Revealing Social Structure from Texts:
Meta-Matrix Text Analysis as a novel method for Network Text Analysis[1]

Jana Diesner, Kathleen M. Carley
Center for Computational Analysis of Social and Organizational Systems
Institute for Software Research International
School of Computer Science
Carnegie Mellon University

Abstract

Texts can be coded and analyzed as networks of concepts often referred to as maps or semantic networks. In such networks, for many texts, there are elements of social structure – the connections among people, organizations, events, and so on. Within organizational and social network theory an approach called the meta-matrix is used to describe social structure in terms of the network of connections among people, organizations, knowledge, resources, tasks and so on. Herein, we propose a combined approach using the meta-matrix model, as an ontology, to lend a second level of organization to the networks of concepts recovered from texts. We have formalized and operationalized this approach in an automated tool for text analysis referred to as AutoMap. We demonstrate how this approach enables not only meaning but also social structure to be revealed through text analysis. We illustrate this approach by showing how it can be used to discover the social structure of covert networks – the terrorist groups operating in the West Bank.

Keywords: network text analysis, meta-matrix, meta-matrix text analysis, mental models, covert networks, social network analysis, semantic networks

[1] 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.

1. Introduction

Texts are a typical source of information about meaning, organizations, and society. Today, a large and growing number of texts is available in electronic form that describes, discusses, or display information about people, the groups to which they belong, the activities in which they engage, and the resources at their disposal. This data and its accessibility motivate the development and investigation of automated techniques for extracting the underlying social and organizational structure from such texts in an effective and efficient way.

In this paper, we present an automated approach to text analysis that can be used to extract the underlying social and organizational structure contained in texts. This approach is based on the following insights. First, texts can be represented as networks of concepts and the connections between them. These concepts refer to ideas, people, resources, organizations, events, etc... Second, many of the items referred to, such as people, are core entities in the structure of groups and organizations; hence, the extracted networks contain representations of the social structure - the entities and relations among them that comprise a group, organization or society. By classifying the concepts into entity classes used in defining social structures and partitioning the extracted networks into sub-networks, we have effectively used network analysis of texts to reveal the social structure represented in texts.

Herein, we describe this approach in detail and explain how we operationalized, formalized, and implemented this approach into a software called AutoMap that enables the analyst to extract social structure from texts. As part of this work, we have operationalized an ontological scheme (based on the meta-matrix proposed by Carley (2002)) for describing social and organizational structure. This ontology is utilized as part of a hierarchical scheme for cross categorizing concepts. In this paper we furthermore demonstrate how analysts can use AutoMap to automatically extract not just networks of concepts and the relations among them, but also classify the concepts and relations between them according to this ontology. This enables the automatic extraction of views of the social structure.

The chapter begins with a brief overview on the model and methods involved. We then describe how we formalized and implemented the combination of the meta-matrix model and the network text analysis technique. This is followed by a substantive example that we provide in order to illustrate this approach for revealing social structure through the analysis of texts by extracting an image of the social structure of the terrorists groups in the West Bank. We conclude with a discussion of the potentials and limitation of our approach.

Note, this paper should not be viewed as a description of the West Bank terrorist groups. We have coded for this paper only a small sample of texts to illustrate the technique. No conclusions for this group should be drawn from the results reported herein.

2. Using Network Text Analysis to Code Texts

In the area of network text analysis, previous research and development have provided computer-supported solutions that enable analysts to gain a window into social structure and meaning as represented in texts. Collectively these approaches enable the analyst to extract networks of concepts and the connections between them from the texts. These networks are sometimes referred to as maps (Carley, 1997b), networks of centering words (Corman, Kuhn, Mcphee & Dooley, 2002), semantic nets (Reimer, 1997), semantic networks (Monge & Contractor, 2001, 2003; Popping, 2003; Ryan & Bernard, 2000), networks of concepts (Popping, 2000), or networks of words (Danowski, 1993). Herein, we refer to such techniques using the general term – network text analysis (NTA) (Carley, 1997b; Popping, 2000). NTA approaches vary on a number of dimensions such as the level of automation, a focus on verbs or nouns, the level of concept generalization, and so on. Nevertheless, in all cases, networks of relations among concepts are used to reveal the structure of the text, meaning, and the views of the authors. Further, these networks are windows into the structure of the groups, organizations and societies discussed in these texts. This structure is implicit in the connections among people, groups, organizations, resources, knowledge, tasks, events, and places.

