured, natural language text data. Data describing the structure and behavior of networks can be collected from a variety of sources, including surveys, interviews, and text data. Often in prior work, text data are analyzed separately from relational data, or are reduced to the fact and frequency of the flow of information between nodes. The latter approach acknowledges the fact that information exchange has taken place, but does not consider the substance of the data. However, we know that without considering the content of text data, we are limited in our ability to understand the effects of language use in networks, including the transformative role that language can play on networks, and the interplay and co-evolution of information and network structure and behavior. Thus, we expect that in bringing together text data and relational data, we will be able to make substantial advances in network analysis. A complicating factor is that sometimes the structure and behavior of networks are encoded in the text data itself. In these cases, network data needs to be extracted from the texts.

For my thesis, I propose to develop, apply and evaluate a set of computational methods that facilitate the joint analysis of relational data and the content of text data. In working towards this goal, I use an interdisciplinary and computationally rigorous approach that combines theory and models from social science and socio-linguistics with methods from natural language processing and machine learning that are based on probabilistic graphical models. The datasets used for this work are the Enron email data, data about research funding from the European Union, and a dataset about the Sudan. The anticipated contributions with this thesis include:

- Provide and evaluate methods that will be integrated into the publicly available software products AutoMap and ORA.
- Clean and normalize public datasets that contain relational data and text data in order to ensure that each node represents one unique social entity and no entity is represented by more than one node.

By using publicly available data and software I ensure the repeatability of the findings. The overall goal with this thesis is to provide methods that support users in collecting rich network data that allow for meaningful and actionable analysis.
1 Introduction

Socio-technical networks are ubiquitous and impact society on many dimensions. Realizing these facts, public administrations, business corporations, funding agencies, and communities of practice, among others, have been facing a common challenge: How can network data be efficiently collected, managed and analyzed such that the relevant properties and behavior of networks and the underlying forces that drive the evolution and dynamics of networks are identified and understood, and respective decision making processes are supported? Often, the data about networks that are needed to address these questions are explicitly or implicitly contained in unstructured, natural language texts data. This dissertation focuses on methods for the efficient collection and meaningful analysis of network data that are derived from or enhanced by information from text data. The collection and storage of texts pertaining to networks has become fast, cheap, and easy. This development has generated a broad need among researchers and practitioners alike for theories, methods, metrics, and tools that support the automated knowledge discovery and reasoning about networks based on text data. The proposed work is motivated by the need for scalable and reliable methods for performing practically useful and relevant network analysis:

Relational data, also referred to as graphs, consist of vertices, also called nodes, and of edges, also called arcs, links or connections that connect the nodes. Additionally, nodes and edges can have weights, attributes, and types, and links furthermore have a direction. Nodes can represent instances of one or more types of entity classes, such as “agents” and “information”, and edges can represent instances of one or more types of relationships, such as “collaboration” or “communication” (Carley, 2002; Wasserman & Faust, 1994, p. 79). Social networks are networks involving only entities of the type agent. Based on the concept of socio-technical systems (Emery & Trist, 1960), the web of interactions within complex societal systems and their infrastructures is referred to as socio-technical networks. Most socio-technical networks exhibit characteristics of complex systems: they are in flux, vary in size, and feature a multitude of interactions and interdependencies between variables that can lead to radical changes in the system’s behavior (Kauffman, 1995). The concept of socio-technical networks includes virtual networks. Prior work stemming from graph theory and computing, among other disciplines, has led to the development of models of the evolution of graphs (Barabási & Albert, 1999; see for example Erdős & Rényi, 1959; Watts & Strogatz, 1998) and a wide range of efficient and scalable solutions for collecting, managing, and analyzing relational data (see for example Newman, Barabasi, & Watts, 2006).

Network data consists of relational data plus additional data that help to contextualize and interpret relational data (Alderson, 2008). Thus, relational data are an indispensable subset of network data, but are insufficient for revealing comprehensive stories about socio-technical networks (Corman, Kuhn, McPhee, & Dooley, 2002). The field of Social Network Analysis (SNA), which originated from the social sciences, especially from anthropology, sociology, and organizations science, has provided a plethora of theories, models and methods for working with network data (see for example Carrington, Scott, & Wasserman, 2005).

The concept of meaningful analysis is defined for the scope of this thesis by building on prior work as follows: in order to allow for meaningful analysis, linked data and data from the web needs to be
transformed into information, and information into knowledge (Parastatidis, Viegas, & Hey, 2009). Applying these steps to the networks domain requires the following two-step process, which is explained in the remainder of this chapter: first going from relational data to network data, and second from network data to knowledge.

1.1 Going from relational data to network data

Transforming relational data into network data requires the enhancement of relational data with additional data. This is typically achieved by bringing together various types or sources of information about a network (Alderson, 2008). This approach can be put into practice by using one or more of the following techniques:

- Include attributes that describe relevant characteristics of nodes and/or edges (Sampson, 1968).
- Consider different views of a network (Krackhardt, 1987).
- Enhance relational data with additional data that help to fix the context of relational data. Such additional data are often referred to as meta-data. Widely adopted types of additional data are temporal and spatial information (Eagle & Pentland, 2006; Snijders, 2001), such as timestamps of events or the geophysical position of entities and edges. Another type are information residing in un- or semi-structured, natural language text data that are authored by entities from within or outside the network (Carley & Palmquist, 1991; Danowski, 1993).

This thesis is confined to the last approach - the usage of text data for constructing and enriching graphs and networks. This approach is motivated by prior work that has shown that although text data and network data are often analyzed separately, for specific domains, datasets and questions, combining these two data sources can lead to a more comprehensive understanding of networks than exploiting only either one source (see for example Bourdieu, 1991; Carley & Palmquist, 1991; McCallum, Wang, & Mohanty, 2007b). The background chapter provides a more detailed discussion of this relationship. While natural language data are generated by humans and are therefore considered as a type of behavioral data herein, meta-data can be generated by people and by technical agents, e.g. in the case of key words for documents. This thesis addresses methods for utilizing human-generated text data pertaining to socio-technical networks, including meta-data. Summarizing the small body of research on bringing together text data and relational data, which is scattered in a rather disjoint fashion across various disciplines, it remains unclear:

- Under what conditions does the joint utilization of text data and relational data enhance the practical usefulness of network analysis? These conditions involve the research question, data, data analysis techniques, and methods for interpreting analysis results.
- What methods and metrics are available, needed and are best suited for jointly utilizing text data and relational data?

Answering the first question requires a substantial body of empirical work. The second question is approachable through interdisciplinary computational work that focuses on methodology. I chose to take the latter route in order to make the thesis a coherent fit within the wider framework of the COS (Computation, Organizations and Society) program. Since I am using three publically available datasets
from different domains (see the Data chapter of details and comparisons), some collateral yet unsystematic and incomplete answers to question one might also be produced.

1.2 Going from network data to insights and actions

Going from networks to knowledge means to perform analyses and further computations on network data to support end-users in answering their questions. This requires the usage of methods and metrics that are appropriate for reasoning about given network data. Sometimes, using generic matrix operations and network analysis measures suffices. In other cases, such as for the analysis of socio-technical networks, the consideration of specific network characteristics, e.g. the types of nodes and edges, is crucial (Cataldo, Herbsleb, & Carley, 2008; Krackhardt & Carley, 1998). In this step, the meaningful analysis of network data can be facilitated by computing outputs based on models and measures that are grounded in theories about the system that the network data represent. The outcome of this process can also be instrumental in building and testing theory (Corman, et al., 2002).

Both steps together help to make the meaning of network data actionable, i.e. extractable, explicitly representable, and usable. This thesis aims to contribute to the actionable meaning of network data by providing methods that can be applied for executed the outlined two-step process:

1. First, the exploitation of text data in order to support the collection of meaningful network data.
2. Second, leveraging social science theory in order to combine text data and network data.

The concept of actionable meaning is closely related to **semantic computing**, which refers to “computing with (machine processable) descriptions of content and intentions” (Parastatidis, et al., 2009). The difference between both concepts is that **actionable meaning** does not require the consideration of intensions, but the practical usability of network data.

Overall, the proposed work is driven by my search for a better understanding of the co-evolution and interplay of the semantics and mechanics of socio-technical networks. My goal with this research is to contribute to the integration of text data and relational data through a computationally rigorous and interdisciplinary approach; aiming to serve and advance the intersection of network analysis, natural language processing, and computing.

2 Background

The focus of this thesis is on the collection and analysis of network data – as opposed to relational data. Both steps are integral parts of the overall network analysis process, which is delineated below. The chapter continues with a review of prior approaches to bringing together text data and relational data. The review focuses on actionable semantics of the resulting data.

2.1 The network analysis process

**Social Network Analysis** (SNA), which refers to the “testing of theories about structured social relationships” (Wasserman & Faust, 1994, p. 17), and **Network Science**, which is defined as “the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena” (National_Research_Council, 2005, p. 28) have both evolved into widely accepted
methods among researchers and practitioners for studying networks. By network analysis I refer to both, SNA and Network Science. Originally, network analysis has mainly served social scientists as a method for retrospectively gaining a rich and thorough understanding of small groups (Mitchell, 1969; Newcomb, 1961; Ryan & Gross, 1943; Sampson, 1968). At the same time, researchers have also been using synthetic and empirical relational datasets to investigate the numerical properties of graphs (Erdős & Rényi, 1959; Simon, 1955). Since a plethora of disciplines started to adopt network analysis, e.g. linguistics, economics, and physics, the scope of methodological applicability network analysis has been extended into a general utility method, much like statistics. Impacted by the evolution from many disciplines, the methodology for performing network analysis is less standardized than research procedures in other disciplines. Synthesizing prior work suggests that the network analysis method comprises the steps shown in Figure 1. Since the individual steps are interdependent, the steps considered herein (indicated by colored frames in Figure 1) can cause recuperations on other steps.

Figure 1: The network analysis process. Green frames indicate the steps considered in this thesis.

