ConText:
Network Generation from Text Data

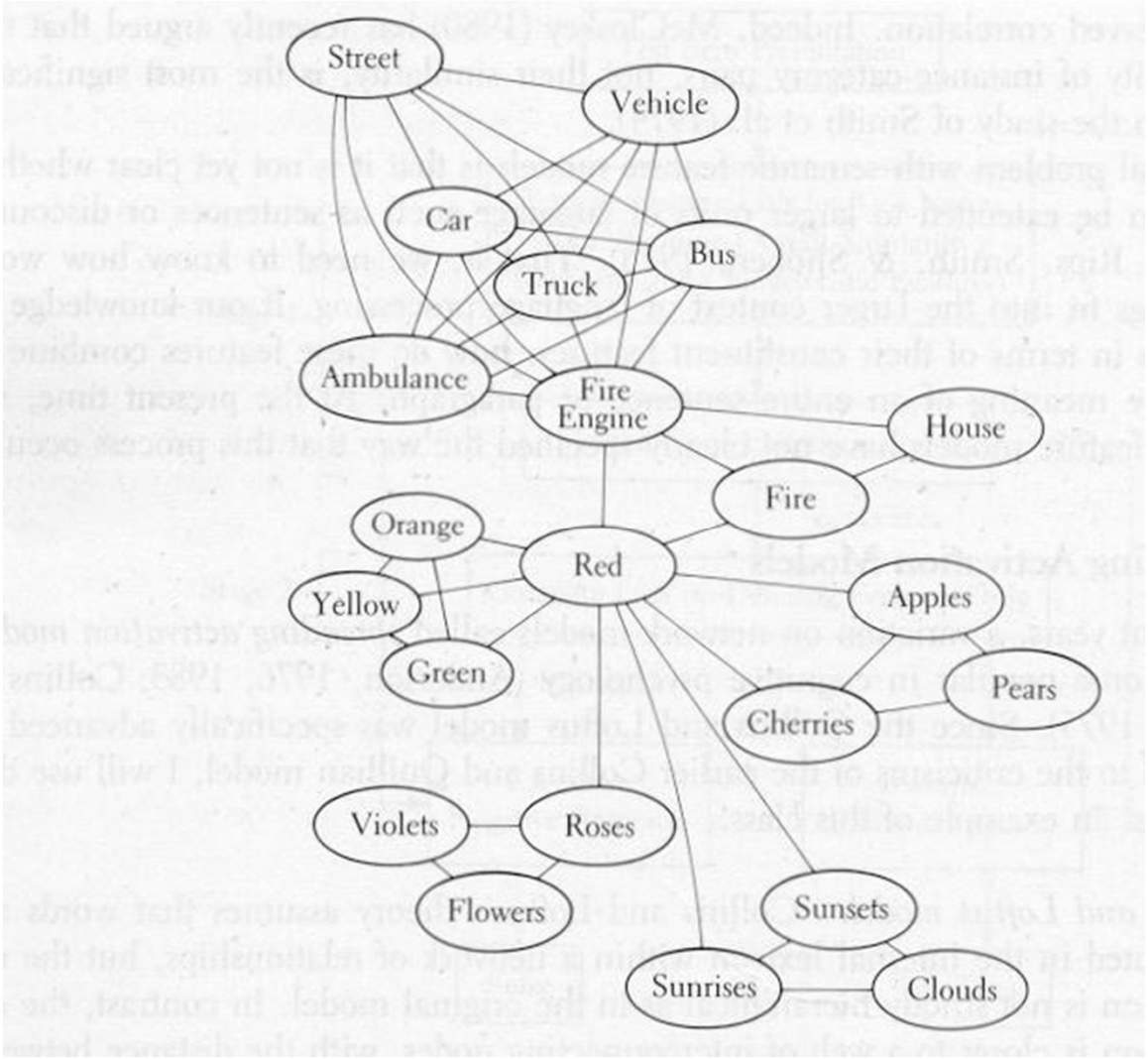
Jana Diesner

Bochum, Apr 2015

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

The concept of Semantic Networks

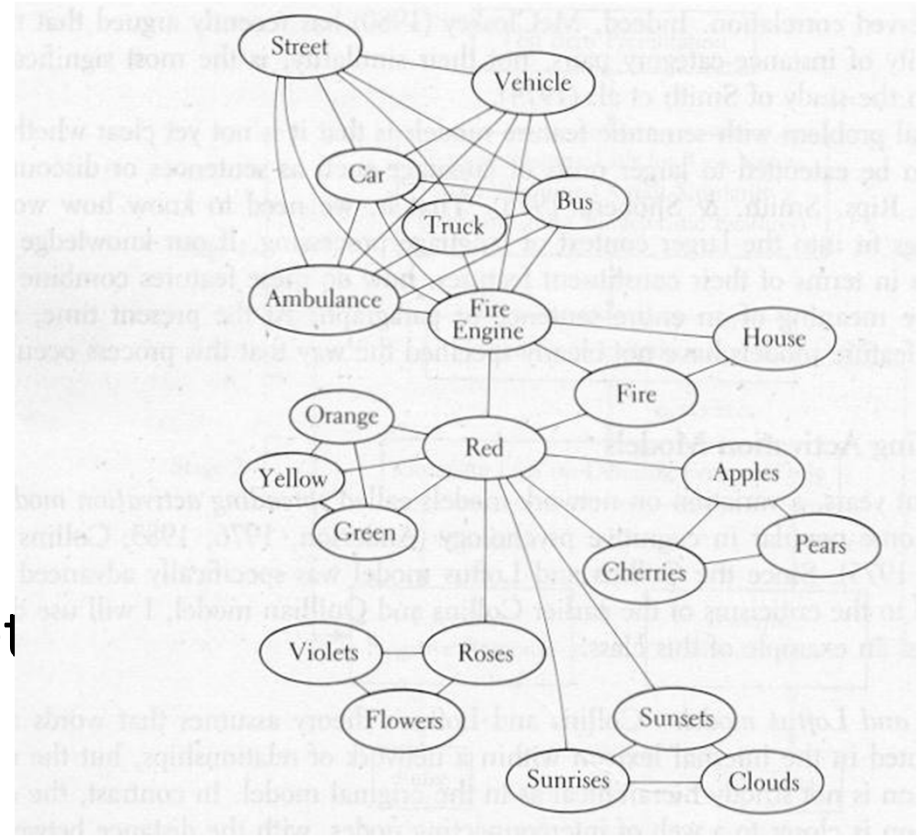


- Structured representations of information and knowledge
- Originally meant to be used for reasoning and inference
  - E.g. theory of spreading activation