NTA is a specific text analysis method that encodes the links between words in a text and constructs a network of the linked words (Popping, 2000). The method is based on the assumption that language and knowledge can be modeled as networks of words and the relations between them (Sowa, 1984). NTA methodologically originates from traditional techniques for indexing the relations between words, syntactic grouping of words, and the hierarchical and non-hierarchical linking of words (Kelle, 1997). The method of NTA enables the extraction, analysis, and concise representation of the complex network structure that can be represented in texts. Furthermore, NTA covers the analytic spectrum of classical content analysis by supporting the analysis of the existence, frequencies, and covariance of words and themes (Alexa, 1997; Popping, 2000). Given these functionalities, computer-supported NTA is one method for analyzing large collections of texts effectively and efficiently. Several NTA methods exist (see bullet item list below, for more details on methods see Popping, 2000; Popping & Roberts, 1997). Many have been applied in empirical settings (see discussion by 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, 1997b; Carley & Palmquist, 1992)
- Network Evaluation (Kleinnijenhuis, Ridder & Rietberg, 1996)
- Word Network Analysis (Danowksi, 1982).

Besides the analysis of textual data, current work also focuses on the visualization of networks extracted from texts (Batagelj, Mrvar & Zaveršnik, 2002).

In this research we concentrate on map analysis. Map analysis systematically extracts and analyzes the links between words in texts in order to model the authors "mental maps" as networks of linked words. Coding texts as maps focuses analysts on investigating the meaning of texts by detecting the relationships between and among words and themes (Alexa, 1997; Carley, 1997a). Maps are a cognitively motivated representation of knowledge (Carley, 1988). In map analysis, a concept is a single idea represented by a single word or a phrase. A statement is two concepts and the relation between them. A map is the network of the statements (Carley, 1997b).

Before continuing, it is worth noting that the terminology in this area is very diverse having come from a variety of disciplines. Thus to orient the reader and help avoid confusion we provide some basic terminology as we will use it herein in Table 1. This will foreshadow the discussion of the procedure we are proposing in this paper.

Table 1. Terminology and associated symbols

Term	**Definition**	**Alternative Terms**	**Examples**
Text	A written work	Sample	Newspaper article, abstract, website, transcribed interview
Concept	Single ideational kernel	Node	Terrorist, terrorism, 9-11, training camp
Text-level concept	Words that appear in text	Word, concept, phrase, named-entity	Abdel Aziz al-Rantisi, terrorism, Palestine, Hamas, captured
Higher-level concept	A word or phrase chosen by the analyst into which other words or phrases are translated	Concept, node	Terrorist, Rantisi
Entity class	Objective category that can be used for classifying concepts; Top level in the ontology	Meta-node, entity, category, concept type, node type	Agent, Organization
Relation	Connection between concepts	Statement, link, tie, edge, connection	Rantisi is in the Hamas
Relation class	Objective category that can be used for classifying relations connecting concepts in entity class a to concepts in entity class b, such that a and b may or may not be distinct.	Relation type, edge type, tie type, sub-network	Social network, is a member of
Map	The network formed by the set of statements in a text.	Network, concept network, semantic net(work), network of concepts or words	See Figures 3 and 4
Meta-matrix	Conceptual organization of concept networks into a set of networks defined by entity classes and relation classes.	Ontology, classification scheme, meta-network	See Tables 2 and 3

3. Using the Meta-Matrix as an Ontology

Since NTA can be used to extract networks of concepts, we can leverage the methods of social network analysis (SNA) to analyze, compare and combine the network of concepts extracted from the texts (see e.g., Scott, 2000; Wasserman & Faust, 1994 for SNA techniques). This provides the analyst with tremendous analytical power (see Hill & Carley, 1999 for illustrative study). If in addition, we cross classify the extracted concepts into an ontology, particularly one designed to capture the core elements of social and organizational structure, we gain the added theoretical power of extracting in a systematic fashion an empirical description of the social and organizational structure. The key would be to design a useful ontology.