1. Specification of a goal, question, or task.
2. Specification of relevant entities (nodes), relations (edges), and network boundaries.
3. Data collection if no data is given.
4. Representation of the relational data as a list, matrix, or graph.
5. Analysis and utilization of relational data. This may entail database operations such as search and retrieval, network analysis, network visualization, network simulation, and generation of input for machine learners, among other processes.
6. Validation of results. Error analysis if applicable.
7. Interpretation of results with respect to step 1. This can include suggesting intervening strategies and policies or formulating, extending or revising theory.

2.2 Collection of network data

Texts are one among many sources for information about networks: network data can be gathered through a variety of methods, most of which can be categorized as surveys (Krackhardt, 1987; Ryan & Gross, 1943), questionnaires (Newcomb, 1961), experiments (Milgram, 1967), observations (Mitchell, 1969), archival data (Burt & Lin, 1977), and simulations (Carley, 1991). Harvesting the web has also
become a widely adopted strategy for collecting relational data, especially about social networks (Parastatidis, et al., 2009). This thesis focuses on supporting actionable semantics of network data for situations in which text data pertaining to socio-technical networks are available. Text data often can be acquired as a by-product of the network data collection processes. Examples for text data that have been used for network analytical purposes include communication related to collaborative work processes (Carley & Palmquist, 1991; Cataldo & Herbsleb, 2008; Danowski & Edison-Swift, 1985), interpersonal communication (Fitzmaurice, 2000), interviews (Carley, 1988; Sageman, 2004), legal documents (Baker & Faulkner, 1993), news coverage (Van Atteveldt, 2008), web sites (Gloor, et al., 2009), and social media such as blogs (Adar & Adamic, 2005), chats (Paolillo, 1999), emails (McCallum, Wang, & Corrada-Emmanuel, 2007a), social networking sites (Downes, 2005), wikis (Chang, Boyd-Graber, & Blei, 2009a), and virtual worlds such as online games (Landwehr, Diesner, & Carley, 2009).

2.3 Text data pertaining to networks

Figure 2 illustrates the various options for the availability of data for any network analysis project. The figure shows which of these options are addressed in this dissertation, and which are not. This thesis considers the enhancement of relational data with information from texts (Figure 2, 2.2) and the extraction of relational data from texts (Figure 2, 3.2). Both steps link the data collection step to the data analysis step of the network analysis process (Figure 1).

2.3.1 Joint availability of texts and networks (Figure 2, case 2.2)

Sometimes, relational data representing the nodes and edges of a network and text data pertaining to the same network are both available (Figure 2, 2.2). I refer to relational data that denote entities and their relations as explicit relational data, and to text data that contain information about a network but are not available in relational form as implicit relational data. Prominent examples for the joint availability of text data and relational data include:

- Surveys that ask respondents not only for information about entities and their interactions (explicit network data, see for example Krackhardt, 1987), but also for answers to open ended questions that further describe the nature of nodes and links (implicit relational data).
- Co-citation networks, where person A is linked to person B if A cited B (explicit) in a paper (implicit), (Hummon & Doreian, 1989).
- Web science studies that combine data on the connectivity between URIs (explicit) with the content of pertinent webpage or blog entry (implicit) (Adar & Adamic, 2005; Kleinberg, 2003).

2.3.2 Availability of text data only (Figure 2, Case 3.2)

Sometimes, texts are the only source of information about a network (Figure 2, 3.2). Most of these cases fall into one or more of the following groups:

- Networks that are inaccessible or unobservable for the researcher:
  - Covert networks such as illegal business coalitions (Baker & Faulkner, 1993) or adversarial organization (Krebs, 2002; Sageman, 2004).
  - Networks that ceased to exist such as former regimes (Seibel & Raab, 2003) or bankrupt companies (Diesner, Frantz, & Carley, 2005).
Mental models of concepts that people conceive in their minds (Klimoski & Mohammed, 1994; Rouse & Morris, 1986).

- Very large networks for which the conduct of surveys within the network boundaries is prohibitively expensive (Burt & Lin, 1977), such as studies of affiliation of individuals with tribes or ethnicities in a country.
- Networks that lack an underlying real-world network or that are nothing more than the data traces that are produced by or within them, such as blogs (Adar & Adamic, 2005). We refer to such data as WYSIWII (What-You-See-Is-What-It-Is) (Diesner & Carley, 2009).

The next section develops the requirements for using texts and networks or texts only such that practically useful and meaningful network analyses are facilitated. Sections 2.5 and 2.6 take these requirements into account in order to define the work I propose.

Figure 2: Availability of data for network analysis

- General case: raw input data for any network analysis project
  - Case 1: relational data only
  - Case 2: relational data plus other data
  - Case 3: Non-relational data only
    - Case 2.1: relational data plus non-text data
    - Case 2.2: relational data plus text data
    - Case 3.1: Non-text data
    - Case 3.2: Text data

Task: Computational Integration
- Graph modification: Existence and/or properties of nodes and edges
- Additional information about texts

Task: Relation Extraction
- Transformation

Process: Making available for processing

Network data as input for meaningful further use such as: database operations, visualization, network analysis, simulation, machine learning
Sources for implicit relational data, for instance web pages, may also contain non-text data, such as images and audio and video material. Like texts, these data might also bear additional information about the network of interest (Figure 2, Case 2.1). While I am not considering implicit relational data that are not text data, the distinction between explicit and implicit relational data and the framework for jointly utilizing them suggested in this work might serve others as a starting point for jointly using non-relational, non-textual data together with explicit relational data.

2.4 Criteria for meaningful and practically useful analysis of word networks and socio-technical networks

A large number of approaches to capturing and exploiting the meaning of network of words and socio-technical networks exist. Many of these approaches have little in common beyond calling the object of interest a semantic network or a socio-technical. This thesis does not provide a comprehensive discussion of the semantics of texts or networks, but reviews prior work stemming mainly from (computational) linguistics, social science and network analysis that addresses practical usability of network data extracted from or enhanced by text data.

The meaning of socio-technical networks has not been the subject of extensive debates. Typically, the meaning of a network is equivalent to the knowledge that are gained by performing the steps of the network analysis process as outlined in Figure 1 (Mohr, 1998).

The meaning of relational representations of language and knowledge has been the subject of long-lasting debates among researchers from linguistics and artificial intelligence (Hirst, 2006; Ogden & Richards, 1923; Woods, 1975). We herein refer to relational representations of language and knowledge as word networks. When these networks represent meaning, they are referred to as semantic networks (Woods, 1975). The words in the underlying text data can occur in sequential form, such as in this document, or as disjoint tokens, such as lists of key words. The nodes in word networks are referred to as concepts. Concepts are abstract representations of whatever people conceive in their minds (Sowa, 1984). Edge formation is based on one or more of the following features of language: syntax (Gerner, Schrott, Francisco, & Weddle, 1994), semantics (Kamp, 1981), word distance (Danowski, 1993), and meaningful relations between concepts as perceived by the analyst (Glaser & Strauss, 1967; Novak & Gowin, 1984; Trigg & Weiser, 1986). A unifying assumption across the various approaches to semantic networks is the emergence of meaning from the context of words as given in the text data or the networks (see for example Carley & Palmquist, 1991; Collins & Quillian, 1969; Griffiths, Steyvers, & Tenenbaum, 2007; Minsky, 1974; Mohr, 1998; Shapiro, 1971; Weaver & Shannon, 1949). According to Hirst (2006), further progress in meaning extraction from texts will require a combined consideration of subjective authorial intent, subjective interpretations of the reader, and the extraction of objective representations of meaning from large-scale corpora.

Synthesizing prior work on the meaning of networks derived from text data suggests that there are two basic approaches: First, some word networks are assumed to be inherently meaningful. This is mainly achieved by applying grounded theory methodology to construct structural models (Glaser & Strauss,
1967), applying ontologies\(^1\) to generate word networks (Berners-Lee, Hendler, & Lassila, 2001), and using structured variables that are motivated by theory (Van Atteveldt, 2008). Second, meaning can be obtained by interpreting network analytical results (Bernard & Ryan, 1998; Mohr, 1998). This thesis does not consider these two approaches as being exclusive, but combines aspects from both: using theory to obtain or enhance network data that is subject to further computations in order to arrive at practically useful results.

Overall, there is no guarantee that network data or their analyses are meaningful. Moreover, it is easy to read patterns and meaning into networks, for example by making heuristic use of network visualizations (Bernard & Ryan, 1998). This limitation can be avoided by starting with a clearly defined research question, collecting appropriate data, and performing formal structural analysis – in short, following the network analysis process outlined in Figure 1. A second limitation applies: for practical purposes, achieving actionable semantics is not the only goal when using texts for network analysis. Prior work in this area suggests that the requirements for competitive computational solutions include three more objectives:

**Table 1: Objectives for competitive solutions to meaningful and practically useful word networks**

<table>
<thead>
<tr>
<th>Objective</th>
<th>Description</th>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actionable meaning</td>
<td>Going from relational data to network data to actionable knowledge.</td>
<td>Contributes to practically useful network analysis.</td>
</tr>
<tr>
<td>Automation</td>
<td>The ability to collect network data automatically as opposed to manually or in a computer supported fashion.</td>
<td>Automation contributes to scalability.</td>
</tr>
<tr>
<td>Abstraction</td>
<td>The ability to use words verbatim and to use concepts that may not occur verbatim in the text data.</td>
<td>Abstraction enables analyses on different levels of granularity and aggregation.</td>
</tr>
<tr>
<td>Generalization</td>
<td>The ability to identify new instances of relevant classes of nodes, edges and attributes as opposed to identifying a deterministic set of instances.</td>
<td>Generalization enables greater flexibility in applying a method to new data and domains.</td>
</tr>
</tbody>
</table>

Table 2 gives an overview on which of these goals are achieved by the different methods for generating word networks. Note that this table reviews methods for generating word networks, not for using them. Most of these methods were not developed to provide input for network analysis, but have been included for the sake of comprehensiveness.

