Using Network Text Analysis to Detect the Organizational Structure of Covert Networks*

Jana Diesner, Kathleen M. Carley
Carnegie Mellon University
diesner@cs.cmu.edu, kathleen.carley@cmu.edu

Abstract

Herein we demonstrate how to get from sets of texts to network representations of covert networks represented in texts. We report on the application of a computer-supported technique that combines network analysis of texts with classifying social and organizational systems into an ontology called the meta-matrix model. The resulting combinatory method is referred to as Meta-Matrix Text Analysis. We apply this technique to detect the social and organizational structure of a Mideastern country. Social agents covered in our coding are people and organizations identified by human subject matter experts to be relevant to intelligence matters in that area.

Contact:
Jana Diesner, Kathleen M. Carley
Center for Computational Analysis of Social and Organizational Systems (CASOS)
Institute for Software Research International (ISRI)
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
Tel: 1-412-268-5866 (Diesner), 1-412-268-6016 (Carley)
Fax: 1-412-268-1744
Email: diesner@cs.cmu.edu, kathleen.carley@cmu.edu

Key Words: covert networks, network text analysis, meta-matrix text analysis, social network analysis, AutoMap, DyNetML, ORA

1. Introduction to the Methodology

Machine readable texts that convey information about covert networks are available on a large scale. In order to extract the organizational structure of covert networks effectively and efficiently from texts, appropriate tools and techniques are needed (Coffman, Greenblatt & Marcus, 2004). Furthermore, current hypotheses and theories about the structure of covert networks are oriented towards dynamic, complex and large-scale systems (Carley, 2003). This requires tools and methods that enable analysts to gain multi-level access to the meaning of textual data.

Network Text Analysis (NTA) is one method for encoding the relationships between words in a text and constructing a network of the linked words (Popping, 2000). The technique is based on the assumption that language and knowledge can be modeled as networks of words and the relations between them (Sowa, 1984). Several NTA methods exist (for an overview see Popping, 2000; Popping & Roberts, 1997, for discussion of empiric studies see Monge & Contractor, 2003), such as Centering Resonance Analysis (Corman et al., 2002), Functional Depiction (Popping & Roberts, 1997), Knowledge Graphing (Bakker, 1987; James, 1992; Popping, 2003), Map Analysis (Carley, 1988; Carley & Palmquist, 1992), Network Evaluation (Kleinnijenhuis, Ridder & Rietberg, 1996), and Word Network Analysis (Danowksi, 1982). Since the terror attacks on September 11, 2001 in the USA, research also focuses on visualizing covert networks extracted from texts (Krebs, 2001; Batagelj, Mrvar & Zaveršnik, 2002; Johnson & Krempel, 2004).

* **Acknowledgement:**
This work was supported in part by the National Science Foundation under grants ITR/IM IIS-0081219, IGERT 9972762 in CASOS, and CASOS – the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University (http://www.casos.cs.cmu.edu). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. government. We thank Dan Wood for his help with collecting and pre-processing the data.

In the research presented herein, we use map analysis. Map analysis systematically extracts and analyzes the links between words in a text in order to model the authors "mental map" as networks of words. Maps are a cognitively motivated representation of knowledge (Carley, 1988). Coding texts as maps focuses analysts on investigating the meaning of texts, because it detects the relationships between and among words and themes (Alexa, 1997; Carley, 1997a). Therefore, map analysis facilitates the analysis of quantitative characteristics of textual data as well as the extraction of meaning from texts (Carley, 1997b). In map analysis, a concept is a single idea represented by one or multiple words (e.g. Government, White House). Concepts are equivalent to nodes in Social Network Analysis (SNA). The link between two concepts is referred to as a statement, which corresponds with an edge in SNA. The union of all statements per texts forms a map (Carley, 1997b). Maps are equivalent to networks. We have operationalized and formalized the map analysis technique and implemented it into the AutoMap software (Diesner & Carley, 2004a).