Such an ontology is implicit in the meta-matrix approach (Carley 2003, 2002; Krackhardt & Carley 1998) to organizational design. Krackhardt and Carley defined an approach to representing the state of an organizational structure at a particular point in time as the set of entities (people, resources, and tasks) and the relations among them. The meta-matrix approach is a representational framework and a set of derived methods for the computational analysis of multi-dimensional data that represent social and organizational systems. The concept of the meta-matrix originates from the combination of:

1. Information processing and knowledge management (Galbraight, 1977; March & Simon, 1958; Simon, 1973; Carley & Hill, 2001).
2. The PCANS approach (Krackhardt & Carley, 1998), which was later generalized by Carley and Hill to include knowledge, events, and organizations (Carley, 2002; Carley & Hill, 2001).
3. Operations research (Carley & Krackhardt, 1999; Carley, Ren & Krackhardt, 2000).
4. Social network analytic techniques and measures (see, e.g., Scott, 2000; Wasserman & Faust, 1994).

The meta-matrix enables the representation of team or organizational structure in terms of entity classes and relations. In principle, this is an extensible ontology such that new entity classes and new classes of relations can be added as needed. Each entity class represents an ontologically distinct category of concepts (or in the social network language, nodes). Each relation class is a type of link between concepts within entity class 1 and 2. For the sake of illustration we use a simple form of the ontology in which we identify four entity classes – People, Resources (or Knowledge/Skills), Tasks or Events, and Groups or Organizations (see Table 2 headers). We choose these entity classes as they are sufficient for illustration and they are critical for understanding the structure of teams, groups and organizations. The reader should keep in mind that it is possible to use different entity classes and still think in terms of the meta-matrix conceptualization (as indeed we do in this paper). The key aspect for our purposes is that the meta-matrix defines a set of entity classes and a set of relation classes. This facilitates thinking systematically about organizational structure and provides a limited hierarchy for structuring the network of concepts.

Table 2: Original meta-matrix conceptualization

Meta-Matrix entities	**People**	**Knowledge/ Resources**	**Events/ Tasks**	**Groups/ Organizations**
People	Social network	Knowledge Network/ Resource Network	Attendance Network/ Assignment Network	Membership network
Knowledge/Resources		Information Network/ Substitution Network	Needs network	Organizational capability
Events/Tasks			Temporal Ordering/ Task Flow/ Precedence	Institutional support or attack
Organizations				Interorganizational network

Based on Carley, 2002, 2003.

Between any two entity classes there can be one or more classes of relations. For example, between people and people we can think of a number of relations including, but not limited to, communication relations, friendship relations, or money exchange relations. To orient the reader, in Table 2, common labels for the network formed by linking the row and column entity classes are identified. The data in a meta-matrix represents the structure of the group or organization at a particular time. It can be analyzed to locate vulnerabilities, strengths, features of the group, to identify key actors, and to assess potential performance. In summary, the meta-matrix approach allows analysts to model and analyze social systems according to a theoretically and empirically founded schema (Carley, 2003). By employing this approach as an ontology we enable the analyst to extract and analyze social systems as described in texts.

4. Combining NTA and Meta-Matrix Approaches

In texts, the links between words (concepts) are implicit. Hence, extracting a network of concepts from a text and classifying this network via the meta-matrix ontology requires an inference process. The links, or relations, between concepts must be extracted based on the semantic, syntactic, and contextual information given in a text (Carley, 1986; Carley, 1988; Popping, 2003). Making the meta-matrix approach available for NTA can provide analysts with a novel technique for extracting textual networks that reveal the relationships within and between the elements that compose a network and that were classified a priori according to the meta-matrix model. The features of the textual data that are relevant to the analyst can then be represented as a network structure of the meta-matrix entity classes and the connections between these classes. Such a network makes the structure of social systems, which is implicitly contained in texts, visible and analyzable.

How did we combine and formalize the meta-matrix approach and the map analysis technique, which is a specific type of NTA? We utilized the meta-matrix model as an extension of NTA in general and of map analysis in specific by instantiating the following 5 step procedure:

1. Concept Identification – Identify the concepts in texts that are relevant to the analyst's research question. As part of this process the analyst may first want to generalize many text-level concepts into higher-level concepts.
2. Entity Identification – Define an ontology for capturing the overall structure described in the text. We use the basic meta-matrix. However, other analysts may wish to adapt this to their research question. Note, step 1 and 2 can be done in either order.
3. Concept Classification – Classify the identified concepts into the relevant entity classes in the meta-matrix. Given the vagaries of the language it

may be that some concepts need to be cross-classified in two or more entity classes.