**Table 2: Basic characteristics of methods for generating word networks**

<table>
<thead>
<tr>
<th>Type of relational representation of natural language text data</th>
<th>Automation No = manual Yes = automated</th>
<th>Abstraction No = verbatim Yes = concepts</th>
<th>Generalization No = deterministic Yes = find new instances</th>
<th>Where do semantics come from? (data means resulting relational data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mind maps (Buzan, 1974)</td>
<td>- No</td>
<td>Yes</td>
<td>No</td>
<td>- In process</td>
</tr>
<tr>
<td></td>
<td>- Computer supported</td>
<td></td>
<td></td>
<td>- In data</td>
</tr>
<tr>
<td>2. Concept maps (Novak &amp; Gowin, 1984)</td>
<td>- No</td>
<td>Yes</td>
<td>No</td>
<td>- In data</td>
</tr>
<tr>
<td></td>
<td>- Computer</td>
<td></td>
<td></td>
<td>- In process</td>
</tr>
</tbody>
</table>

---

\(^1\) Ontologies specify the set of possible elements and relations between elements in a given domain.
<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Generation:</th>
<th>Usage:</th>
<th>supported</th>
<th>supported</th>
<th>In data</th>
<th>Through network analysis</th>
<th>Through inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.</td>
<td>Qualitative text coding according to Grounded Theory (Glaser &amp; Strauss, 1967; Richards, 2002)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>In data</td>
<td>Through network analysis</td>
<td>Through inference</td>
</tr>
<tr>
<td>4.</td>
<td>Hypertext (Trigg &amp; Weiser, 1986)</td>
<td>Computer supported</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>In data</td>
<td>Through network analysis</td>
<td>Through inference</td>
</tr>
<tr>
<td>5.</td>
<td>Definitional semantic networks including text coding by using ontologies (Berners-Lee, et al., 2001; Fellbaum, 1998)</td>
<td>Generation: no</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>In data</td>
<td>Through network analysis</td>
<td>Through inference</td>
</tr>
<tr>
<td>6.</td>
<td>Knowledge representation in artificial intelligence and assertional semantic networks (Shapiro, 1971; Woods, 1975)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Through inference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Mental Modeling according to Spreading Activation (Collins &amp; Loftus, 1975; Collins &amp; Quillian, 1969)</td>
<td>Computer supported</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Through theory</td>
<td>Through network analysis</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Case Grammar and Frame Semantics (Fillmore, 1982; Fillmore, 1968)</td>
<td>Generation: no</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>In data</td>
<td>Through statistical analysis</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Frames (Minsky, 1974)</td>
<td>Computer supported</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>In data</td>
<td>Through network analysis</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Semantic Grammars (Franzosi, 1989; Roberts, 1997a)</td>
<td>Computer supported</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Through network analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Discourse Representation Theory (Kamp, 1981)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Through network analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Centering Resonance Analysis (Corman, et al., 2002)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Through theory</td>
<td>Through network analysis</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>Mental Modeling according to Map Analysis (Carley &amp; Palmquist, 1991)</td>
<td>Computer supported</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Through network analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>Event Coding in political science (King &amp; Lowe, 2003; Schrodt, Yilmaz, Gerner, &amp; Hermick, 2008)</td>
<td>Computer supported</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Through statistical analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>Semantic network in communication science (Danowski, 1993; Doerfel, 1998)</td>
<td>Computer supported</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Through network analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>Relation Extraction in NLP (Brin, 1999; Bunescu &amp; Mooney, 2007; Janas &amp; Schwid, 1979)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Through network analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>Probabilistic graphical models (Howard, 1989; Pearl, 1988)</td>
<td>Generation: no</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Through network analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This review suggests that approaches based on probabilistic graphical models (group 18) have the greatest potential for satisfying the requirements of actionable meaning, automation, abstraction and generalization for generation word networks – with one caveat: further computations on the relational data are needed to arrive at meaningful results. Once word networks are available, they can be
analyzed. Network analyses can be placed on a spectrum ranging between graph analysis and network analysis as shown in Table 3.

Table 3: Spectrum of network analysis methods

<table>
<thead>
<tr>
<th>Network analysis spectrum</th>
<th>Analysis of relational data</th>
<th>Analysis of network data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>- Find, formally describe and model evolution and properties of graphs through computational analyses - Advance theories evolution and properties of graphs</td>
<td>- Find, explain, interpret and model causes and implications of evolution and properties of networks - Advance understanding or theories about cognition and behavior of entities in network</td>
</tr>
<tr>
<td>Analysis procedure</td>
<td>- Focus on data analysis w.r.t to a research question</td>
<td>- All steps of the network analysis process (Figure 1)</td>
</tr>
<tr>
<td>Scalability</td>
<td>- Focus on large-scale systems</td>
<td>- Methods, metrics and tools typically scale up to analysis of small to moderately sized systems - Theories about networks of any size, include large-scale</td>
</tr>
<tr>
<td>Typical application domains</td>
<td>- Technical infrastructures such as telecommunication networks and the internet (Barabási &amp; Albert, 1999; Eagle &amp; Pentland, 2006) - Online social networks and social media (Adamic &amp; Huberman, 1999; Leskovec, Kleinberg, &amp; Faloutsos, 2007) - Other sizable socio-technical networks such as geopolitical entities (Auerbach, 1913; Bass, 1969; Newman, Strogatz, &amp; Watts, 2001; Simon, 1955)</td>
<td>- Social sciences and organizations science, foci on: - Analysis of small and medium-size social groups over time, e.g. people who live or work together (Milgram, 1967; Sampson, 1968) - Adoption and diffusion of innovation (Coleman, Katz, &amp; Menzel, 1966; Kraut, Rice, Cool, &amp; Fish, 1998) - Formal and informal communication (Monge &amp; Contractor, 2003) - Information processing and learning processes of individual and groups (Carley &amp; Palmquist, 1991)</td>
</tr>
</tbody>
</table>

Ultimately, users are interested in combining the advantages of both sides of this spectrum in order to gain a formally describable, generalizable, rich and meaningful understanding of networks of any size (Corman, et al., 2002; Hirst, 2006). This high-level goal can be broken down into a set of smaller goals as shown in Table 4. Since table 4 focuses on the features that are relevant for this thesis, it is not comprehensive in its specification of the objectives functions of network analysis methods: abstraction is only one among other features that account for the appropriateness of a network representation with respect to a research question. Determining the appropriateness of a network representation also requires decisions about the existence, weight, probability, type and attribute of nodes and links, and the directionality of links.

Table 4: Objectives for competitive solutions to meaningful and practically useful socio-technical networks

<table>
<thead>
<tr>
<th>Objective</th>
<th>Description</th>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actionable meaning</td>
<td>- Going from relational data to network data to actionable knowledge.</td>
<td>- Contributes to practically useful network analysis.</td>
</tr>
<tr>
<td>Automation</td>
<td>- The ability to collect and analyze network data automatically as</td>
<td>- Automation contributes to scalability.</td>
</tr>
</tbody>
</table>
Out of the given objectives, this thesis focuses on the appropriateness of network representations: the handling of natural language text data that pertain to networks (marked bold in Table 4). One frequently used solution in this case is to reduce text data to the existence, strength or probability of nodes and edges. For example, the fact that two people exchanged an email is represented as a link between these two agents; with the link weight representing the sum of the number of emails exchanged between these two people. The same strategy is being used for analysis of relational data on phone calls, microblogging, and text messaging services (see for example Eagle & Pentland, 2006). This efficient abstraction approach is adequate and sufficient if the content of the texts does not add to the understanding of the network (Alderson, 2008). Also, it preserves private information to some degree. Shannon and Weaver (1949, p. 8) capture this notion by stating that for the “transmission of a finite set of discrete symbols”, “information must not be confused with meaning”, thus “the semantic aspects of communication are irrelevant for to engineering aspects”. This approach has shown to be adequate for example for representing the flow of current through power grids, amounts money or products through trading partners, and data packets through routers. In order to formalize such transmission processes, computing has provided the concept of protocols, which specify the rules of connection and data transfer between senders and receivers. However, Alderson (2008) argues that even though knowing which entities exchanged what and how much of it might provide useful input to relational data analysis, additional information, such as the bandwidth and queuing capacities of nodes, might be essential for gaining even a basic understanding of networks. Moreover, it has been shown that considering the content of text data can enhance and alter the understanding of a network. Section 2.5 discusses prior work that supports this statement. In these cases, reducing text and communication to the existence and weight of nodes and edges is inappropriate. Shannon and Weaver (ibidem) note that measuring the successful, i.e. effective, conveyance of meaning is harder than determining the accuracy of symbol transmission.

Table 5 shows a mapping of the objective functions of network analysis methods (Table 4) to the main families of network analysis methods. This table is based on my review of prior work in network analysis. These families are non-exclusive, meaning that single studies can use features from multiple families. This table focuses on the collection and analysis of data, not the usage of the results.
Table 5: Basic characteristics of families of methods for collecting and analyzing graphs and networks

<table>
<thead>
<tr>
<th>Families of network analysis methodologies</th>
<th>Automation of data collection</th>
<th>Formalization of nodes and edges</th>
<th>Representation of natural language data: Reduction = abstraction of content to presence and weight of nodes or edges</th>
<th>Where does meaning come from?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Anthropological studies, qualitative studies, observations (Johnson, Boster, &amp; Palinkas, 2003; Mitchell, 1969)</td>
<td>No = manual supported</td>
<td>Yes and No</td>
<td>Appropriate</td>
<td>In network data</td>
</tr>
<tr>
<td>2. Classical SNA studies of small groups and organizations (Milgram, 1967; Newcomb, 1961; Sampson, 1968)</td>
<td>Originally: no</td>
<td>Yes</td>
<td>Appropriate to Reduction</td>
<td>In network data</td>
</tr>
<tr>
<td>4. Diffusion and adoption studies (Coleman, et al., 1966; Ryan &amp; Gross, 1943)</td>
<td>Originally: no</td>
<td>Yes</td>
<td>Appropriate to Reduction</td>
<td>In network data</td>
</tr>
<tr>
<td>6. Studies of social media and social networking sites (Barabási &amp; Albert, 1999; Eagle &amp; Pentland, 2006)</td>
<td>Yes</td>
<td>Yes and No</td>
<td>Reduction</td>
<td>In raw data, not in relational data in case reduction happened</td>
</tr>
</tbody>
</table>

Table 5 suggests that for all of the network analysis methods considered, computer supported and fully automated techniques for collecting and analyzing are available (for more information on network analysis software see Huisman & Van Duijn, 2005). Furthermore, most of these families use the small set of canonical node and edge classes for which most network analysis models, methods, metrics and tools have been developed. The vast majority of network analysis studies uses one node class (one-mode network), typically of the type agent or organization, and one edge class (uniplex network). The edge classes often represents a social relationship, such as acquaintance or friendship, a professional relationship, such as advice seeking or reporting to, or some type of communication exchange, such as sending a message to. Occasionally, studies consider two (two-mode network) or more (multi-mode network) node classes, and/ or more than one edge class (multiplex network). In the last years, the set of supported node classes has been extended to also include entities that represent instances of the
what (event), when (time), where (location), why (opinions, emotions) and how (resources) of events (Carley, 2002; Krackhardt & Carley, 1998). Similar trends have been occurring in NLP: research has moved beyond extracting names of people from texts to also distill other entities referred to by a name, mainly organizations and locations (Bikel, Schwartz, & Weischedel, 1999), and to distill emotions and opinions from texts (Wiebe, 2000).