If the research goal is to reveal social structure from texts, one strategy is to furthermore classify the concepts in maps extracted from texts into an ontology that models social systems as the entities the system is composed of and the connections between these entities. Such entities are, for example, people, groups that people are affiliated with, activities they engage in, resources at their disposal, etc.. The meta-matrix model provides one ontology for modeling multi-mode and multi-link social and organizational structure (Carley 2003, 2002; Krackhardt & Carley 1998). Previously, we have described an approach for combining the map analysis methodology with the meta-matrix model (Diesner & Carley, 2004b). The resulting integrative technique we refer to as Meta Matrix Text Analysis (ibid). This technique enables analysts to extract not only networks and meaning, but also social and organizational structure from texts. We have integrated the Meta Matrix Text Analysis technique into the AutoMap software. Entities of the meta-matrix that we implemented in AutoMap are agents, knowledge, resource, task/events, organizations, location, actions, roles, and attributes. Herein, we present the application of Meta Matrix Text Analysis to detect and analyze the structure of a covert network, which is a Mideastern country, from a set of texts.

2. Data

The Mideast covert network data set consists of 247 texts collected at CASOS. Of those texts 202 were gathered from LexisNexis Academia via exact matching Boolean keyword search. Search terms were 116 names of people and organizations identified by subject matter experts (SMEs) to be of critical importance to intelligence related matters during the past 25 years in a Mideastern country. We sorted the retrieved texts per person and organizations by relevance and selected the top most articles. The media we searched were major papers, magazines and journals. Sources for the 45 other texts in the Mideast data set were non-classified trial transcripts, excerpts from books, scientific articles, and texts from web pages. The time frame of our data set ranges from articles published in 1977 to 2004. The Mideast set contains 12952 unique concepts and 126496 total concepts. The number of unique concepts considers each concept only once per text, whereas the number of total concepts also considers repetitions of concepts per text.

3. Network Text Analysis Setup

Before NTA is run, texts can be pre-processed in order to condense the data to the concepts that are relevant for answering the research question. Therefore, pre-processing simplifies the task of making meaningful interpretations and comparisons across texts. In this study we focus on terms related to the social and organizational structure of a Mideastern country.

The first pre-processing technique we applied is deletion. Deletion removes non-content bearing concepts such as conjunctions and articles from texts (Carley, 1993). Our delete list has 178 entries. With that we removed 14.1% of the unique concepts and 41.4% of the total concepts from the texts. In a second step we stemmed the data. Stemming converts each concept into its related morpheme (Jurafsky & Martin, 2000: 83, 654). As a result, the number of unique concepts was further reduced by 27%, whereas the number of total concepts remains unchanged. Next we used AutoMap's Named-Entity Recognition functionality in order to generate a list of proper names, numerals, and abbreviations contained in the data (Magnini, Negri, Prevete & Tanev, 2002). We used the resulting list of 4539 Named-Entities as a basis for indexing agents, organizations, places, and events in two thesauri. A thesaurus in general is a two-columned collection that associates text-level concepts with higher-level concepts (Burkart, 1997; Klein 1997). The text-level concepts represent the content of a data set, and the higher-level concepts represent the text-level concepts in a generalized way. Thesaurus creation and application was the fourth pre-processing strategy we used. When applying a thesaurus, AutoMap searches the texts for the text level concepts specified in the thesaurus and translates matches into the related higher level concepts. First we built a generalization thesaurus with 1605 entries, which served two main purposes:

- Translate various (mis)spellings, aliases, and synonyms of names into core ids of Named Entities.

- Convert multiple-concept expressions into one concept (e.g Council of Guardians into Council_of_Guardians).

After this procedure, we built a meta-matrix thesaurus that classifies the concepts in the pre-processed texts into entity classes of the meta-matrix. Since one concept can represent several meta-matrix entity classes, a meta-matrix thesaurus can consist of more than two columns. Only concepts relevant to the research question need to be associated with meta-matrix entities in such a thesaurus. The application of a meta-matrix thesaurus enables the ontology-based extraction of social and organizational structure reflected in texts. Table 1 provides quantitative information on the meta-matrix thesaurus and the impact of its application to the data.

