Methodological Workshop
Semantic Networks: Applications in Social Science and Humanities

May 18-19 2013
Centre for German and European Studies
St. Petersburg State University
Universitetskaya nab. 7/9
St. Petersburg
Tel./ Fax: +7 (812) 324 0885
E-mail: info@zdes.spbu.ru

Semantic network analysis is a mature yet still quickly advancing sub-domain of social science aiming to develop theories and techniques to better understand beliefs, meaning, identities, knowledge, discourse and language use in a wide range of application domains. A range of methods and software products have been developed to accomplish these tasks. These methods allow us to capture, model, analyze and visually represent the cognitive structures of organizations, groups and individuals, to compare them across scenarios, genres and domains, to trace their dynamics, and to relate them to other types of social structures. This workshop includes intense methodological lectures and training to provide participants with advanced "hands-on" experience of up-to-date semantic network analysis methods.

Speakers:
Jana Diesner, University of Illinois at Urbana Champaign, USA
Johanne Saint-Charles, Université du Québec à Montréal, Canada
Pierre Mongeau, Université du Québec à Montréal, Canada
Vitaliano Barberio, Vienna University of Economics and Business, Austria

Sessions format: 1 h. for presentations, 30 min. for questions.

Working language: English.

Equipment: to fully benefit from the workshop participants are invited to bring a laptop with Automap, ORA and UCINet preinstalled. However, all live walk-through exercises will be screen-projected.

Readings will be e-mailed to the participants prior to the workshop.

Timetable

<table>
<thead>
<tr>
<th>Time</th>
<th>Saturday May 18</th>
<th>Sunday May 19</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.00 – 10.15</td>
<td>Introduction</td>
<td>10.00 – 10.15</td>
</tr>
<tr>
<td>10.15 – 11.45</td>
<td>Session 1</td>
<td>10.15 – 11.45</td>
</tr>
<tr>
<td>11.45 – 12.00</td>
<td>Short break</td>
<td>11.45 – 12.00</td>
</tr>
<tr>
<td>12.00 – 13.30</td>
<td>Session 2</td>
<td>12.00 – 13.30</td>
</tr>
<tr>
<td>14.30 – 16.00</td>
<td>Session 3</td>
<td>14.30 – 16.00</td>
</tr>
<tr>
<td>16.00 – 16.15</td>
<td>Short break</td>
<td>16.00 – 16.15</td>
</tr>
<tr>
<td>16.15 – 17.45</td>
<td>Session 4</td>
<td>16.15 – 17.45</td>
</tr>
<tr>
<td>17.45 – 18.00</td>
<td>Short break</td>
<td>17.45 – 18.00</td>
</tr>
<tr>
<td>18.00 – 19.30</td>
<td>Session 5</td>
<td>18.00 – 20.00</td>
</tr>
</tbody>
</table>

Tutorial 6 h.
Session 1
10.15 – 11.45

From Words to Networks: The Importance of Methodological Choices for Extracting Semantic Networks from Texts

Jana Diesner, University of Illinois at Urbana Champaign, USA

Semantic networks are structured representations of information and knowledge that can be constructed based on text data, among other sources. Extracting semantic networks from texts - a process also known as relation extraction - requires humans to make a plethora of choices about suitable methods, subroutines and parameter settings. While many relation extraction methods have been applied across various corpora and domains, there is a lack of understanding of the impact of these choices on the resulting network data and analysis results.

In this talk, I first provide a comparative overview on common methods and tools for constructing semantic networks based on text data. Next, I will present on our research on applying common relation extraction methods to three large-scale, open-source corpora from the business, geopolitical and science domain and comparing the retrieved networks in terms of their structural properties. This comparison allows us to identify commonalities and differences between the following methods: First, we used thesaurus-based text coding, which requires a dictionary that associates text terms with entity classes that are relevant for representing real-world networks. Following the traditional approach used in the humanities and social sciences, we built thesauri in a semi-manual, semi-automated fashion, and then linked the identified entities based on their proximity. The second method resembles the first one, except for that we automatically generated thesauri by applying an entity extractor that we built by using supervised machine learning techniques. Third, we constructed network data from metadata, e.g. index terms and keywords provided by humans and text mining algorithms. Fourth, we collaborated with subject matter experts to build network data that can be considered as ground-truth data. I show how the considered methods can be combined such that we can gain a more comprehensive understanding of networks.