Finally, Table 5 shows that regarding the handling of natural language text data, for four out of the six method families, the insight that could gained through network analysis might be lessened due to the inappropriate reduction of texts to existence, strength or probability of nodes and edges. This thesis aims to provide methods for representing the content of text data, e.g. the substance of communication data, in network data with the ultimate goal of contributing to the transformation of relational data to network data and meaningful data analysis.

2.5 Work proposed regarding the joint availability of texts and networks

When texts pertaining to networks are available, users have four choices:

1. Disregard text data for network analysis.
2. Analyze texts separately from relational data.
3. Abstract the presence and flow of text data through the network to the existence and weight of nodes and links if appropriate.
4. Consider the content of text data for network analysis.

Options 1 and 2 are unproblematic if the content of the text data does not contribute to the understanding of the network. Option 3 is a valid choice if the abstraction does not reduce relevant information that can be gained from network analysis. This raises two questions. These questions, answers to them, and solutions to how to fix open problems are presented in Table 6, along with a specification of which questions are further investigated in this thesis, and which are not.

Table 6: Research questions and selection of question for thesis

<table>
<thead>
<tr>
<th>Research Question</th>
<th>How can users figure out if the consequences of options 1 to 3 will apply?</th>
<th>If the consequences of options 1 to 3 apply, how can the content of text data be considered for network analysis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer</td>
<td>They cannot due to a lack of systematic studies that answer this issue. Previously developed theories and methods do not provide decision support mechanism for assessing whether including text data for a network analysis project will be useful or not prior to performing all steps of the network analysis process.</td>
<td>When the existence, weight and likelihood of nodes and links are insufficient proxies for content, the content of texts data needs to be represented in the network. In these cases, “we cannot reduce communication to message transmission” (Corman, et al., 2002).</td>
</tr>
<tr>
<td>Leads to the next research question</td>
<td>Formalize a procedure that assesses the usefulness of text data for network analysis beyond the applicability to a single dataset, domain, and research question.</td>
<td>How can text data and relational data be jointly utilized in a systematic and computational fashion such that the understanding of the socio-technical network of interest is enhanced in an actionable way beyond the exploitation of only text or relational data?</td>
</tr>
<tr>
<td>How to approach research question?</td>
<td>A solution to this task needs to emerge from a large body of empirical work.</td>
<td>Based on the review of methods for generating networks from words (Table 2), use methods based on probabilistic graphical models. Based on the review of prior work on combining texts and networks (Table 5 and 2.5.), use theory about the relationship between network position and language use to inform the generation of word networks.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Approaches to considering the content of texts build on the idea that “travelling through the network are fleets of social objects” (Danowski, 1993, p. 198), such as language, norms, practices, and other types of human behavior and interaction (Bourdieu, 1991; Eckert, 1998). Prior research has provided empirical evidence for the usefulness of considering language or text data when analyzing socio-technical networks: Milroy (1987) and Milroy (1985) show how different types of social relationships, namely kinship, friendship, working together, and being neighborhood, and the structural position of people in the local social networks resulting from these relationships, are indicative of certain roles. These roles relate to a person’s attitude towards changes in the way people speak; leading to impacts on sociolinguistic patterns of the adoption, diffusion and rejection of vernacular. Eckert (1998) shows how in groups that are formed for a certain purpose, linguistic style is continuously developed and shared by its members; leading to greater homogeneity of language use in communities of practice over time. Roth and Cointet (2009) found out that for different types of collaboration communities, different relationships between social capital, measured as degree centrality of authors, and semantic capital, operationalized as highly central documents, apply: for social networks among co-publishing scientists, social capital and semantic capital show a significant positive covariance. The researchers could not confirm the same trend for bloggers, where poor semantic capital does not translate into low social capital. Fitzmaurice (2000) uses historic data to investigate how strategies alliances between individuals, who may have opposing agendas but share a goal, impact their use of language. Guiffre (2001) revealed a positive self-reinforcing relationship between the stylistic perceptions of artists as expressed in reviews authored by art critics and the decisions made by gallery owners about concurrently exhibiting work by different artists; ultimately leading to more or less successful careers of artists. Ehrlich at al. (2007) provide an analysis of SmallBlue; an engine that searches the social network of IBM’s employees. They found that enhancing social network data with semantic information as derived from people’s blog entries, emails, chats, bookmarks and other social media sources can support the performance of expert finder systems, especially when searching for experts on very specific and narrowly defined problems. McCallum et al. (2007b) show how the clustering of individuals based on text data can outperform clustering based on explicit relational data.

In summary, empiric evidence and a few theoretical models exist for the relationship between language and socio-technical networks. Also, several techniques and tools for supporting parts of the process of combining text data and relational data have been developed and used. Combining texts and networks requires interdisciplinary work at the intersection of natural language processing, network analysis,
computing, and maybe other fields. This intersection still forms a small area of interest rather than an accepted domain where people have established standards for methods and evaluations. However, this area is gaining momentum. No coherent methodology for this integration process has yet emerged and been widely accepted. However, a common approach to combining text data and network data is to identify salient terms from texts, e.g. by computing the tf*idf metric per term, picking the terms with the highest score, and linking agent nodes to nodes representing the terms that the agents are affiliated with (Gloor & Zhao, 2006). While these data can readily serve as input to regular network analysis, the network data construction process does not consider theories about the relationship between networks positions and language use (Corman, et al., 2002). I propose to mitigate this limitation by developing a method that based on the review of methods for generating word networks (Table 2) and the review of methods for integrating texts and networks (Table 5, section 2.5) brings together theory suggested by Milroy and Milroy (1985; Milroy, 1987) about the relationship between network position and language with probabilistic graphical models in order to inform the generation of word networks. The ultimate goal with these methods is to contribute to the usage of the content of text data to qualify and interpret structural information in socio-technical networks and vice versa.

2.6 Work proposed regarding the extraction of relational data from texts

When network data is needed, but text data are the only source of information about a networks, the network structure can to be approximated, i.e. sufficiently accurately extracted from texts (McCallum, 2005). This task is referred to as Relation Extraction (REX). Several methods for performing REX have been developed in Natural Language Processing (NLP) and Computational Linguistics (CL). These methods exploit lexical and morphological (Woods, 1975), syntactic (Janas & Schwind, 1979), semantic (Fillmore, 1968), logical (Shapiro, 1971), and proximal (Danowski & Edison-Swift, 1985) information from texts. High performing and scalable methods typically combine multiple NLP and CL techniques and involve probabilistic and machine learning methods (McCallum, 2005; Van Atteveldt, 2008). Tremendous progress in the accuracy of REX has been achieved over the last years (see for example Brin, 1999; Bunescu, 2007; Etzioni, et al., 2004; McCallum, et al., 2007b; Zelenko, Aone, & Richardella, 2003). These advances are largely due to REX competitions that were initialized and funded by the US government, who provided benchmark datasets and developed rigorous REX evaluation metrics (Doddington, et al., 2004; Grisham & Sundheim, 1996). At a minimum, REX involves three steps which are typically performed in the following order:

1. Data preprocessing: includes subroutines such as chunking (partitioning texts into semantic units such as sentences) and reference resolution.

2. Node identification: this task has been studied in NLP and Information Extraction (IE) under the label of Named Entity Recognition (Bikel, et al., 1999), an also in political science (Schrodt, et al., 2008).

3. Edge identification: in academic research, this process often assumes that node identification has already happened (Chang, et al., 2009a).

The extracted relational data may represent the underlying network accurately or not. Accuracy is not a binary concept, but is typically measured as a percentage. In order to evaluate the accuracy of the retrieved relations, two methods are typically employed:
The Gold Standard test compares the distilled network data against some ground truth text data that has been previously annotated with entity and relationship by trained human experts. The manual or computer-supported generation of reliably labeled data is expensive: humans trained for this task can identify and annotate about five to ten relations or events per hour, and up to 40 relations a day (Schrodt, 2001; Schrodt, et al., 2008). Fortunately, various annotated datasets have been provided through nationally funded projects and were made publically available through institutions such as the Linguistic Data Consortium (LDC). However, the complexity of annotating data for REX has lead to compromises: most standard REX datasets denote relations only on sentence level, because the consistent identification, disambiguation and mark up of entities and relations on the document or even the corpus level might be cognitively too complex for humans to perform (Corman, et al., 2002). Also, the number of classes of entities and relations considered for REX is often fairly small – typically, REX applications are constrained to identifying people, places, and organizations, and the existence of relationships between those entity classes (multi-mode, uniplex networks).