4. Perform Map Analysis –Automatically extracting the identified concepts and the relations among them from the specified texts. This results in a map or conceptual network. Since the concepts are classified by entity classes, the resulting concept network is hierarchically embedded in the ontology provided by the meta-matrix. In essence then, there are two networks. First, there is the concept network where the nodes are concepts (many of which are higher-level concepts). Second, there is the entity network where the nodes are the entity classes and the links are the sum of the links from all concepts of entity classes to all concepts of entity classes. Finally, there is the network (embodied in the meta-matrix thesaurus), connecting concepts to entity classes.
5. Graph and Analyze Data – The final step is to take the extracted data for each text, the network, and graph and analyze it in general and by cells in the meta-matrix. As part of this analysis, the resultant networks from different texts can be combined and compared. Note, the analysis can occur at the concept network level (map analysis), the entire meta-matrix level (meta-matrix text analysis), and the sub-cell level (sub-matrix text analysis).

We refer to these five steps as the method of meta-matrix text analysis. With this novel technique we hope to contribute towards the analysis of complex, large-scale data and social systems and providing profound multi-level access to the meaning of textual data. We note that these steps begin to bridge the gap between NTA and a more interpretive analysis of texts. The meaning of concepts is revealed by virtue of what other concepts they are connected to. In the meta-matrix approach, the meaning of concepts is revealed both by what other concepts they are connected to and by what type of entity classes into which they fall.

5. Implementation of Meta-Matrix Text Analysis

We have implemented the formalization of the technique of meta-matrix text analysis in a network text analysis tool called AutoMap (Diesner & Carley, 2004). AutoMap is a software that helps analysts to extract, analyze, represent, and compare mental models from texts. The tool performs computer-supported content analysis, map analysis, meta-matrix text analysis, and sub-matrix text analysis. The latter two types of analysis we discuss in this section. The more classic methods of content analysis and map analysis were previously described in Carley and Palmquist (1992) and Carley (1997a).

Steps 1 to 3 in meta-matrix text analysis may involve a thesaurus. 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. Thesauri are created by reading a set of texts, using pre-defined material, and/ or deriving pairs of concepts and higher-level concepts from theory (Burkart 1997; Kelle 1997, Klein, 1997; Zuell & Alexa 2001). The terminology of a thesaurus depends on the content and the subject of the data set.

Thesauri play a key role in AutoMap coding. AutoMap in performing a classic content analysis or map analysis can utilize a generalization thesaurus. In this thesaurus, the analyst can reclassify words into other words on the basis of shared meaning, spelling errors, aliases, etc.. Further, phrases that refer to a single ideational kernel – such as weapons of mass destruction – can be reclassified as a single concept – WMD. When texts are pre-processed by AutoMap, using a generalization thesaurus, idiosyncratic differences in writing style, multi-word-concepts and wording errors can be eliminated. This generalization process facilitates identifying true conceptual similarities and differences across texts. The creation and application of a generalization thesaurus is step 1 – concept identification – in the coding procedure described in the last section.

When AutoMap is used to perform a meta-matrix text analysis a second type of thesaurus can also be employed. This second thesaurus, the meta-matrix thesaurus, contains the classification of concepts into the entity classes in the meta-matrix. When texts are processed with a meta-matrix thesaurus the organizational structures described in the text can be extracted. Since one concept might be indicative of several meta-matrix entity classes, a meta-matrix thesaurus can consist of more than two columns. For example, the concept military falls into two entity classes – Organization and Resource. The specific entity and relation classes used for the meta-matrix approach in this paper are presented in Table 3.

Table 3: Meta-matrix model formalization used in AutoMap – entity classes and relation classes

Meta-Matrix Entities	**Agent**	**Knowledge**	**Resources**	**Tasks/ Event**	**Organizations**	**Location**
Agent	Social network	Knowledge network	Capabilities network	Assignment network	Membership network	Agent location network
Knowledge		Information network	Training network	Knowledge requirement network	Organizational knowledge network	Knowledge location network
Resources			Resource network	Resource requirement Network	Organizational Capability network	Resource location network
Tasks/ Events				Precedence network	Organizational assignment network	Task/Event location network
Organizations					Inter-organizational network	Organizational location network
Location						Proximity network

Note, in applying the meta-matrix conceptualization to terrorist groups we have extended the original conceptualization (see Table 2) by treating Knowledge and Resource as separate entities (Carley & Reminga, 2004) and by adding Location as a primary entity. Further, we generalized People into Agent to reflect the fact that often names are not known and people are identified by action such as victim_killed. Since this is an extensible ontology, these changes pose no harm to the underlying theory. We did this extension as knowledge, resources, and location are meaningfully unique entities for research in the area of covert networks. By extending the meta-matrix as shown in Table 3, we have done step 2 – entity identification – in the coding procedure described in the last section.