Collins and Loftus (1975)

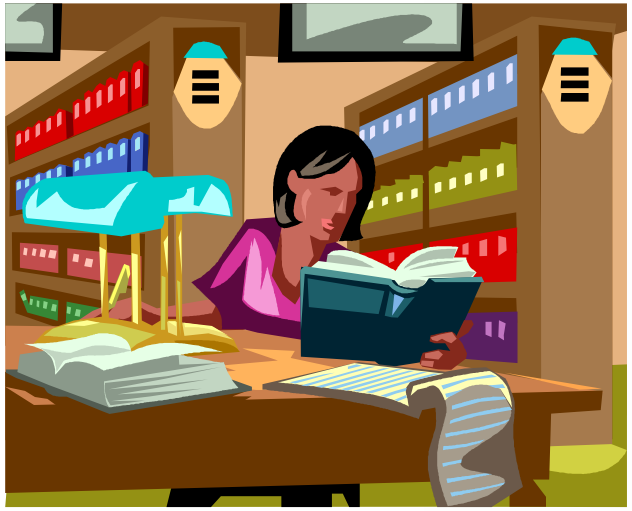
Semantic Networks: Usage

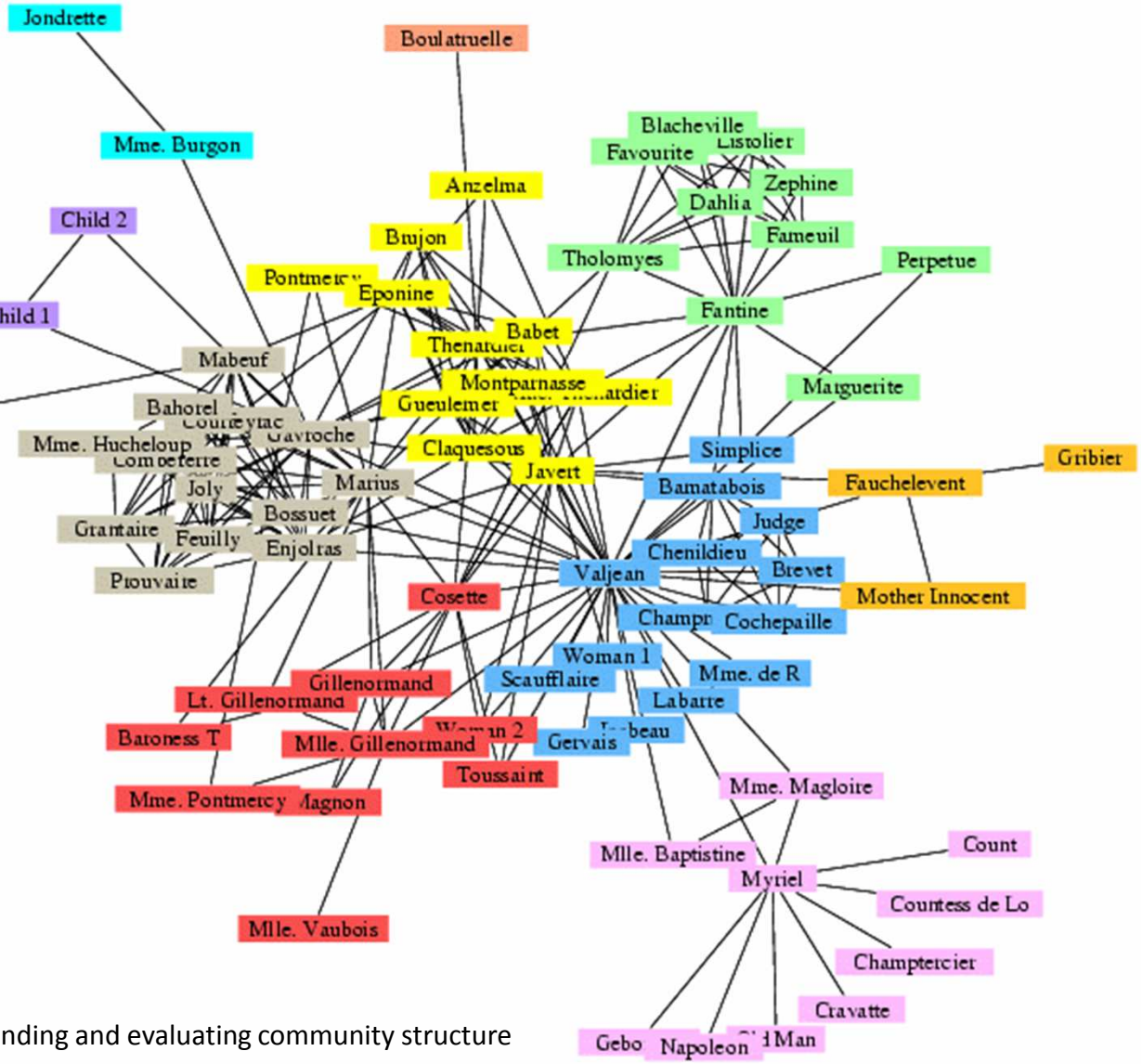
- **Summarization**: retrieve a concept's ego-network to distill the essence of some data in concise and structured form
- **Disambiguation**: network clustering to identify different aspect of the meaning of a concept
- **Prediction**: forecast the set of concepts that will be evoked when a certain note is activated