Another factor limiting the reliability of relation extraction methods and the resulting semantic networks is the propagation of errors throughout multi-step analysis procedures and pipelines. I will present our findings from a series of empirical experiments that we conducted to find answers to the following questions: How much variation in the structure and properties of semantic networks is due to the error rates of the involved sub-routines? Does increasing the accuracy of these techniques actually matter for network analysis results?

Readings:

Diesner J (2013) From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data. Künstliche Intelligenz / Artificial Intelligence. DOI: 10.1007/s13218-012-0225-0


Jana Diesner, University of Illinois at Urbana Champaign, USA

One of the most promising yet under-researched areas in network analysis is the meaningful joint consideration of text data and network data. We know that without considering the content of text data produced or shared by network participants and by analyzing text data and network in a disjoint fashion we are limited in our ability to understand the effects of language use in networks. This includes the transformative role that language can play in networks, and the interplay and co-evolution of information on one side and the structure and behavior of networks on the other side.

In this talk, I first discuss the common prior approaches that people have been using when both, text data and network data, are available as a source of information about networks. Next, I present our methodological advancements to these prior techniques and demonstrate the resulting changes in analytic capabilities and analysis results. I show how extracting content nodes from groups of structurally equivalent agents can be an appropriate strategy for enabling the comparison of content produced or shared by network participants. These equivalence classes can represent a variety of social roles and positions. In contrast to that, extracting content nodes from groups of structurally coherent agents can be suitable for enabling the enhancement of social networks with content nodes. I report on our results from applying these to various text corpora; focusing on data about scientific communities and language use in these communities. The findings include a comparison of the outcomes of topic modeling; an efficient and unsupervised information extraction technique, to the results from alternative methods, including entity extraction based on supervised machine learning.

Johanne Saint-Charles, Université du Québec à Montréal, Canada

A sociosemantic network may be conceived as a specific type of cognitive social network in which the relationships between individuals are based on their discourse similarities. In this workshop, we will present how to generate a social network based on semantic similarities between texts produced by individuals (answers to interviews, open-ended questions in a survey or any other written work). We will first present and explain the main steps of the procedure leading to the creation of a measure of semantic similarity. Through this presentation, we will also discuss the impact the choice of the semantic units has on the results. We will then show how to use this measure to create a sociosemantic network. Finally, we will show how to use an on-line semantic analysis program called Semato. If time permits, participants will be invited to perform some analysis themselves with this program.
Readings:

Short break
16.00 – 16.15

Session 4
16.15 – 17.45

**Linking social and sociosemantic networks: the challenge of choices**

*Johanne Saint-Charles, Université du Québec à Montréal, Canada*

*Pierre Mongeau, Université du Québec à Montréal, Canada*

It is an accepted postulate that sociosemantic networks should be related to social networks. But this relationship is not as straightforward as it may seem. We have been studying “transdisciplinary communities” whose objectives are to bring together people from various walks of life to address complex issues in the field of environmental health. In such communities, it is worth exploring how “relationships and discourse” are linked.

Data from these studies consist of “traditional” social networks gathered with sociometric questionnaires (collaboration, advice, trust, support, etc.) and text obtained through open-ended questions. We will discuss some social network measures (e.g. centralities, sub-groups, equivalence, etc.) that can be used to explore how social networks and sociosemantic networks are linked. We will present and explain some of these measures, illustrate the types of result they bring about and discuss the underlying theoretical assumptions.