Alternatively, the extracted relations can be validated by subject matter experts (SME) for their resemblance of the actual relations of interest (King & Lowe, 2003). However, for scalable real-world applications, network data resulting from REX are often too voluminous and too complex to be vetted for their accuracy by humans. To make things worse, in some cases, neither ground truth data nor SMEs are available.

Overall, comparative accuracy assessments based on data from different corpora, and even more so domains, are rare. In summary, what is needed are automated, scalable and accurate REX methods. One direction of valuable contributions in this area would be to incrementally improve an existing REX technique, or to develop a new one. Several groups are actively working on both of these tasks. I am not proposing to go this route for my thesis, but to tackle two closely related problems that tie together natural language processing and network analysis:

1. I propose to provide a method and respective implementation for the extraction of socio-technical network data from text data. The focus of the method will not be on accuracy improvement, but on generating network data adhere to the premise of “actionable meaning”. Based on my methods and literature review, I decided to put this goal into action by combining machine learning methods based on probabilistic graphical models (Table 2) with theory from social science (Carley, 2002; Krackhardt & Carley, 1998) that suggests a set of node and edge classes that are relevant for socio-technical networks and pertaining to which analytical measures including and beyond the generic set of graph analysis measures exists.

2. I propose to investigate the impact of REX on network analysis results.

The methods section describes how I plan to put these goals into practice. Task one is motivated by the fact that social and organizations science have developed inherently meaningful models of socio-technical networks that consider and require more classes of nodes and edges than those considered for the Named Entity Recognition in REX. Extending REX to find and classify instances of node classes that are theoretically grounded and for which validated analytical metrics are available can increase the practical usefulness and interpretability of network analysis results.
Task two is motivated by the following rational: The steps involved in REX (preprocessing, node identification, edge identification) are typically performed in the given order, but are not independent of each other. Decisions made in one step have shown to impact the results obtained from one or more of the subsequent steps (Bernard & Ryan, 1998; Carley, 1993; Roberts, 1997b; Roth & Yih, 2002). For example, if a node is assigned to the wrong node class in the entity identification stage, the relation extraction stage will inherit this error, which can lead to misclassified links, inaccurate network data, and ultimately to inaccurate results and interpretations. This error propagation is not visible for the end-user: while the relational data resulting from REX are unambiguous, the extraction process is not; meaning that using different subroutines can lead to different results. To make things worse, high performing REX techniques comprise a mixture of subroutines for each of the three REX steps, and these sub-routines also exhibit interaction effects. Consequently, not only the selection of different REX methods or algorithms (for a comparison see Table 2) can lead to the suggestion of different network structures, but the choices regarding more fine-grained computational subroutines can also induce differences. However, comparative REX studies typically report variances on the general methods or algorithm level; typically showing that some reasonable baseline as well as a previously presented method A are outperformed by a more recent method B by or a tweak to method A. I am starting to address the problem by arguing that the choices for and dependencies between sub-routines of REX can induce variances in the resulting network data that are independent of phenomena pertinent to the network, but are nothing more than technical artifacts. While efforts are put into improving the accuracy of REX methods, these effects and interdependencies are insufficiently investigated and understood. Schrod (2001) already pointed out this gap in research. He recommended improving network data analysis techniques rather than increasing REX accuracy, because accuracy gains might not cause significant differences in the final network structure.

To give an example, let’s assume a statistically significant accuracy increase in REX of two percent. This would be a substantial contribution from an NLP point of view. This increase might significantly alter the network structure and values of network analytical measures, maybe way more than two percent. If that was true, further improvements in REX from NLP could move scientific progress ahead more quickly than improvements on the network analytical side. If, however, the REX accuracy increase does not cause any significant change in network structure and properties, investing into improving the accuracy of REX might not be a central focus for network analysis. Which case is likely to apply we don’t know. To the best of my knowledge, the relationship between variances in REX accuracy and variances in network structure has not been addressed yet.

In summary, I suggest that it is crucial to know the exact amount of impact of each subroutine and subroutine interaction involved in REX on the resulting structure and properties of relational data extracted from texts. The outcome of this work will contribute to:

- The informed and reliable extraction of relational data from texts.
- Increase of control over multi-stage analytical processes.
- Drawing of reasonable conclusions and meaningful inferences from the resulting data.
- Even though most automated REX methods have been developed for a specific domain, corpus, and target function, many of them share a large portion of sub-routines for pre-processing and
node and edge identification. Precisely understanding the impact of widely used subroutines can move us closer to a higher comparability and generalizability of REX methods and tools.

Who cares about the outcome of this work? Rigorous investigations of the performance of text-to-network transformation engines beyond accuracy rates are important for understanding the degree to which a technique or tool is loaded with assumptions and decisions that impact the results: End-users of readily available REX packages need to be given the opportunity to know about the fact and strength of these effects. Engineers and scientists can take these effects into consideration when developing solutions for REX and integrating them with network analysis.

3 Data

This section describes the three datasets that were selected for the dissertation based on the data’s properties. These data were freely available from the internet when they were acquired. While much of the recent work on combining text analysis and network analysis investigates the properties and benefits of computer supported collaborative work and the usage of social media, the focus with this thesis is on networks involving conflict, tension and competition. Even though this thesis focuses on methods, knowledge about the properties of such networks might provide a valuable complement to the insights about networks of cooperation and collaboration.

3.1 Enron

The Enron email dataset was originally released online by the Federal Energy Regulatory Commission (FERC) in May 2002. FERC made the data available in order to allow the public to understand why FERC had started investigations into Enron. This dataset contains a large number of emails from individuals who were not involved in any of the actions that were subject of the Enron investigation.

Each email contains three sources of network data:

- Explicit relational data, i.e. the email addresses of the senders and receiver(s) as given in the email headers.
- Implicit data about the relationships between people as represented in the email bodies.
- Additional/meta data, such as a time stamp and a subject line.

FERC collected a total of 619,449 emails from 158 Enron employees, mainly from senior managers. The original version of the dataset had a variety of integrity problems. Next, Leslie Kaelbing from MIT purchased the data. The data was then acquired by researchers from SRI, notably Melinda Gervasio, who fixed many of the integrity problems and released their version of the dataset online. In March 2004, William Cohen from CMU put the data online for research purposes. Cohen’s version of the dataset contains 517,431 distinct emails from 151 unique users. These emails are organized in 150 user folders with a little less than 4700 subfolders; summing up to a size of about 400Mb. Some messages were deleted in response to requests from affected employees. Invalid email addresses for which a recipient was specified were converted to addresses of the form “user@enron.com”, and to

---

2 This section is based on prior descriptions of the Enron email data in (Diesner & Carley, 2005; Diesner, et al., 2005).
“no_address@enron.com” where no recipient was specified. Further consistency checks done by Andres Corrada-Emmanuel from the University of Massachusetts via applying the MD5 digest to email bodies revealed that the corpus actually contains 250,484 unique emails from 149 people. We started off with using the version of the Enron data that was provided by Jitesh Shetty and Jafar Adibi from ISI. The researchers from ISI had refined and normalized the dataset by dropping emails that were blank, duplicates of unique emails, contained only junk data, or were returned by the system due to transmission failures. The resulting corpus consists of 252,759 emails organized in 3000 user defined folders from distinct 151 people. The ISI group put the Enron data in a MySQL database which contains four tables; one for employees, messages, recipients and reference information. We chose this version of the corpus for our work, because the normalization done to it seemed appropriate to us and was well documented, and the structure met our needs.

One major problem with this dataset is that nodes represented email addresses, not people. We have started to correct for this issue for by using publically available meta-data in order to map e-mail address(es) to individuals. These meta-data contained information on the location of the Enron branch that people worked in as well as their job title. So far, we were able to map 1,234 email addresses to 557 distinct individuals, with the number of email addresses per person ranging between 1 and 17. The average for this number is 2.2, and the standard deviation is 1.9. The number of emails for which both, the sender and at least one receiver, can be mapped to a unique and disambiguated individual is 52,866. In total, we have mapped 24,825 emails (9.8%) and 797,569 instances of email addresses (39.5%) to actual people (Diesner, et al., 2005). However, this process is still incomplete and needs to be further continued as part of this dissertation.

3.2 Sudan dataset

Mappings of networks among socio-technical entities in the Sudan are confined to a small number of qualitative studies (Elageed, 2009; Lobban, 2008). Additional information on these networks might reside in the huge amount of publically available text data about Sudan, such as news wire data, research reports, legal documents, and web pages published by various authors and agencies with different agendas and validity. We have collected a corpus of about 40,000 articles that mainly include reports from SMEs about the Sudan plus about ten years worth of coverage from the “Sudan Tribune” newspaper, which is published in France. In the following I refer to this collection of unstructured, natural language text data as the Sudan corpus. Extracting and analyzing network data from this corpus is a deliverable for a multi university research initiative (MURI) that the CASOS Center is part of. The main goals with this MURI are to:

- Develop theories and computational techniques for modeling the adaptive behavior of groups in asymmetric threat environments.
- Identify and investigate various dimensions of socio-technical networks in Sudan with a focus on culture.
- Deliver tools for “rapid ethnographic assessment”.

Within this scope, I will contribute a computational solution for extracting socio-technical network data from the Sudan corpus. This specific task is representative for situations in which people need to distill
concise information about the relevant entities and their relationships of interest from text data, and where the definition of what is “relevant” and “of interest” varies depending on the research question, dataset, time and place. What is generally needed in such situations is the transformation of text data into concise reductions and abstractions of the original material, and utilizing network data are one strategy of working towards this goal.

The validation of the revealed data is often difficult to impossible. For this thesis, Dr. Richard Lobban, who is a professor of anthropology and African studies at Rhode Island College, a leading SME on Sudan, and a member of the aforementioned MURI, agreed to verify the relational data that I will extract from the Sudan corpus. The CASOS Center also has multiple different relational datasets about the Sudan which I will use as a point of comparison:

- Tribal affiliation network which manually compiled by Dr. Lobban.
- Hereditary-linguistic tribal affiliations network which manually compiled by Dr. Lobban
- UN tribal data with tribes to topics information

We do not expect the networks extracted from texts to resemble these networks, but are interested in the differences and commonalities.