Griffiths, Steyvers & Tenenbaum 2007

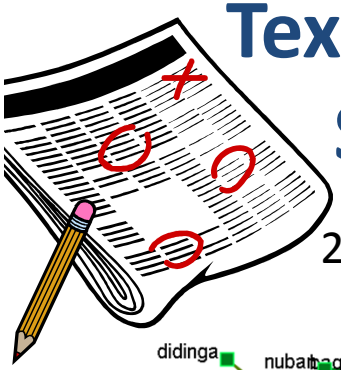
Text based Networks: Association Networks

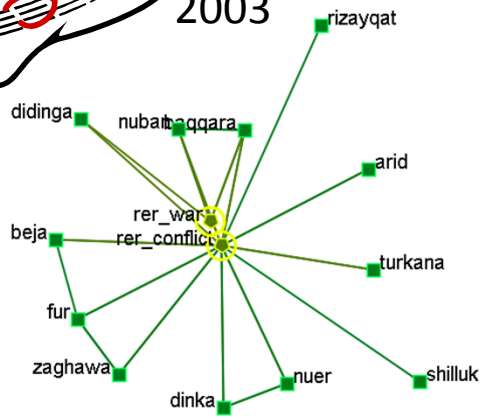


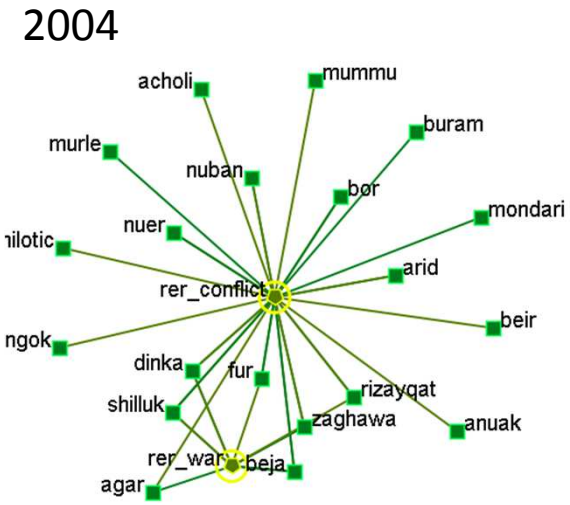


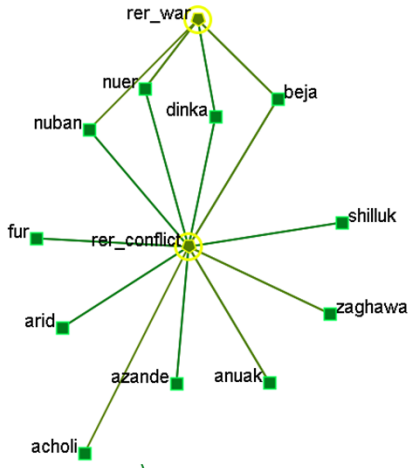
Les Miserables: M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69, 026113 (2004).

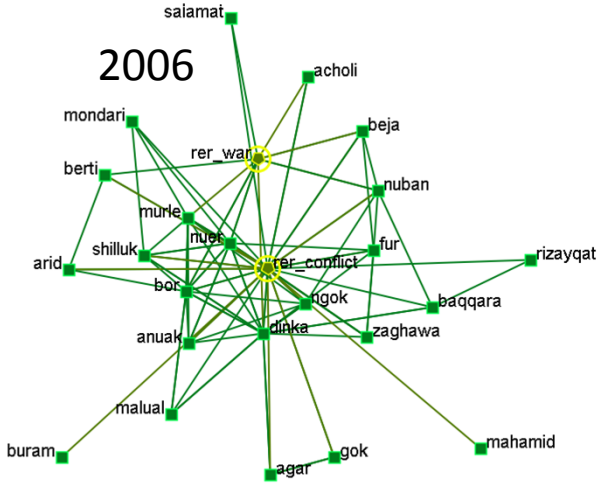
Text based Networks: Relation Extraction: Social Networks (of tribes in Sudan)