Table 1: Meta-Matrix Thesaurus and impact on data

Meta-matrix Entity	Unique # of entity analyzed	Total number of entity analyzed	Percentage of texts analyzed entity occurs in	Total number of entity linked into edges	Percentage of text linked entity occurs in
agent	415	2703	95.5%	4243	94.7%
knowledge	186	1611	89.5%	1937	83.8%
resource	274	2228	87.0%	2603	80.6%
task-event	74	1024	76.1%	1184	66.0%
organization	309	3746	96.4%	4274	94.7%
location	282	3888	98.0%	4545	96.4%
role	257	3584	99.6%	4960	97.2%
attribute	230	2878	91.5%	3184	89.5%

After pre-processing the data, we specified the statement formation settings that determine how to link concepts into statements (for detailed information about coding choices in AutoMap and their impact on map analysis results see Diesner & Carley, 2004a). We used entire texts as a coding unit, a window size of four, and rhetorical adjacency for the delete list and the meta-matrix thesaurus. For the generalization thesaurus adjacency does not apply, because we did not choose the thesaurus content only option for it. After setting up and applying the pre-processing and statement formation rules, which together form the coding scheme, we run meta-matrix text analysis on the data set.

4. Text Analysis Results

Statements between meta-matrix entities were formed form eight types of entities or nodes. These entities on average linked into 22 unique concepts per text, raging from 2 to 63, and 56 total statements, ranging from 2 to 403. The number of unique statements considers each statement only once per text, whereas the number of total statements also takes into account repetitions of statements. In total, 12465 statements were identified in our data set. Figure 2 (next page) provides an overview on the distribution of these 12465 edges across the meta-matrix. Maps generated with AutoMap are digraphs in order to adequately represent the inherently directed structure of texts. The entity links into the least number of edges in comparison to all other entities, which is task/event, and the two entities that form most often edges, which are role and attribute, have a higher outdegree than indegree. For all other entities, the indegree is higher than the outdegree.

Once we have understood what sections of the meta-matrix are covered to what extent by our data, we are interested in analyzing particular sections of the meta-matrix in detail. In order to do this, we perform Sub Matrix Text Analysis (Diesner & Carley 2004b). This technique distils sub-network from the meta-matrix. Sub-networks are, for example, membership networks (agent by organization, for a list of sub matrix labels see Diesner & Carley, 2004b) or organizational assignment networks (organization by task/event). Figure 2 displays the social network (agent by agent) from our data. The circular sub-network in the lower right corner that is not connected to entities out of the circle shows the persons who were charged with the Khobar Tower Bombing in Saudi Arabia in 1994. Though the entire graph is rather sparse, there are only 2 isolated agents, which are Hassan Al-Banna and Abdullah Ramazanzadeh. The visualizations were created by outputting the union of all links between the specific meta-matrix entities from all maps in DyNetML format (Tsvetovat, Reminga & Carley, 2004) with AutoMap and visualizing this file with ORA's Visualizer (Carley & Reminga, 2004). With ORA the matrices can be further analyzed (Carley, Diesner, Reminga & Tsvetovat, 2004). We run ORA's Intel report on the social network data. The results of this suggested that the person with the highest Cognitive Demand (0.0915), Degree Centrality (0.0915) and Betweenness Centrality (0.0463) was Mohammad Khatami. Abdallah Al-Jarash has the highest Clique Count (12.0), and Mustafa Al-Qassab had the highest number of Simmelian Ties (0.0282).