### 3.3 Research Funding

Policies and legal regulations obligate federal funding agencies to publicize information about the allocation of resources to people and ideas. This development has contributed to the transparency of state-level decision making processes. As a byproduct, analysts are provided with large and rich sources of information about who collaborates with whom on what. I have collected the set of research proposals that the European Union (EU) has accepted for funding under the “Framework Programmes for Research and Technological Development”, short Framework Programmes. The dataset contains about 60,000 proposals and is publically available from the Community Research and Development Information Service (CORDIS). The Framework Programmes (FP) were started in 1984 by the Research Council of the EU with the goal of stimulating and enabling competitive research in the European Research Area (ERA). The FPs have been continued since then, with the 7th FP currently in place.

For each funded project, the CORDIS database provides the name, affiliation, and contact information for the project coordinator. Project coordinators are the equivalent to principal investigators in the US. The same information is provided for all additional collaborators on the projects. Further specified for each project are the start date and end date, costs and amount of funding awarded, completion status, and various key words and index terms. Finally, the data include three fields of unstructured, natural language text data, which contain the title, description, and additional information for each project. The length of text data per project varies greatly; ranging from concise summaries spanning a few dozen words to lengthy descriptions of the background, methodology and technical details. The completeness of entries in CORDIS also varies across and within FPs. Table 7 gives an overview on the size and completeness of projects per FP. Here, a “project” is defined as a CORDIS database entry for which at least a unique identification number is provided. Based on this definition, CORDIS contains 55,972 projects for FPs 1 through 6. A project is counted as a “project with text” if it includes an “objective” and “general information” that together are more than ninety characters long. I established this heuristic in
order to disregard text fields that only contain headers, such as “Research objectives and content:”, but no further content. A project is considered as a “project with person” if for at least one person specified, a non-empty and valid entry is available in the name field for either the coordinator or any collaborator. Valid entries disregard entries such as “NOT AVAILABLE”, “Address”, and “TBC”.

As already pointed out for the Enron dataset, one major challenge with this dataset is the consolidation of the various instances and spellings of people’s names into one consistent name per actual individuals. In order to identify the various ways in which a person is referred to, a data-driven set of rules and heuristics needs to be developed as part of this dissertation. I plan to use data from FP1 to FP6 only, because entries for FP7 are still being added such that a dataset for FP7 will be incomplete. Furthermore, if hypotheses and methods are generated, they can be further validated in the future with data from FP7.

Table 7: Quantitative information about the FPs

<table>
<thead>
<tr>
<th>FP</th>
<th>Time range</th>
<th>Number of projects</th>
<th>Ratio of projects with text</th>
<th>Ratio of projects with person</th>
<th>Ratio of projects with text and person</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1984 - 1987</td>
<td>3283</td>
<td>82.7</td>
<td>78.3</td>
<td>71.1</td>
</tr>
<tr>
<td>2</td>
<td>1987 - 1991</td>
<td>3884</td>
<td>79.9</td>
<td>63.0</td>
<td>57.9</td>
</tr>
<tr>
<td>3</td>
<td>1991 - 1994</td>
<td>5529</td>
<td>76.8</td>
<td>65.7</td>
<td>60.8</td>
</tr>
<tr>
<td>4</td>
<td>1994 - 1998</td>
<td>15061</td>
<td>79.9</td>
<td>93.1</td>
<td>74.5</td>
</tr>
<tr>
<td>5</td>
<td>1998 - 2002</td>
<td>17629</td>
<td>75.4</td>
<td>95.6</td>
<td>72.3</td>
</tr>
<tr>
<td>6</td>
<td>2002 - 2006</td>
<td>10255</td>
<td>96.9</td>
<td>90.4</td>
<td>87.4</td>
</tr>
</tbody>
</table>

3.4 Limitations of datasets

Proposals rejected by the EU for funding are not available from CORDIS or any other public source. Furthermore, it cannot be assessed based on public information how complete the CORDIS database is. Similar limitations apply to the Enron and Sudan data: For Enron, it is unclear how many emails are missing. For Sudan, it is hard to determine which texts contain reliable and accurate information about the system of interest. To further generalize this point, for none of the datasets that I chose to work with I can say based on public information how incomplete, error-prone or biased the data are. While these caveats clearly limit scientific investigations and conclusions, they represent common situations of data availability and quality: when studying large and hard to access systems, we have to work with the data that can be collected, not the data that we would wish to have. That being said, it is crucial to clearly state any limitations of the data and respective consequences for analyses and reasoning. Furthermore, it is essential to make the best effort possible to normalize and clean the data. For example, for Enron, email addresses need to be mapped to people, and for Funding, instances of unique people need to be associated with a unique key identifier per person.

3.5 Comparison of datasets

Table 8 compares the datasets along various dimensions. Even though the datasets stem from different content domains - namely industry, politics, and science - they share a few characteristic: all three datasets represent long-term, large scale collections of unstructured, natural language text data (implicit
network data). For all other dimensions listed in Table 8, two out of the three datasets share one characteristic, which I hope will contribute to the comparability of my results.

Prior research suggests that for all three datasets, the formation and cohesion of groups might be driven by external pressures such as scarce resources and struggle for power more so than by group-internal characteristics such as a shared identity and the desire to collaborate (Fitzmaurice, 2000). These properties have shown to foster the development of strategic alliances. For situations in which groups need to balance concealment and coordination, prior research has provided empirical evidence for network structures that differ from those in situations of open collaboration (Baker & Faulkner, 1993).

Table 8: Comparison of datasets

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Enron</th>
<th>Sudan</th>
<th>Funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Business: Innovation and Crime</td>
<td>Politics: Conflict and Crime</td>
<td>Science: Innovation, Collaboration, Competition</td>
</tr>
<tr>
<td>Subdomain</td>
<td>Crisis management</td>
<td>Crisis management</td>
<td>Diffusion of innovation, strategic alliances</td>
</tr>
<tr>
<td>Social network</td>
<td>Explicitly given in emails headers</td>
<td>Implicit in texts</td>
<td>Explicitly given in project descriptions</td>
</tr>
<tr>
<td>Semantic</td>
<td>Implicit in email bodies</td>
<td>Implicit in texts</td>
<td>Implicit in abstracts</td>
</tr>
<tr>
<td>information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>50,000 emails</td>
<td>40,000 texts</td>
<td>60,000 proposals</td>
</tr>
<tr>
<td>Time span</td>
<td>6 years</td>
<td>12 years</td>
<td>25 years</td>
</tr>
<tr>
<td>Original access to data</td>
<td>Internal</td>
<td>Public</td>
<td>Beginning: internal If funded: Public</td>
</tr>
<tr>
<td>Intended audience</td>
<td>Addressee(s)</td>
<td>The public Analysts</td>
<td>Program managers Scientific community</td>
</tr>
<tr>
<td>Style</td>
<td>From informal interpersonal to formal business</td>
<td>Journalistic Scientific, analytical</td>
<td>Scientific</td>
</tr>
</tbody>
</table>

### 4 Methods

This thesis comprises three projects. Table 9 summarizes which methods are used for which project and dataset. The following sections explain these projects and methods in detail.

Table 9: Overview on data by project

<table>
<thead>
<tr>
<th>Method used per project</th>
<th>Data</th>
<th>Project I: Relation Extraction</th>
<th>Project II: Impact of Relation Extraction on network structure and measures</th>
<th>Project III: Combine Network Analysis and Topic Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enron</td>
<td>Not needed</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Funding</td>
<td>Not needed</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
4.1 Relation Extraction: combining a model from social science with supervised machine learning to extract meaningful relational data

As discussed in the background chapter, machine learning methods that are based on probabilistic graphical models seem to be the most promising approach for automated and predictably accurate relation extraction, for locating and classifying previously unseen instances of node and edge classes, and for abstracting verbatim expressions to more general concepts. From an NLP/ML point, a common key question in this context is: What is the comparatively most accurate algorithm that serves this purpose, and how to construct it? Points of comparison are a simple baseline solution and the best-performing alternative algorithms. In contrast to that, from a network analysis point of view, one key question is: What nodes and edges need to be extracted in order to answer a research question, and how to do that? Answering this question is a mandatory step in the network analysis process (Figure 1), and has to be answered before the NLP/ML-centric question becomes applicable. This thesis aims to provide a computational solution that puts the network-oriented key question into practice. I plan to combine:

- the meta-network model, which is a theoretically grounded model originating from network science that defines the basic elements of socio-technical networks with
- existing probabilistic graphical models from ML for solving the actual node identification and classification task

such that the extracted relational data allow for subsequent reasoning and the computation of empirically validated metrics for which interpretations have been previously defined. The meta-network model (Carley, 2002) is an extension of the PECANS model (Krackhardt & Carley, 1998). I plan to use the meta-network model as an ontology that defines the types of nodes and edges that need to be extracted data. Furthermore, the model needs to be applied to the text data by using a method that does not require the a priori specification of instances of node and edge classes or rules for identifying these instances. In the Information Extraction field, a similar version of this task has been previously approached by using machine learning methods to:

- Identify and classify instances of people, organizations and locations; a process also referred to as Named Entity Recognition (NER) (Bikel, et al., 1999).
- Identify the relations between the named entities (Miller, Fox, Ramshaw, & Weischedel, 2000; Zelenko, et al., 2003). Solutions to this step sometimes assume that the named entities are already given. I am not making this assumption.