The analyst can use none, either or both types of thesauri – generalization and meta-matrix – to analyze texts with AutoMap. Moreover, the thesauri can be used in either order. Although in general, the analyst may find it useful to first create a generalization thesaurus, then a meta-matrix thesaurus. Building these thesauri can be done iteratively as new texts are added to the available text set as AutoMap minimizes the cost of coding and recoding texts. The larger the corpus of texts being analyzed, the more time is saved.

When using the meta-matrix thesaurus, AutoMap allows the analyst to associate a text-level concepts or higher order concepts from the generalization thesaurus with one, multiple or no entity classes, and to add user-defined entity classes. This process of associating concepts with entity classes is step 3 – concept classification – in the previously identified coding procedure.

When AutoMap applies the meta-matrix thesaurus, it searches the text set for the concepts denoted in the meta-matrix thesaurus and translates matches into the corresponding meta-matrix entity classes as specified by the analyst. When performing meta-matrix text analysis, AutoMap links the meta-matrix entity classes in the texts that were pre-processed with a meta-matrix thesaurus into statements, and builds one concept networks per text that is cross coded in terms of the meta-matrix, thus resulting also in a meta-matrix. This automated network creation is step 4 – perform map analysis – in the previously identified coding procedure.

As noted, the resultant networks can be analyzed at varying levels during step 5 – graph and analyze data. For example, the analyst might be interested in seeing and analyzing the networks of the text-level concepts that represent all or only some of the meta-matrix categories. We implemented this functionality as sub-matrix text analysis. Each cell in Table 3 (see above) denotes a sub-matrix. Sub-matrix text analysis distils one or several sub-networks from the meta-matrix and presents text-level concepts or, if a generalization thesaurus was applied, concepts or higher order in the chosen entity classes. This routine enables a more thorough analysis of particular sections of the meta-matrix, such as Agent by Agent networks (social networks), or Organization by Resource networks (organizational capability networks). When performing sub-matrix text analysis, AutoMap links the concepts representing the meta-matrix entity classes selected by the analyst into networks.

With the implementation of meta-matrix text analysis and sub-matrix text analysis into AutoMap we hope to contribute to the investigation of the network structure of social and organizational systems that are represented in texts. With these techniques we aim to provide a reasonable extension of the base technology of computer-supported network text analysis and a practical implementation of the meta-matrix model. In the next section we demonstrate how these novel techniques can help analysts to detect the meaning and underlying social structure inherent to textual data in order to answer related research questions.

6. Illustrative example of application of Network Text Analysis

To demonstrate the meta-matrix approach to NTA we use a small application data set of 18 texts. Each text will be coded using the proposed approach and the AutoMap software.

6.1. Data

This text sample is a sub-sample drawn from a larger text collection that consists of 191 texts collected at CASOS about six major terrorist groups that operate in the West Bank. These groups are the Al Aksa Martyrs Brigades, Al Fatah, Al Qaeda, Hamas, Hezbollah, and the Islamic Jihad. We gathered the texts from LexisNexis Academia via exact matching Boolean keyword search for each of the groups. The media that we searched with LexisNexis were The Economist, The Washington Post, and The New York Times. The time frame of our data set ranges from articles published in 2000 to 2003. We sorted the retrieved texts by relevance, screened the top most texts, and selected up to three texts per organization and year for our dataset. The sub sample from this corpus that we work with in this paper consists of one text per terror group from each medium from 2003 (Table 4). This sub sample of 18 texts contains 3035 unique concepts and 13141 total concepts. The number of unique concepts considers each concept only once, whereas the number of total concepts also considers repetitions of concepts. The reader should keep in mind that the small size of this data set and the fact that the texts were chosen across groups rather than within is likely to lead to more overall concepts and fewer relations among them. A discussion of Hamas and Yassin may be unlikely to refer to a discussion about Al Qaeda and bin Laden; whereas, it is more likely to refer to Rantisi.