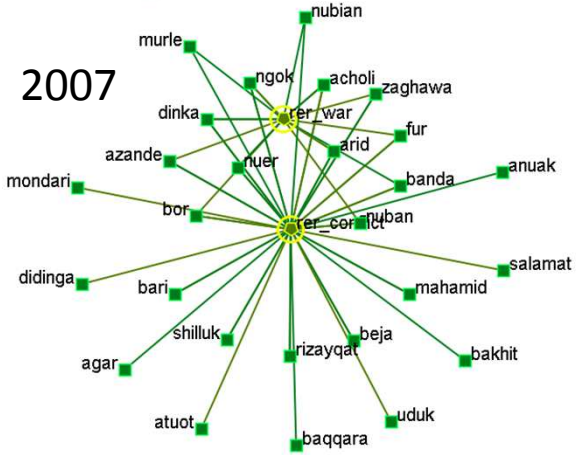


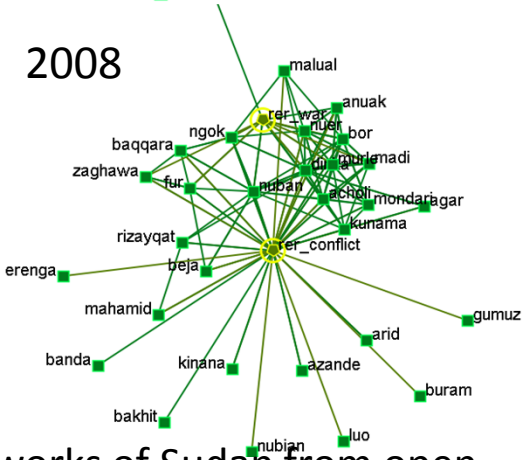






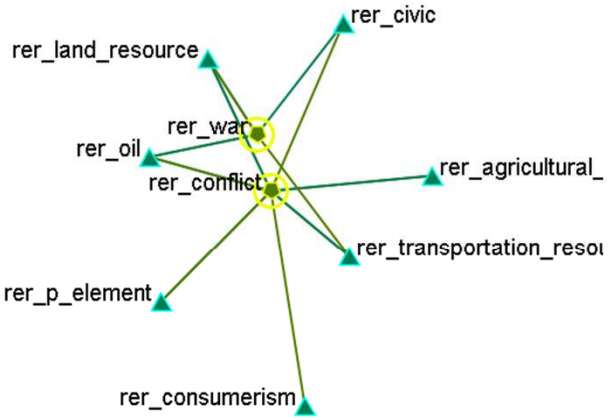


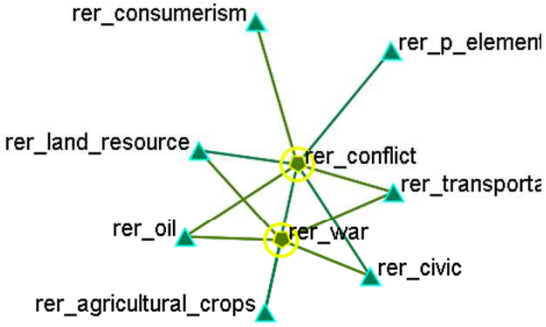


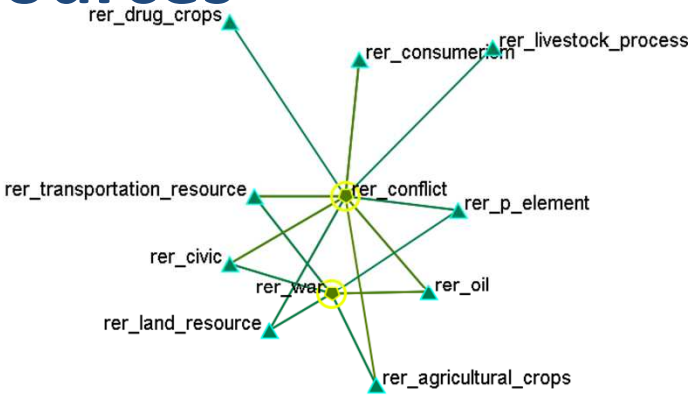


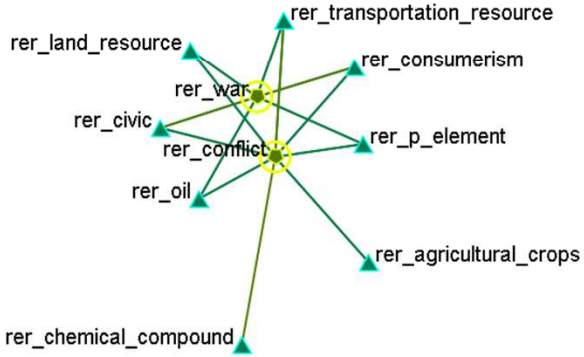
Diesner J, Carley KM, Tamabyong L (2012) Mapping socio-cultural networks of Sudan from open-source, large-scale text data. Journal of Computational and Mathematical Organization Theory (CMOT) 18(3), 328-339.

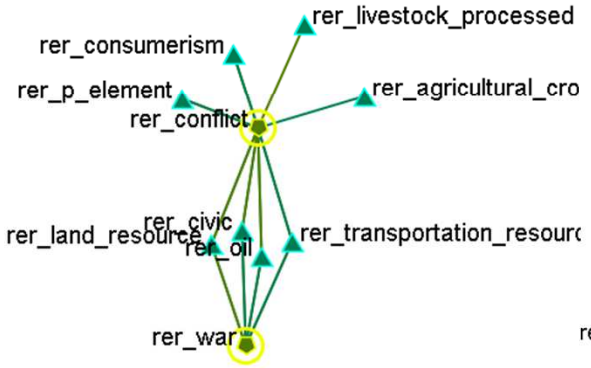
Text based Networks: Relation Extraction: Organizations and Resources

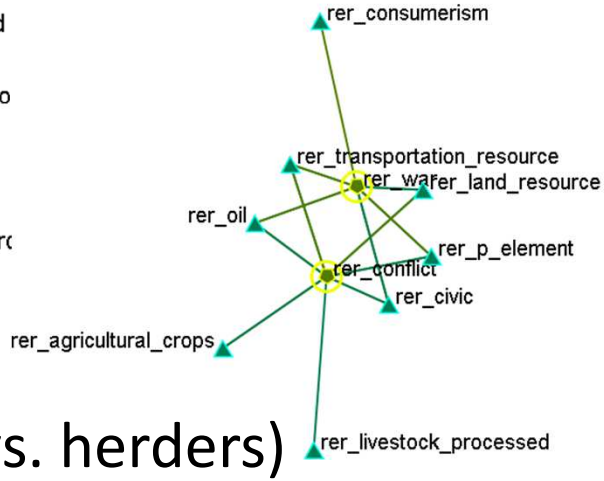












- Conflict: Agriculture, Livestock (farmers vs. herders)
- War: Land Resource (concept of *dar*)
- Conflict and War: Oil, Civic, Transportation

Text based Networks: Meta Data Networks

to reform the NDP should be forewarned:climate change and crony capitalism are excellent reasons to reboot thepro-corporate R850-billion infrastructure spend, and to redirect the funds tomeet basic infrastructure needs. But Molewa's defunding of her climateresponsibilities gives an accurate indicator of why reform is highly unlikely without a rapid increase in eco-social activism. Patrick Bond's Politics ofClimate Justice was recently named as one of the Guardian's10 leading books on the topic