Both steps together are referred to as REX. Identifying instances of and relations between agents, groups, and locations in text data does not always suffice for representing socio-technical networks and studying their structure and behavior: the PECANS model suggests that the entity classes resources and tasks are also relevant. The PECANS framework also provides a set of matrix operations for combining networks of different types and a corresponding set of network analytical metrics that are tailored to the considered classes and networks. For example, examining the congruence between technical interdependencies between tasks and resources on one side, and the collaboration between people who execute these tasks and use these resources on the other side requires the identification of
instances of people, tasks and resources (Cataldo, et al., 2008). The PECANS framework has been previously validated and practically used for a variety of domains and datasets. The meta-network model (Carley, 2002) extends the PECANS model by considering additional node classes, namely knowledge and time, and additional metrics that can be computed on networks that use these node classes. These measures have also been previously validated (Carley, Reminga, Storrick, & DeReno, 2009) and serve the ultimate goal of supporting richer interpretations than the standard suite of “content-free” network analysis measures does. The majority of these classic measures, such as density, different types of centrality, and diameter, were designed for either social networks (agent to agent) or generic graphs (any type of nodes and edges). Zooming out from the specific labels of node classes to a model of wider and more flexible applicability, the meta-network model and respective measures keep being developed and adjusted to accommodate network data that represent the who (agent, organizations), what (task, event), when (time), where (location), why (emotions, beliefs) and how (resources, knowledge) of events. If such data need to be collected from texts, the following tasks apply:

1. Support end-users in extracting instances of node that represent the who (agent, organizations), what (task, event), when (time), where (location), why (emotions) and how (resources, knowledge) of actual or fictional scenarios and that may or may not be referred to by a name. The resulting data can be analyzed by using methods and tools that are capable of reasoning about these classes, e.g. the meta-network related metrics in the ORA tool (Carley, et al., 2009).

2. Provide developers with a modifiable and extensible solution for building models that are based on partially or completely different ontologies. For example, if somebody wants to extract instances of plants and animals from ecological literature in order to build a food network, the learner needs to accept training data annotated with the customized set of categories, and based on that generate a model that can be applied to new data with predictable accuracy.

I propose to deliver a computational solution that satisfies both requirements by leveraging the Conditional Random Fields (CRF) approach (Lafferty, McCallum, & Pereira, 2001). CRF are a discriminative supervised learning technique that is capable of finding global optima in sequential data with respect to a target function. CRF have shown to be advantageous over alternative algorithms from the same family of ML techniques, such as Maximum Entropy Markov Models, because they do not suffer from the labeling bias effect, which can severely hurt relation extraction accuracy (Dietterich, 2002). While training a CRF has a high time complexity due to performing global search with a reasonably sized gradient, applying the model is fast and scales to very large text sets. The resulting technology tuned for the meta-network system will be made publically available in the AutoMap software (Carley, Columbus, DeReno, & Diesner, 2008). The resulting learner will be built such that it can be re-used to train models based on modified or new ontologies. All technologies will be made available as part of the AutoMap and ORA executables.

4.1.1 Training and testing data

Supervised learning requires labeled training data. While there is no tagged data available that uses the meta-network categories, a subset of the categories used in the BBN Pronoun Coreference and Entity Type Corpus (BBN in the following) (Weischedel & Brunstein, 2005) can be mapped to the meta-network categories. The resulting model will be applied to the Sudan data in order to extract relational data that
represents socio-technical networks in the Sudan. Both, BBN and the Sudan corpus, consist of news articles, but from different regions.

4.1.2 Evaluation

Two evaluation methods will be used: First, I will build and evaluate a model by using CRF and performing tenfold cross validation on BBN. The accuracy of node identification and classification will be determined by using the Gold Standard test, where the model is tested on previously unseen yet presumably correctly labeled data. I will report how accurately the resulting system performs in comparison to other systems built on the same data that use the original BBN categories.

Second, I will use the model to extract relational data from the Sudan corpus. For this step, the nodes also need to be linked. Different node linkage techniques will be used as described in the next section (chapter 4.2). These data will be evaluated by a group of subject matter experts on the Sudan (for details see the Sudan data section in chapter 3). Their evaluation will be based on:

- The underlying text data.
- Expert knowledge about the Sudan regardless of the underlying text data.

The outcome of the tenfold cross validation and the feedback from the SME’s will then be used to perform error analysis and can be used to improve the learner.

4.1.3 Contributions

The contributions that I hope to make with this project are:

- Integration of a model from network analysis with methods from ML and NLP to facilitate the efficient and predictable accurate extraction of network data from large corpora.
- Support end-users in exploiting the retrieved network data by computing metrics that consider the substance of the node types.

The difference between my solution and prior NER solutions are:

- The consideration of entities that may or may not be referred to by a name.
- The consideration of vague and fuzzy entity classes, such as knowledge and resources.

Both types of information are important in the context of socio-technical networks.

4.2 Impact of relation extraction on network data

The computational steps involved in REX impact not only the extraction accuracy, but also the structure and properties of network data. While accuracy is the classic evaluation criterion for REX – at least from a NLP/ML point of view, the impact of REX on network data and respective analysis results is insufficiently understood. If significant increases in extraction accuracy do not cause significant changes in network structure, focusing on further boosting accuracy might not be the most promising direction for advancing REX. If however minor changes in extraction accuracy result in significant changes in network structure, increasing accuracy further has significant effects beyond the REX task. I plan to test the sensitivity of network structure and measures computed on these structures to variations in REX methodology. I will conduct a series of controlled experiments to determine the delta between using
different REX routines and sub-routines. Table 10 details the routines considered for these experiments. Ultimately, one would wish to have a comprehensive knowledgebase of methods-induced biases that everybody can draw from when using or developing REX. I am not promising to deliver such a knowledgebase, but I will get work started in this direction by determining how much of a difference the usage of a selected subset of very widely used REX sub-routines makes for subsequent network analyses. In order to achieve this goal, I will solve the following three tasks:

1. Identify a set of computational subroutines that are an integral part to a variety of REX method.
2. Quantifying the strength and boundaries of the variation in network structure and network-analytical results that are due to these subroutines by performing controlled experiments.
3. Communicate the identified effects to end-users in an understandable fashion.

I will contrast REX results generated by using substantially different REX methods as opposed to methods originating from the same family of approaches, such as supervised machine learning only. I chose this wide scope to enable a broad understanding of the robustness of REX techniques. Identification of sub-routines has not yet been complete, but started: state of the art REX methods run reference resolution as part of the pre-processing step. Reference resolution involves two sub-routines: Anaphora resolution aims to find an antecedent that a pronoun most likely refers to (Sidner, 1979). Coreference resolution identifies the various instances of an entity and associates them with each other (Hobbs, 1979). Both, node and edge identification, can be impacted by reference resolution. I will solve tasks two and three for both subtypes of reference resolution.

Table 10: Entity and link extraction techniques considered for network robustness experiments

<table>
<thead>
<tr>
<th>Technique</th>
<th>Impact on node extraction</th>
<th>Impact on edge extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reference resolution: anaphora resolution (Sidner, 1979) and</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>coreference resolution (Hobbs, 1979).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Thesauri, positive filters, negative filters, semantic grammars. All</td>
<td></td>
<td>Not applicable</td>
</tr>
<tr>
<td>members of family of deterministic approaches that are widely used in</td>
<td></td>
<td></td>
</tr>
<tr>
<td>political science (Schrodt, et al., 2008), social sciences (Carley, 1993),</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and (computational) linguistics (Roberts, 1997a).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Parts of speech tagging and shallow parsing. Used as stand-alone</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>technique (Corman, et al., 2002; Schrodt, et al., 2008) or integral part</td>
<td></td>
<td></td>
</tr>
<tr>
<td>of more complex NLP systems.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Windowing, a proximity based technique originating from</td>
<td>Not applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>communication science (Danowski, 1993).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Supervised machine learning methods,: Conditional Random Fields</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>(Lafferty, et al., 2001).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2.1 Data

For the experiments, I plan to use the text data portion of all three datasets. This procedure supports the comparison of methods-induced differences across different domains, sources and writing styles (see Table 8 for comparison of datasets).
4.2.2 Evaluation

The following comparisons of the resulting data will be performed:

1. Against ground truth where available, i.e. BBN.
2. Within and across each dataset, determine the difference between:
   2.1. For pre-processing techniques: using a routine versus not using it
   2.2. For detecting nodes and edges: using a technique versus using each alternative technique.

Table 11 shows which metrics and dimensions will be used for the comparisons.

Table 11: Dimensions of comparison of relational data

<table>
<thead>
<tr>
<th>Network level</th>
<th>Quantitative measures</th>
<th>Qualitative measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>- Number</td>
<td>- Identity of N highest scoring nodes</td>
</tr>
<tr>
<td></td>
<td>- Centrality (degree, betweenness, closeness, eigenvector)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Clustering coefficient</td>
<td></td>
</tr>
<tr>
<td>Graphs</td>
<td>- Number of edges</td>
<td>- Identity of N highest scoring edges</td>
</tr>
<tr>
<td></td>
<td>- Density</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Connectivity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Centralization (degree, betweenness, closeness, eigenvector)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Average path length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Number of triads</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Reciprocity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Number of components</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Fragmentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Upper Boundedness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Hamming distance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Quadratic assignment procedure (QAP)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- For Sudan data: tribal affiliation and hereditary-linguistic tribal affiliations network built by SMEs</td>
<td></td>
</tr>
<tr>
<td>Meta-networks, can be extracted by using methods 2. (predefined preprocessing material) and 5. (machine learning) in table 10</td>
<td>- Meta-network metrics available in ORA. This step builds upon the techniques developed in the previous project (4.1., relation extraction by using meta-network) and provides an additional validation of these techniques.</td>
<td>- Identity of N highest scoring nodes and edges</td>
</tr>
</tbody>
</table>

4.2.3 Contributions

I plan to contribute empirical quantitative and qualitative comparisons of the impact of computational solutions for relation extraction on network analysis data and results. This contribution will advance the understanding of the impact of methodological choices on network data beyond accuracy rates.
4.3 Computational integration of network-centric theory of language change and text analysis

Previous research on the relationship between the dynamics of language and social networks suggest an impact of the general position of individuals in a network on their motivation or ability to bring about innovations and change with respect to the current state of affairs in their environment (Bourdieu, 1991; Milroy & Milroy, 1985; Milroy, 1987): people are likely to introduce and spread innovations if they maintain a plethora of weak ties, are marginal to any adopting group, and don’t consider the element of change as a significant network marker (for the notion of strong and weak ties see Granovetter, 1973). People who are situated at the core of cliques and hubs can afford and tend to resist impacts that deviate from the group’s norms and that originate from outside the network. I will leverage the theory about the diffusion of language change originating from socio-linguistics to identify the set of individuals whose text data will be analyzed.