Table 4: Dataset – number of texts that terror group appears in

Source	Aksa	Fatah	Hamas	Hezbollah	Islamic Jihad	al Qaeda
The Washington Post	2	1	2	1	1	2
The New York Times	1	2	3	2	2	1
The Economist	1	2	4	1	2	1
Total	**4**	**5**	**9**	**5**	**5**	**4**

This text set is a suitable illustrative example because the detection of covert networks such as terrorist groups is one application domain for meta-matrix analysis (Carley, Dombrowski, Tsvetovat, Reminga & Kamneva, 2003). Since texts are a widely used source of information about terrorist groups, a technique for pulling networks classified according to the meta-matrix scheme from this type of data is needed. The results of this sample study are neither a valid indication of these terrorist groups nor a formal validation of the method of meta-matrix text analysis, but show what information the analyst can gain from this novel technique.

6.2. Data pre-processing (Concept Identification)

The quality of the map (or network) extracted from the text can be enhanced by pre-processing the data prior to running analysis: Text pre-processing condenses the data to the concepts that capture the features of the texts that are relevant to the analyst. This technique is also the first step in the procedure of performing meta-matrix text analysis (see section 4). In a previous publication we have described text pre-processing strategies and results with AutoMap in detail (Diesner & Carley, 2004). As a first pre-processing technique we applied a delete list customized for this dataset[1]. Deletion removes non-content bearing concepts such as conjunctions and articles from texts (Carley, 1993). This reduces the number of concepts that the analyst needs to consider when forming thesauri. Then we stemmed the texts with the AutoMap stemmer, which is based on the Porter Stemmer (Porter, 1980). Stemming detects inflections and derivations of concepts in order to convert each concept into the related morpheme (Jurafsky & Martin, 2000: 83, 654). Stemming simplifies the process of constructing a generalization thesaurus and can often eliminate spelling errors and typos. Then we used AutoMap's Named-Entity Recognition functionality. Named-Entity Recognition retrieves concepts such as proper names, numerals, and abbreviations contained in a text set (Magnini, Negri, Prevete & Tanev, 2002). This technique helps to index agents, organizations, places, and events and facilitates building the meta-matrix thesaurus. There were 591 named entities in our dataset. This list of named entities was used to:

1. Translate relevant phrases into a unit that will be recognized as a single concept.

    Examples:

    Holi War into Holy_War. The apparent misspelling of Holi results from stemming.

    Golan Height into Golan_Heights.

2. Translate people's names, various versions of their names as they appear in the data set, aliases and synonyms that these people use into the organization that this person is associated with.

    Examples:

    Dr. Abdel Aziz Rantisi and Dr. Rantisi into Aziz_Al-Rantisi, who is a member of Hamas.

    Mahmoud Abba and Abu Mazen into Mahmoud_Abbas, who is a member of the Palestinian Authority.

3. Translate various spellings of a group and synonyms for groups into one unique name of the related group or organization.

    Examples:

    Hizbullah into Hezbollah.

    Islamic Resistance Movement into Hamas.

6.3. Thesaurus creation

The resulting 170 pairs of associations of text-level concepts with higher-level concepts formed the generalization thesaurus. As noted, a generalization thesaurus translates text-level concepts into higher-level concepts. A single higher-level concept typically has multiple text-level entries associated with it in a thesaurus. For example, Imad Falouji (the higher-level concept), a Hamas member, appeared in the text set as Imad Falouji and Mr. Falouji (two related text-level concepts). The more text-level entries associated with a higher-level concept the greater the level of generalization being employed by the analyst.

Since no pre-defined thesaurus was available to us that would have matched terrorism related concepts to meta-matrix entity classes, we built a second generalization thesaurus. After applying this generalization thesaurus, we built and applied a second generalization thesaurus with 50 entries that translates people's names into organizations or more abstract groups with which these people are associated. We used 4 basic guidelines:

1. Members of the six terrorist groups that the data set focuses on into the related terrorist organization.

   Examples:

      Aziz_Al-Rantisi into Hamas.

2. Representatives of the governments of various countries into Country's Name_Government.

   Examples:

      Palestine, into Palestinian_Authority.

      Omar_Sulieman into Egypt_Government

      Mahmoud_Abbas into Palestinian_Authority.

3. People's names into organizations or abstract groups that they belong to.

   Examples:

      Hans_Blix, Kofi_Annan, and Michael_Chandler into UN.