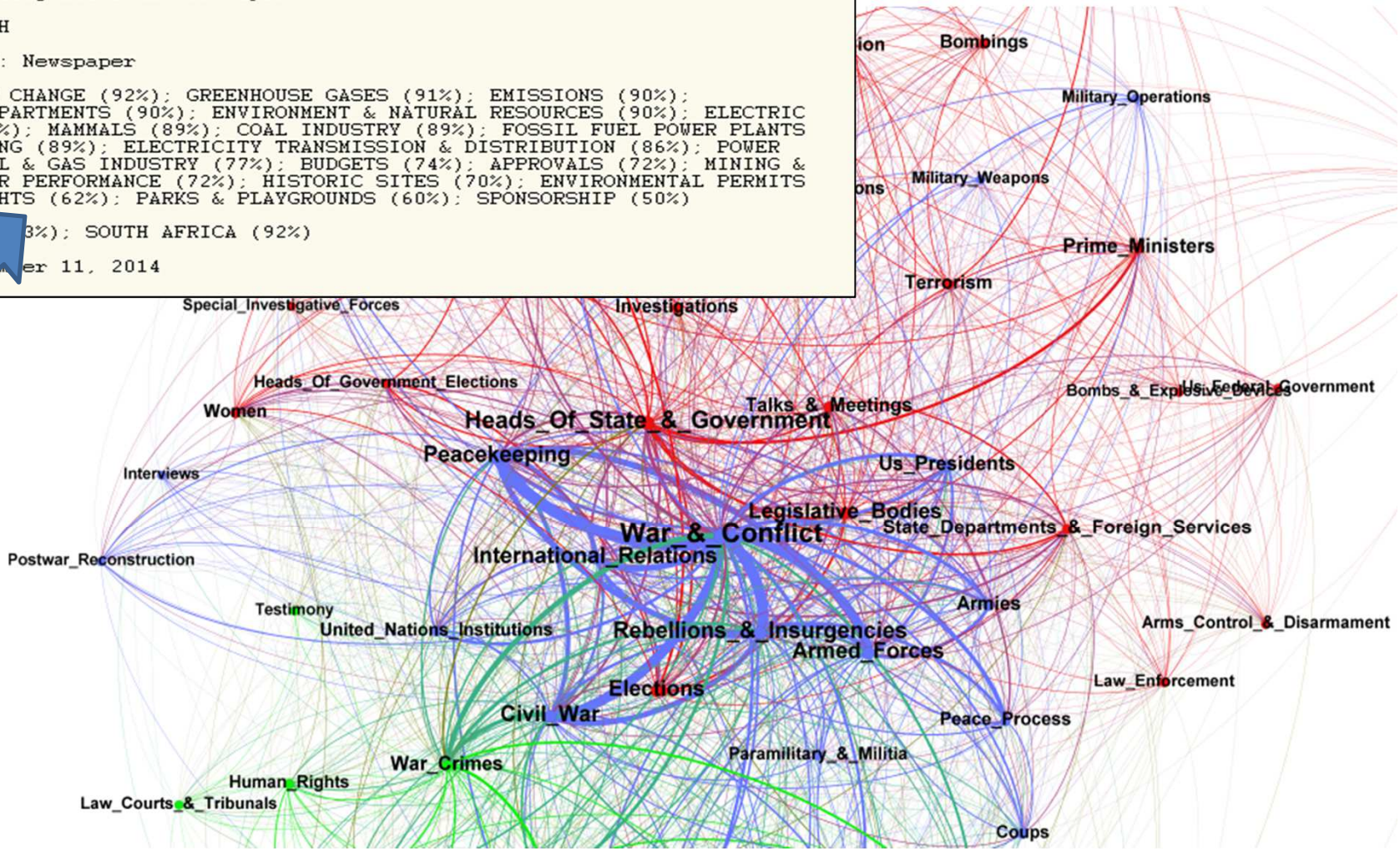
LANGUAGE: ENGLISH

PUBLICATION-TYPE: Newspaper

SUBJECT: CLIMATE CHANGE (92%); GREENHOUSE GASES (91%); EMISSIONS (90%); ENVIRONMENTAL DEPARTMENTS (90%); ENVIRONMENT & NATURAL RESOURCES (90%); ELECTRIC POWER PLANTS (89%); MAMMALS (89%); COAL INDUSTRY (89%); FOSSIL FUEL POWER PLANTS (89%); COAL MINING (89%); ELECTRICITY TRANSMISSION & DISTRIBUTION (86%); POWER PLANTS (77%); OIL & GAS INDUSTRY (77%); BUDGETS (74%); APPROVALS (72%); MINING & EXTRACTION SECTOR PERFORMANCE (72%); HISTORIC SITES (70%); ENVIRONMENTAL PERMITS (67%); WATER RIGHTS (62%); PARKS & PLAYGROUNDS (60%); SPONSORSHIP (50%)

COUNTRY: AFRIC 3%); SOUTH AFRICA (92%)

LOAD-DATE: er 11, 2014



Basic Types of Information in Text Data

- **Morphology**: structure of words
  - E.g. spelling, inflections, derivations
- **Syntax**: relationships between words
  - e.g. parts of speech tagging
- **Semantics**: meaning of language
  - e.g. word sense disambiguation, grammars
- **Pragmatics**: language in context and social use of language
  - e.g. sentiment analysis, discourse analysis
- **Relation Extraction (this lab)**: borrows from all of the above

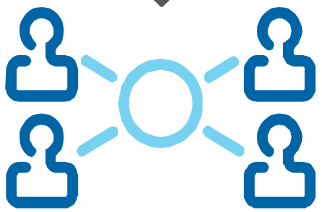
Methods for Constructing Networks of Words