Figure 1: Number of statements in meta-matrix cells

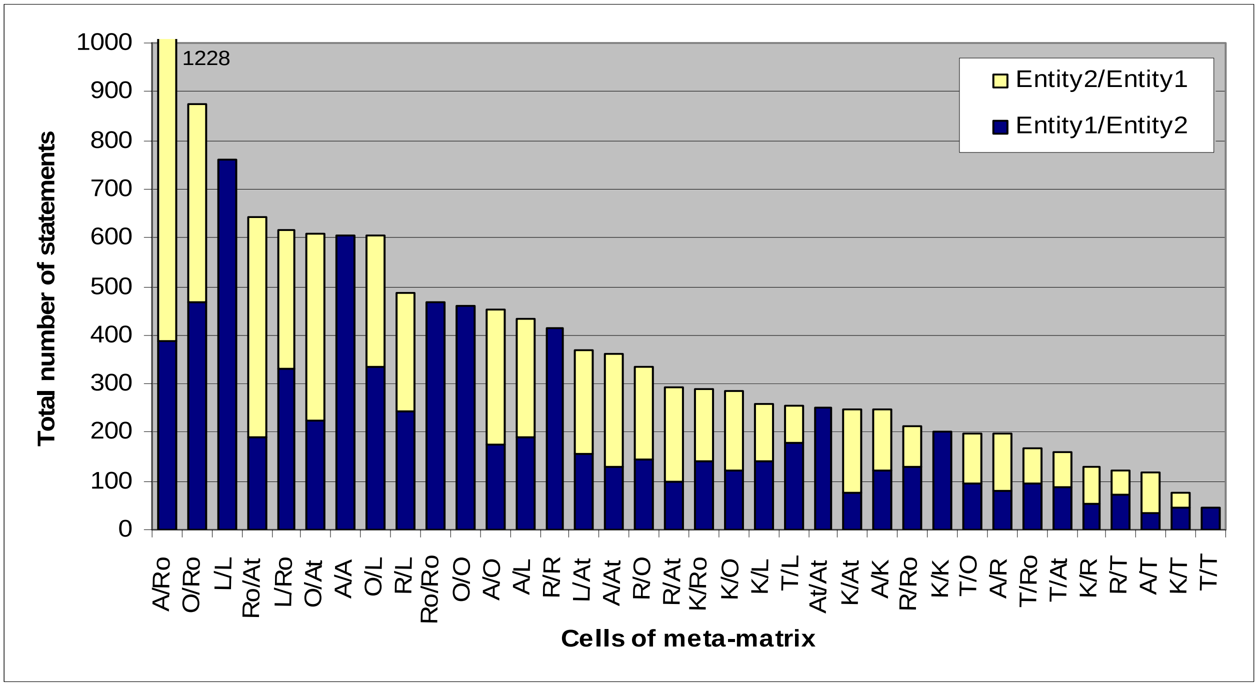
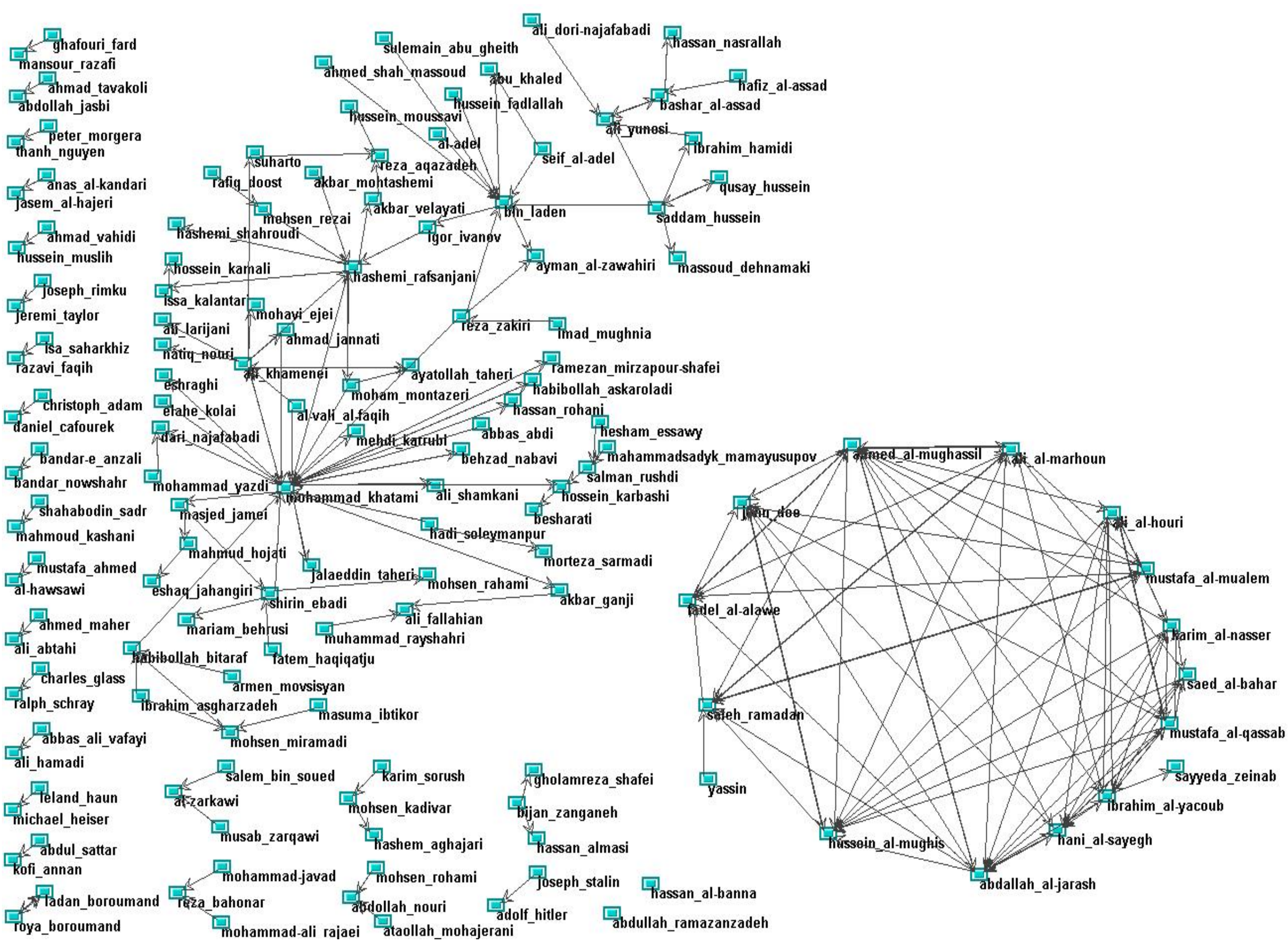