      Hanadi_Jaradat and Saed_Hanani into Suicide_Bomber.

      Haviv_Dodon, Muhammad_Faraj

      Samer_Ufi into Victim Killed.

In doing this, the basic principle we were applying was to retain as specific actors – those who appeared to play primary roles, whereas, secondary actors were reclassified by their role such as victim. Thus, not all names of people that can be associated with a group were translated into the related group. We applied this strategy in order to enable us to retranslate the entity class Agent, to which we assigned these names in the meta-matrix thesaurus that we applied after the second generalization thesaurus, into the names of key players relevant to us in a sub-matrix text analysis that can be run after the meta-matrix text analysis. Names that we decided not to match with an organization are for example Osama bin Laden, Yasser Arafat and Ariel Sharon. This level of maintenance of detail of information always depends on the research question or goal. Our goal e.g., was to detect the network structure of terrorist groups.

After finishing the generalization process[2] we built and employed a meta-matrix thesaurus. In order to support the analyst in matching text-level concepts against meta-matrix categories, AutoMap offers the options to a) load a list of all unique concepts appearing in the text set into the left most column of the meta-matrix thesaurus or b) save a list of a union of all unique concepts on a directory of the analyst's choice. In the next step the analyst has to manually go through this list and to decide whether or not to associate each single concept with meta-matrix categories. Our dataset contained 2083 unique concepts after applying the generalization thesaurus. 303 of these unique concepts we assigned to a single entity class in the meta-matrix, and 23 of them to two entity classes (Table 5, sum of column one). A total of 1780 of the 2083 unique concepts we did not assigned to any meta-matrix entity class, but they were kept as non-categorized concepts. The creation of a meta-matrix thesaurus is step 3, concept classification, in the procedure of performing meta-matrix text analysis (see section 4).

In the next step we applied the meta-matrix thesaurus to the data set[3], and ran a meta-matrix text analysis on the pre-processed text set[4]. This technique forms step 4, perform map analysis, in the procedure of performing meta-matrix text analysis (see section 4).

Table 5: Creation and application of meta-matrix thesaurus (sorted by Frequency)

Category	Cumulated sum of assignment of concepts to the entity classes in the meta-matrix thesaurus	Cumulated sum of appearance of entity classes in texts after application of meta-matrix thesaurus	Cumulated sum of linkage of concepts associated with meta-matrix entity classes into statements
Organization	48	569	434
Location	81	404	404
Agent	54	250	217
Resource	75	261	188
Task-Event	27	168	146
Knowledge	41	134	128

7. Characteristics of the textual networks as meta-matrices (Graph and Analyze Results)

In this section we report the results of the meta-matrix text analysis and sub-matrix test analysis we ran on our sample data set. This task is step 5 in the procedure of performing meta-matrix text analysis (see section 4). The intent in this section is to illustrate the type of results and graphs possible using the proposed meta-matrix approach to NTA, not to present a comprehensive analysis of terrorist networks. In doing this example, we will analyze the: 1) frequencies of the unique and total concepts and statements, 2) frequencies of the unique and total statements that were formed from concepts associated with meta-matrix entity classes, and 3) the distribution of statements formed from meta-matrix entity classes across the data set.

For our analysis we considered the six meta-matrix entity classes in Table 3. Therefore, we have six unique entity level concepts. Considering only concepts that fall into one or more of these categories, we found an average of 99.2 total concepts per text, ranging from 37 to 163. Based on these concepts, on average of 18.9 unique statements (ranging from 8 to 29) and 45.7 (ranging from 12 to 84) statements were formed per text. Thus, on average, each unique statement appeared 2.4 times per text. Theoretically, each text could contain up to 36 unique statements. The theoretic maximum would be achieved if there existed at least one concept associated with each entity, and at least one concept of each entity formed a statement with at least one concept in each other entity class. The multiple occurrence of unique statements is expressed in the number of total statements.