	Automation	Abstraction	Generalization
1. Mental Models (Spreading Activation) (Collins & Loftus 1975)			
2. Case Grammar and Frame Semantics (Fillmore 1982, 1986)			
3. Discourse Representation Theory (Kamp 1981)			
4. Knowledge representation in AI, assertional semantic networks (Shapiro 1971, Woods 1975)			
5. Centering Resonance Analysis (Corman et al. 2002)			
6. Mind maps (Buzan 1974)			
7. Concept maps (Novak & Gowin 1984)			
8. Hypertext (Trigg & Weiser 1986)			
9. Qualitative text coding (Grounded Theory) (Glaser & Strauss 1967)			
10. Definitional semantic networks incl. text coding with ontologies (Fellbaum 1998)			
11. Semantic Web (Berners-Lee et al. 2001, Van Atteveldt 2008)			
12. Frames (Minsky 1974)			
13. Semantic Grammars (Franzosi 1989, Roberts 1997)			
14. Network Text Analysis in social science (Carley & Palmquist 1991)			
15. Event Coding in pol. science (King & Lowe 2003, Schrodt et al. 2008)			
16. Semantic networks in comm. science (Danowski 1993, Doerfel 1998)			
17. Probabilistic graphical models (Howard 1989, Pearl 1988)			

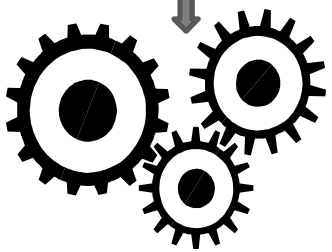
Diesner, J. (2012). Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts. CMU-ISR-12-101, Carnegie Mellon University.

Text Data → **Network Analysis** → **Applications**

Need: scalable, reliable, robust methods & technologies



Network representation



Analysis tools

- Answer substantive questions about social systems
- Visualizations
- Fill databases for archiving, search, retrieval
- Forecasting: Explore future behavior and what-if scenarios

Putting it all Together

- What happens once we have extracted relational data from texts?
  - Load data into network analysis toolkit and perform analysis, visualization, simulation, etc…
- Relation extraction from texts can be an alternative or supplemental relational data collection technique when:
  - Classic methods for collecting network data fail
  - Analysis of large amounts of data is constrained

Bravo! You have passed the primer in network text analysis!

Hands-on part: Types of techniques

1. **Network Construction**
    1. **From Meta Data**
    2. **From Text Data**
        1. **Node identification (and classification)**
            1. Dictionary-based (manually and/ or automatically built)
        2. **Edge identification (and classification)**
            1. Identification: Proximity based approach (co-occurrence)
        3. **Text pre-processing**
            1. Natural Language Processing (NLP) techniques: precondition for finding meaningful information, incl. representations of nodes and edges, in text data
            2. Selection of relevant entities (positive filter: dictionary, entity extraction) or removal of irrelevant information (negative filter: delete list) given the research question, data, domain

Extracting Network data from meta data

Construction of Meta Data Networks

- Exploit meta data information that come along with data
- Example: LexisNexis Academic
- How to:
  - Data curation: lexisnexis
    - Pre-process: split up, clean up, put into relational database
  - Network construction: meta data
    - Generate networks: construct weighted networks of people, organizations, locations, information
- Visualize resulting networks in Gephi

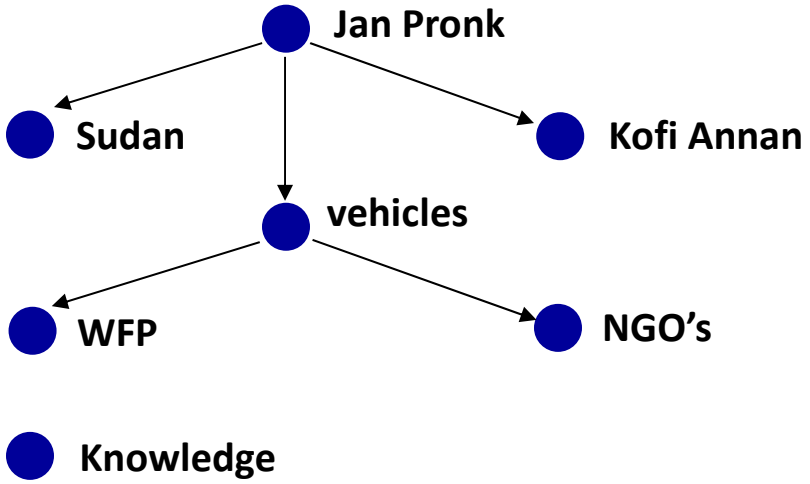
Extracting network data from text data

Association Networks and Relation Extraction: One-mode and Multi-mode Networks

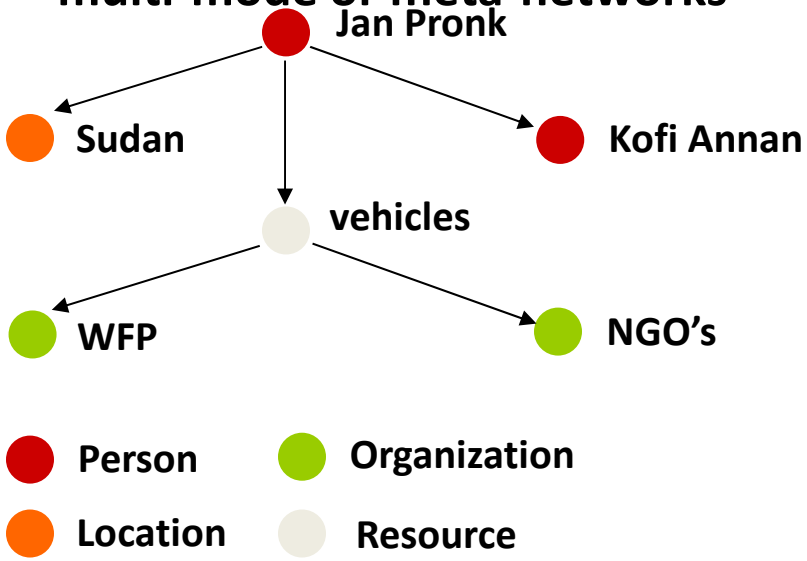
Example from UN News Service (New York), 12-28-2004: "Jan Pronk, the Special Representative of Secretary-General Kofi Annan to Sudan, today called for the immediate return of the vehicles to World Food Programme (WFP) and NGOs."