Figure 2: Social network (agent by Agent)



5. Discussion and Conclusion

In the research presented herein we demonstrated how to get from texts to networks that represent covert systems. We have shown that the social and organizational structure of a covert network can be extracted from textual data. The model of the covert network that we detected however needs to be compared to empiric data in future research in order to validate the techniques we applied. Also it needs to be compared to other covert networks in order to enhance interpretability of SNA measures generated by ORA.

The Network Text Analyzes we performed in order to reveal the network's structure and properties is supported by AutoMap in a computer-supported fashion. The phase that required the most manual yet low demanding work was the construction and refinement of the thesauri. In future research, we will look at techniques for further automating this process.

We have started separating the data into various time periods and analyzing the dynamics of the network over time (Carley, Diesner, Reminga & Tsvetovat, 2004). In the future we will report on results of this work.

References

Alexa, M. (1997). *Computer-assisted text analysis methodology in the social sciences*. ZUMA-Arbeitsbericht 97/07.

Bakker, R.R. (1987). *Knowledge Graphs: Representation and Structuring of Scientific Knowledge*. Dissertation. University Twente.

Batagelj, V., Mrvary, A., & Zaveršnik, M. (2002). Network Analysis of Texts. In T. Erjavec, & J. Gros (Eds.), *Proc of the 5th International Multi-Conference Information Society - Language Technologies*. (pp. 143-148). Ljubljana, October 2002. Jezikovne tehnologije / Language Technologies, Eds.. Ljubljana, 2002.

Burkart, M. Thesaurus. In M. Buder, W. Rehfeld, T. Seeger, & D. Strauch (Eds.), *Grundlagen der praktischen Information und Dokumentation: ein Handbuch zur Einführung in die fachliche Informationsarbeit*, (pp. 160 – 179). 4th edition. München: Saur.

Carley, K.M. & Reminga, J. (2004). *ORA: Organization Risk Analyzer*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report.

Carley, K.M. (1988). Formalizing the Social Expert's Knowledge. *Sociological Methods and Research*, 17(2), 165-232.

Carley, K.M. (1997a). Extracting team mental models through textual analysis. *Journal of Organizational Behavior* 18: 533-558.

Carley, K.M. (1997b). Network Text Analysis: the network position of concepts. In Carl W. Roberts (Ed.), *Text analysis for the social sciences*, (pp. 79-102). Mahwah, NJ: Lawrence Erlbaum Associates.

Carley, K.M. (2003). Dynamic Network Analysis. In R. Breiger, K.M. Carley, & P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, (pp. 133-145). Committee on Human Factors, National Research Council, National Research Council.

Carley, K.M. 2002. Smart Agents and Organizations of the Future. In L. Lievrouw & S. Livingstone (Eds.), *The Handbook of New Media:* 206-220. Thousand Oaks, CA: Sage.

Carley, K.M. (1993). Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis. In Marsden P. (Ed.), *Sociological Methodology*, 23: 75-126. Oxford: Blackwell.

Carley, K.M., & Palmquist, M. (1992). Extracting, Representing, and Analyzing Mental Models. *Social Forces*, 70 (3), 601-636.