For text analysis, I will use topic modeling - an unsupervised machine learning technique that reduces the dimensionality of text data to an unlabeled set topics (Griffiths, et al., 2007). Each topic comprised a set of words where the weight per word indicates the strength or likelihood of the association of a word with the topic. The assignment of words to topics is non-exhaustive and non-exclusive, meaning that not all texts terms are descriptive for topics while certain terms or phrases may be indicators for multiple topics. Topic modeling has several desirable properties:

- Cost efficient: since the learning is unsupervised, no labeled ground truth necessary.
- Scalability: Scale up to very large corpora.
- Word sense disambiguation: capable to identify different meanings of a word by considering the word’s context.

Taking topic modeling to the structural level, Chang and his colleagues have used the Latent Dirichlet Allocation (LDA) technique (Blei, Ng, & Jordan, 2003) to suggest link labels for untyped links in semantic networks (2009a). Mimno and McCallum (2008) suggest that while in the vanilla version of LDA, any observed and descriptive features of the text data are generated based on an assumed latent probabilistic graphical model, conditioning topics on the observed data instead of generating the data might be more efficient. Based on this rationale, they develop the Dirichlet-Multinominal Regression (DMR) technique as an extension to LDA. The key idea with DMR is the assumption and computation of distributions per topic not only over words, but also over metadata that provide additional information about documents. Thus, DMR eases the consideration of various types of metadata on the text data, such as the date or publication venue of a text document. I propose to use this technique in order to condition words on network roles. More specifically, I aim to not to learn a topical profile per individual and/or document as done in prior work (McCallum, et al., 2007b), but to identify topics per groups of individuals who occupy theoretically grounded roles in socio-technical networks. I will operationalize the roles of change agents and preservation agents by using network analytical measures that are indicative of these roles. There is no canonical set of metrics and value ranges per metric that describes these roles. Based on my literature review (Carley, et al., 2009; Wasserman & Faust, 1994), the measures shown in Table 12 will be considered. Indicate value ranges for these measures are not defined in the literature, and therefore will be identified by using a data-driven approach.
Table 12: Operationalization of roles based on network analysis

<table>
<thead>
<tr>
<th>Measure</th>
<th>Brief Definition</th>
<th>Change agent</th>
<th>Preservation agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree centrality</td>
<td>Sum of direct links per node</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Betweenness centrality</td>
<td>Number of shortest paths between any pairs of nodes passing through node i</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Closeness centrality</td>
<td>Average distance of a node from all other nodes</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Boundary spanner</td>
<td>high on betweenness and low on degree</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Cut point</td>
<td>removal of node results in increase in number of components</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Triads</td>
<td>number of triads centered at each node</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>density of subgraph induced by a node's ego network</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Clique Count</td>
<td>number of distinct, maximally complete subgraphs of size three or more to which a node belongs</td>
<td>Low</td>
<td>Low to high</td>
</tr>
</tbody>
</table>

The identified topics will be used to enhance the graph: topics will be added as nodes such that a two-mode, agent-to-knowledge network is created. ORA provides metrics that can be computed on this type of network. The roles will be added as agent attributes such that they can be considered for further analysis. These analyses will not be part of the thesis. Links will be labeled with the salient topics identified for the set of authors per documents.

The following methodological extension to topic modeling on texts generated by people in socio-technical networks seem relevant, but will not be pursued in this thesis: First, topic modeling can be performed on the entire corpora to identify people with similar topical profiles. It can then be checked if an explicit relationship is denoted for these people or not. This technique might help to identify missing links, redundant links, and competitors. Second, when topic modeling is performed on the whole corpus, the identified topics can be labeled, which is typically done by using the term most strongly associated with a topic. The topics that the authors of the documents actually address can then be compared against predefined keywords for these documents. This comparison can serve the evaluation of the alignment or mismatch between top-down definitions of topics, such as the categories given for research funding proposals or email subject lines, against the categories that emerges from the actual data.

4.3.1 Data

For this project, I will use the Enron data and the Funding data.

4.3.2 Evaluation

While the methods for performing social network analysis and topic modeling involve rigid procedures and explicit computations, the interpretation of topic modeling results is a non-standardized process that leaves room for reading meaning into the outputs (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009b). Even though we have some expertise in a few research domains and also some experience in acquiring funding for research, we are not qualified to evaluate topic models regarding grants awarded
since 1984 across a wide range of research domains. Therefore, I will interpret the results to illustrate the usage of the method rather than to formulate empirical conclusions. A domain-specific analysis of the results would require the judgment of subject matter experts. The outputs will be evaluated by:

- Comparing the topic models per role against each other.
- Comparing the topic models per role against topic models derived from randomly selected people who do not assume any of the given roles.

4.3.3 Contributions

As a result form this project I will deliver a methodology that can be used to investigate the following question: What topics are addressed by the set of individuals who assume the network-based roles of change agents and preservation agents, and how do these topics differ? Text data and relational data are often either analyzed in a disjoint fashion, or are reduced to the fact and frequency of the transmission of objects or data through a network. This common approach limits the insights that can be gained from relational data analyses if the text data and the network data complement each other such the joint consideration of both sources leads to a more comprehensive view of the underlying system. The goal with this project is to provide a methodology that aims to address this limitation and that can be applied to others domains and datasets different from the ones used in this study.

5 Contributions and limitations

This proposal outlines the motivation, background, data, and planned work for the development, analysis and evaluation of computational methods that facilitate the joint consideration of text data and relational data. These methods are intended to support users in collecting rich network data that allow for meaningful and actionable analysis. In working towards this goal I use an interdisciplinary and computationally rigorous approach that brings together theory and models from socio-linguistics and social sciences with methods from machine learning and natural language processing. Table 13 provides an overview on the anticipated contributions and limitations. I will make the scope and limits of the applicability and generalizability of my findings explicit in the thesis. Overall, the proposed methods for integrating text data and relational data are small steps towards the greater goal of jointly utilizing both types of data as they represent different dimensions of behavior in socio-technical networks. The outcome of this work could serve as building blocks for more comprehensive frameworks and theories that might emerge in the future.

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Cleaning and normalization of public datasets, i.e. the Enron dataset and research funding dataset. This step is needed to ensure that each node represents one unique social entity (disambiguation) and no unique entity is represented by more than one node (reference resolution). I use public datasets for which I do not know how incomplete, error-prone and biased they are. These caveats limit scientific investigations and conclusions, but represent a common case for the availability of text and network data. Using other datasets that differ in genre, content domain and other dimensions might result in different findings. This assumption</td>
<td></td>
</tr>
</tbody>
</table>
### Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
</table>
| Relation Extraction: method and implementation for extracting network data from texts where:  
- nodes are identified and classified according to the meta-network model  
- edges are identified based on node proximity as baseline.  
These routines will be made available in AutoMap, which is available at no charge for research purposes. The resulting network data will be output in DyNetML, which is an open standard that can be used as input to ORA, which is also available at no charge for. | The computational work is mainly based on probabilistic machine learning techniques. Choosing different families of methods, such as rule-based techniques, can lead to different results. |

### Evaluation

Evaluation and comparison of relation extraction techniques and subroutines with respect to network structure and network analytical measures.

### Methodology for:

- Combining network-centric theory about language change and topic modeling.  
- Modifying social network based on identified topics. This methodological step will be made available in AutoMap, and the resulting network will be output in DyNetML.

The validation of the outcome of topic modeling is hard since there are no ground truth and subject matter experts on the studied domains available.

### 6 Thesis Outline

1. Introduction
2. Background  
   2.1. The network analysis process  
   2.2. Collection of network data  
   2.3. Text data pertaining to networks
3. Data  
   3.1. Enron data  
   3.2. Sudan data  
   3.3. Research funding data  
   3.4. Comparison and limitations of datasets
4. Relation Extraction: Computational integration of meta-network model and supervised machine learning to extract network data from texts  
   4.1. Method  
   4.2. Results  
   4.3. Evaluation  
   4.4. Contributions
5. Impact of relation extraction on network data  
   5.1. Experimental design  
   5.2. Results  
   5.3. Evaluation
5.4. Contributions

6. Computational integration of network-centric theory of language change and unsupervised machine learning
   6.1. Method
   6.2. Results
   6.3. Evaluation
   6.4. Contributions

7. Contributions

8. Limitations and future work

9. References
7 Thesis Timeline

Data collection
Enron, Sudan, Funding data set (2 months)
  • Scrape data from public sources (done)
  • Manage in RDBMS (done)
  • Cleaning and normalization (2 more weeks for Enron and Funding)

Project: Relation Extraction (2 months)
  • Implement reusable learner to identify nodes based on meta-network model (50% done)
  • Connect model to node linkage routine (proximity based as baseline)
  • Evaluate prediction accuracy
  • Error analysis

Project: Impact of relation extraction on network data (3 months)
  • Empirical identification of impact of relation extraction method and involved sub-routines on relational data and respective analytical measures
  • Network comparisons

Project: Integration of network-centric theory of language change and topic modeling (4 months)
  • Operationalization of roles and respective network analysis and robustness analysis
  • Link role identification to topic modeling
  • Use topics to enhance graph (nodes, edges, and attributes of nodes and edges)

Write-up and synthesis of findings (1 month)
  • Defense in Spring 2011
Acknowledgements

This work was supported by the following grants: National Science Foundation (NSF) IGERT 9972762, ONR ROE N00014-08-11186, ONR MURI N00014-01-0610, ARL Alion ST M1193391, ARI W911W07C0063, and AFOSR MURI FA9550-05-1-0388. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the Army Research Lab, the Army Research Institute, the Office of Naval Research, the Department of Defense, or the U.S. government. I am furthermore grateful to my committee for guiding me through this process.

References


CORDIS. Community Research and Development Information Service from [http://cordis.europa.eu/search](http://cordis.europa.eu/search)


39