extract relational data:

one mode or semantic networks



multi-mode or meta-networks

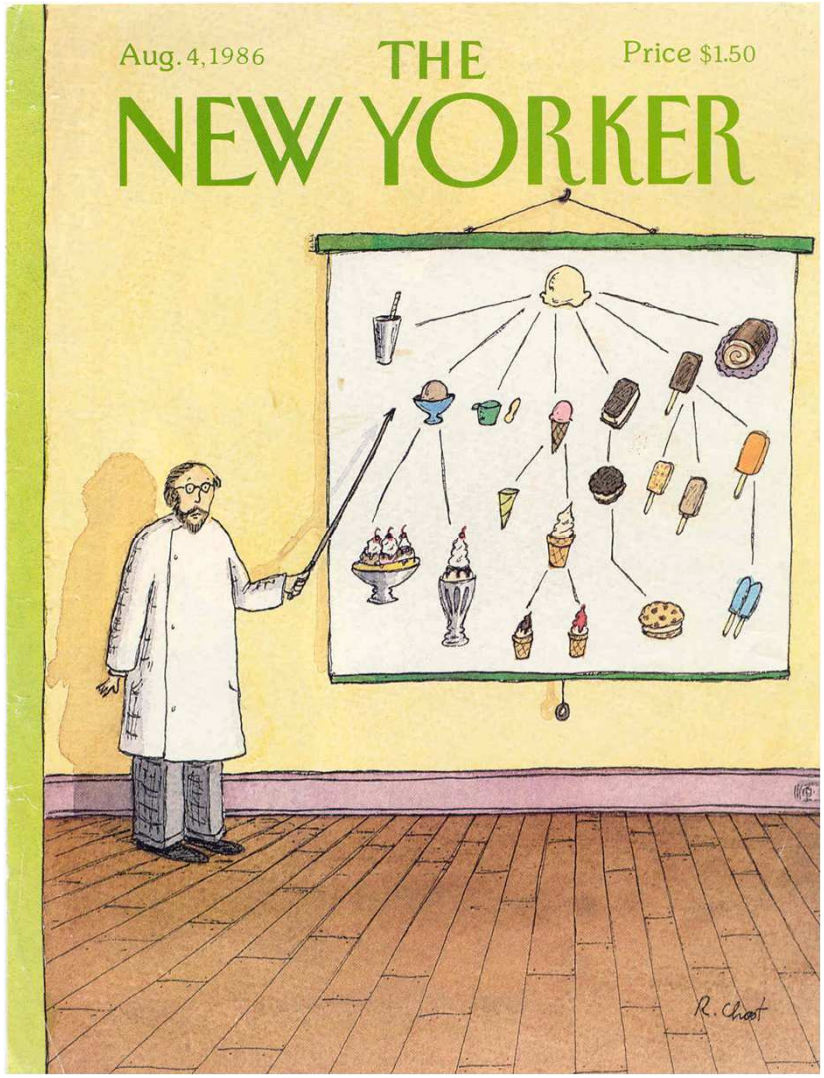


Identify relevant nodes

Classify identified nodes

Classification Schema

- Ontology:
  the study of being or existence
- Taxonomy:
  practice & science of classification



Dictionary: Data Structure

- Text term (incl. n-gram), concept, entity class
- Functions: disambiguation (1,2), consolidation (3,4), n-gram concatenation (3,4)
- Examples:
  1. Apple, apple, organization
  2. apple, apple, resource
  3. Barack Obama, Barack_Obama, agent
  4. Barack Hussein Obama, Barack_Obama, agent

Dictionary: Usage

- How to:
  - ConText: Text Analysis: codebook application
  - See impact of codebook on text data
    - » Helps to refine codebook
  - Options:
    - Normalization vs. entity class tagging
    - Refinement (replace and insert) versus positive filter
    - Positive filter: notion of adjacency

Dictionary: Construction

- From rule-based and deterministic methods to probabilistic and machine-learning based methods
- ConText:
  - Load raw data, no preprocessing (needs proper syntax if possible)
  - Text Analysis: Analysis: Entity Detection
    - Use as baseline, refine

Dictionary: More details

- How to use them:
  - List relevant entities: serves as positive filter
    - Then construct one-mode network, aka semantic network
  - Cross-classify relevant entities with ontological categories
    - Then construct multi-mode network
- Limitations of manual approach:
  - Tedious, incomplete, outdated, deterministic
- Help for constructing thesauri:
    - Computer-supported: terms with high (weighted) frequency
    - Automated: entity detection
    - External sources (e.g. CIA World Fact Book, WordNet)
    - Other automated techniques, e.g. Bootstrapping

Codebook Construction: Corpus Statistics

- Cumulative frequency: Bag of Words
- How to:
  - Load texts into ConText
  - Create Corpus Statistics
  - This is one dimension of salience, prominence, importance
  - What are other dimensions?