Carley, K.M., Diesner, J., Reminga, J., Tsvetovat, M. (2004). An Integrated Approach to the Collection and Analysis of Network Data. *In Proc of the (NAACSOS) 2004 Conference*, Pittsburgh, PA

Coffman, T., Greenblatt, S., & Marcus, M. (2004). Graph-based technologies for intelligence analysis. *Communications of the ACM* 47(3): 45-47.

Corman, S.R., Kuhn, T., Mcphee, R.D., & Dooley, K.J. (2002). Studying Complex Discursive Systems: Centering Resonance Analysis of Communication. *Human Communication*, 28(20), 157-206.

Danowski, J. (1982). A network-based content analysis methodology for computer-mediated communication: An illustration with a computer bulletin board. In R. Bostrom (Ed.), *Communication Yearbook*, (pp. 904-925). New Brunswick, NJ: Ransaction Books.

Diesner, J., Carley, K.M. (2004a). *AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-100. URL: http://reports-archive.adm.cs.cmu.edu/anon/isri2004/abstracts/04-100.html.

Diesner, J., Carley, K.M. (2004b - in print). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. In V.K. Naraynan, & D.J. Armstrong (Eds.), *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*. Harrisburg, PA: Idea Group Publishing.

James, P. (1992). Knowledge Graphs. In R.P. van der Riet, & R.A. Meersman (Eds.), *Linguistic Instruments in Knowledge Engineering* (pp. 97-117). Amsterdam: Elsevier.

Johnson, J.C., & Krempel, L. (2004). *Network Visualization: The Bush Team in Reuters News Ticker 9/11-11/15*. Journal of Social Structure, 4(1). URL: http://www.cmu.edu/joss/content/articles/volume5/JohnsonKrempel/

Jurafsky, D., & Marton, J.H. (2000). *Speech and Language Processing*. Upper Saddle River, New Jersey: Prentice Hall.

Klein, H. (1997). Classification of Text Analysis Software. In R. Klar, & O. Opitz (Eds.), *Classification and knowledge organization: Proceedings of the 20th annual conference of the Gesellschaft für Klassifikation e.V.*, (pp. 255-261). University of Freiburg, 1996, Berlin, New York: Springer. 1997.

Kleinnijenhuis, J., De Ridder, J.A., Rietberg, E.M. (1996). Reasoning in Economic Discourse: An application of the Network Approach in Economic Discourse. In C.W. Roberts (Ed.), *Text Analysis for the Social Sciences* (pp. 79-102). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Krackhardt, D., & Carley, K.M. 1998. A PCANS Model of Structure in Organization. In *Proceedings of the 1998 International Symposium on Command and Control Research and Technology Evidence Based Research*: 113-119, Vienna, VA.

Krebs, V.E. (2002). Mapping Networks of Terrorist Cells. *Connections* 24(3): 43-52. URL: http://www.sfu.ca/~insna/Connections-Web/Volume24-3/Valdis.Krebs.L2.pdf

Magnini, B., Negri, M., Prevete, R., & Tanev, H. (2002). A WordNet-based approach to Named Entities Recognition. In *Proceedings of SemaNet'02*: *Building and Using Semantic Networks*, (pp. 38-44). Taipei, Taiwan, August 2002.

Monge, P., & Contractor, N. (2003). *Theories of Communication Networks.* Oxford: University Press.

Popping, R. (2000). *Computer-assisted Text Analysis*. London, Thousand Oaks: Sage Publications.

Popping, R. (2003). Knowledge graphs and Network Text Analysis. *Social Science Information* 42(1): 91-106.

Popping, R., & Roberts, C.W. (1997). Network Approaches in Text Analysis. In R. Klar, & O. Opitz (Eds.), *Classification and Knowledge Organization: Proceedings of the 20th annual conference of the Gesellschaft für Klassifikation* (pp. 381-898). University of Freiburg, Berlin, New York: Springer.

Sowa, J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Tsvetovat, M., Reminga, J., & Carley, K.M. (2004). *DyNetML: Interchange Format for Rich Social Network Data*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, CASOS Technical Report CMU-ISRI-04-105. URL: http://reports-archive.adm.cs.cmu.edu/anon/isri2004/abstracts/04-105.html.