Short break
17.45 – 18.00

Session 5
18.00 – 19.30

**A method to extract social structure and organizational identity from text and two practical applications**

*Vitaliano Barberio, Vienna University of Economics and Business, Austria*

Discourse - the social use of language - interfaces organizations and their both internal and external audiences. As such it mirrors shared understandings of social roles and positions - social structure - on the one hand, and the meaning of being part of a collective - organizational identity - on the other hand. We present a method for the systematic extraction and network representation/analysis of social structure and organizational identity from natural language text. The method is illustrated with two practical case studies. The first uses data collected from the website of the city of Vienna (AT) in 2011, the second uses data from the university president elections held in Bologna (IT) in 2008. The practical implication are that: a) strategic communication can be represented and improved with network methods and, b) intelligence on social organizational structures can be done with public data.
May 19

10.15 – 20.00 (with breaks)

Tutorial

From Words to Networks - Extraction and Analysis of Semantic Network Data from Text Data

Jana Diesner, University of Illinois at Urbana Champaign, USA

This interdisciplinary workshop introduces a selection of fundamental theories, concepts, methods and applications for bringing together text analysis and network analysis. Participants will learn how to conduct data analysis at the nexus of these areas in an informed, systematic and efficient fashion, and how to:

- Construct one-mode and multi-mode semantic networks from unstructured, natural language text data.
- Visualize and analyze semantic network data.
- Interpret semantic network analysis results.

Throughout the workshop, practical applications from the academic, administrative and business domain will be discussed. At the end of the workshop, participants will be able to design and conduct research projects for scholarly and commercial use in these fields.

Semantic networks are structured representations of information and knowledge, and are often constructed based on textual data. Socio-technical networks represent interactions between social agents, infrastructures and information. These networks can also be extracted from text data and are then often also referred to as semantic networks. The functioning and dynamics of all of these networks involve the continuous production, processing and flow of information. Information about semantic networks is often explicitly or implicitly contained in text data, such as news wire data, scientific information such as publications and patents, communication data such as conversations transcripts and emails, self-presentations such as mission statements and annual reports, and social media data such as tweets and wikis. Using text data to construct or enhance network data has been used to answer questions such as:

- Who is talking to whom, and about what?
- What are the mental models of individuals or groups about certain topics?
- How do memes and innovations emerge and spread in society and online?
- Who are the key entities in a network?
- What benefits and risks result from an observed network structure for an organization and its wider context?

The main component of this workshop is to teach to participants practical, hands-on skills in working with text analysis and network analysis tools. Participants will perform basic natural language processing techniques on the lexical, syntactic and semantic level including:

Readings:


- Pre-process texts with techniques such as reference resolution, stemming and parts of speech tagging.
- Identify salient concepts and themes from single documents and entire text collections.
- Create and apply codebooks, which are also known as dictionaries or thesauri.
- Locate and classify entities that can serve as nodes for networks. We will move beyond the classic set of entity classes (people, organizations, locations) to also consider other classes that relevant for studying social processes and culture, e.g. tasks, resources and knowledge.
- Relation Extraction, linking entities into edges based on various criteria.

Participants will also perform basic network analysis techniques, including:

- Manipulate and visualize network data.
- Compute basic network metrics on the graph and node level.
- Identify meaningful groups and clusters of nodes.

Going from texts to networks involves some principles and strategies originating from computer science that are not only applicable to the task at hand, but to a wide range of problems. These principles and strategies are referred to as “Computational Thinking” - a basic skill like reading, writing and arithmetic that is crucial for solving problems and understanding human behavior across fields (Wing 2006). In this workshop, participants are introduced to Computational Thinking and practice applying this way of thinking.

**Software:** Prior to the workshop, participants will receive an email with links to the software tools that will be used for the workshop. Participants are invited to bring laptops to the workshop. If one cannot bring a laptop he/she will still fully benefit from the workshop as we screen-project all live walkthrough exercises. At the workshop, a tutorial document and further learning resources will be provided.

**Data:** Participants can work with the sample data provided by the instructor and/or bring their own data. If participants bring their own text data, a sample of not more than 20 text documents of less than two pages in length is recommended, in line with network data with not more than 200 nodes. The tools we use scale up to larger data sets, but large-scale data might not be practical for training purposes.

**Readings:**

**Mandatory:**


**Optional:**


http://www.faculty.ucr.edu/~hanneman/nettext/.


The instructor is available for pointing participants to further readings in their areas of interest at jdiesner@illinois.edu.