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley

Codebook Construction: Corpus Statistics: TF*IDF

- What determines word's importance in corpus?
  - Discriminating and distinguishing
- tf = term frequency (importance of term within document)

$$tf = \frac{cumulative\ occurrence\ of\ term\ x\ in\ document\ y}{total\ number\ of\ terms\ in\ document\ y}$$

- idf = inverse document freq. (importance of a term in corpus)

$$idf = \log \frac{total\ number\ of\ documents\ in\ corpus}{total\ number\ of\ documents\ containing\ term\ x}$$

*tfidf = tf * idf*

- tfidf: strategy and measure
  - High if tf = high and df = low
  - High for signal, low for noise
- How to:
  Generate: Concept
  List: Union

Codebook Construction: Stop words

- Stop words listed in delete list
- Serves as negative filter (remove from text data what's contained in list)
- How to:
  - Text Analysis: preprocessing: remove stop words
  - How to construct a delete list?
    - Use predefined lists (data folder)
    - Construct your own, one entry per line (incl. n-grams)
  - Notion of adjacency (direct vs. rhetorical, which maintains original distance of words)

Text pre-processing: Stemming

- Detects inflections and derivations of concepts
- Converts each term into its morpheme
- How to:
  - ConText: Text Analysis: preprocessing: stemming
- Two families of stemmers:
  - Porter (rule-based): high efficiency, poor human readability (ConText)
  - Krovetz (dictionary-based): lower efficiency, better human readability

Porter, M.F. 1980. An algorithm for suffix stripping. *I 14* (3): 130-137.

Krovetz, Robert (1995). *Word Sense Disambiguation for Large Text Databases*. Unpublished PhD Thesis, University of Massachusetts.

Text pre-processing: Parts of Speech (POS) Tagging

- Task: Assign single best tag or parts of speech (POS) to each word in a corpus (e.g. noun, verb, adjective, or adverb)
- What is the challenge here?
  - Ambiguity resolution (can) - words match multiple tags depending on their context
- ConText: text analysis: preprocessing: part of speech tagging

Church, K. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. 2nd Conference on Applied Natural Language Processing, Austin , TX, 136-143.

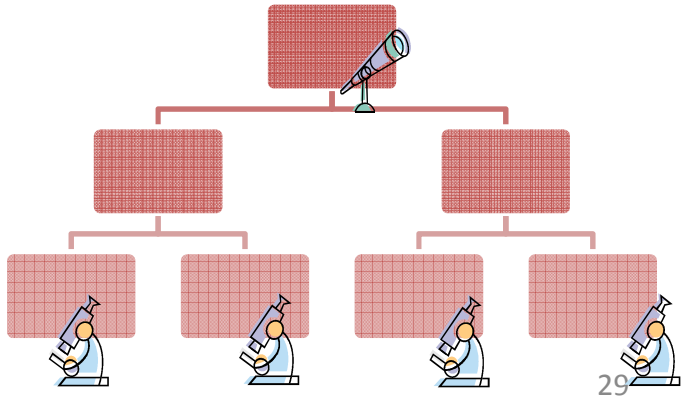
Diesner, J., & Carley, K. M. (2008). *Looking under the hood of stochastic part of speech taggers (No. CMU-ISR-08-131R): Carnegie Mellon University, School of Computer Science, Institute for Software Research.*

Finding salient terms: N-grams

- Meaningful multi-words units
- How to:
  - Generate: generalization thesaurus: bigram
  - Sort by decreasing frequency and decreasing tfidf, pick suitable entries, removed duplicates

Text pre-processing: When to stop?

- When to stop? ("criteria")
  - Orcam's razor:
    - 14th-century English logician and Franciscan friar, William of Ockham.
    - Aka lex parsimoniae (law of parsimony)
    - Basic idea: All other things being equal, the simplest solution is the best.
    - Why does it matter?
      - Don't want: Overfitting
      - Want: Generalizability



Linking nodes: Approaches

- **Syntax and surface patterns** (Fillmore, Schrodt)
  - Linguistics: parsing trees
- **Logical and Knowledge Representation in Artificial Intelligence** (Shapiro)
  - first order calculus, predicate logic (quantifiers)
- **Distance based** (Danowski)
  - Communications: distance in text (windowing) or in space (Euclidean)
- **Probabilistic, learning from data** (McCallum)
  - Machine learning techniques: probabilistic (Bayesian), kernels (N-dimensional similarity), graphical models (hidden markov models, conditional random fields), boot strapping

Cites and summary in Diesner, J., & Carley, K. M. (2010). Relation Extraction from Texts (in German, title: Extraktion relationaler Daten aus Texten). In C. Stegbauer & R. Häußling (Eds.), Handbook Network Research (Handbuch Netzwerkforschung) (pp. 507-521). Vs Verlag.

Distance-based node linkage

- How to:
  - Right hand side of codebook construction panel
  - ConText: network construction: codebook a
    - One mode (semantic network) versus multi mode
  - Set parameters:
    - Codebook
    - Aggregation: one network per document or for all documents
    - Distance: an appropriate window size
    - Unit of analysis: where to stop the sliding window

References

- Recommended further readings related to this session:
  - McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. ACM Queue, 3(9), 48-57.
  - Diesner, J., Carley, K. M. (2011): Semantic Networks. In G. Barnett (Ed), Encyclopedia of Social Networking, (pp. 595-598). Sage Publications.
  - Diesner, J., Carley, K. M. (2011): Words and Networks. In G. Barnett (Ed.), Encyclopedia of Social Networking, (pp. 958-961). Sage Publications.

References

- Mental Models:
  - Johnson-Laird, P. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.
  - Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, 20, 403-437.
  - Rouse, W. B., & Morris, N. M. (1986). On looking into the black box; prospects and limits in the search for mental models. *Psychological Bulletin*, 100, 349-363.

Thank you!
Q&A

- For any questions, comments, feedback, follow-up-now and in the future:
  Jana Diesner
  Email: jdiesner@illinois.edu
  Phone: ++1 (412) 519 7576
  Web: http://people.lis.illinois.edu/~jdiesner