Across the 18 meta-matrices extracted from our sample texts, 822 total statements were formed within and between the cells of the meta-matrix (see Table 6 for distribution of total statements across meta-matrix). Notice that the upper and lower triangle of the meta-matrix in Table 6 are not symmetric. For example, in Table 6 from Resource (row) to Organization (column) there are a total of 23 statements but from Organization (row) to Resource (column) there are a total of 35 statements. Indeed, there is no need for symmetry as the relations between concepts (edges between nodes) found with AutoMap are directed, which is inherently pre-defined by the directed structure of language. The results in Table 6 show that concepts associated with each meta-matrix entity class appear approximately as often in posterior positions of statements (last row in Table 6) as in anterior positions (last column in Table 6). Thus, the indegree or receptivity of a meta-matrix entity class approximately equals the outdegree or expansiveness of the class. This is due, in part, to the use of proximity in the text to place links among concepts and reflects, if anything, the lack of overly stylized sentential form.

Table 6: Number of links (total number of statements) between meta-matrix categories

Meta-Matrix	Agent	Know-ledge	Resource	Task-Event	Organi-zation	Location	Sum
Agent	24	8	8	12	55	12	**119**
Knowledge	10	18	9	3	20	11	**71**
Resource	8	9	39	11	23	20	**110**
Task-Event	13	7	9	10	20	17	**76**
Organization	58	23	35	19	90	44	**269**
Location	9	10	17	25	47	69	**177**
Sum	**122**	**75**	**117**	**80**	**255**	**173**	**822**

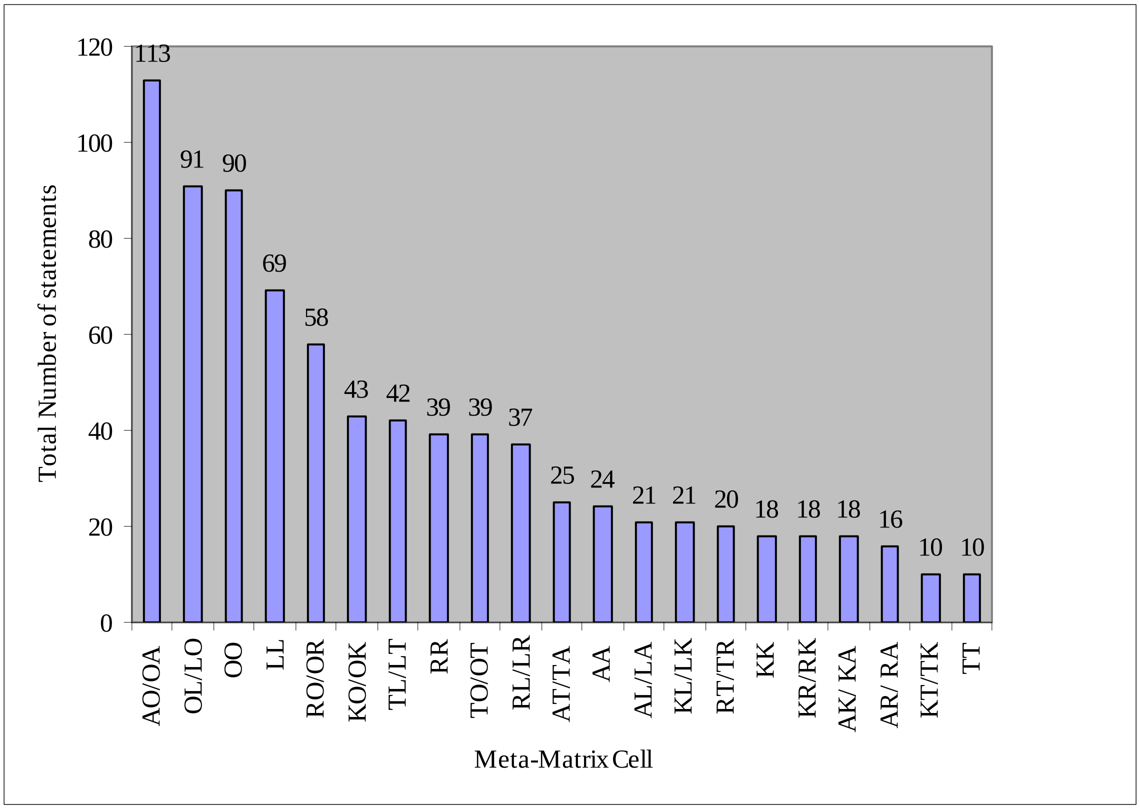
Within the meta-matrix, the entity class that linked most frequently to other entity classes is Organization (179 links), followed by Location (108), Agent (95), Resource (71), Task-Event (66), and Knowledge (53). If we do not look at these absolute values, but at percentages of the linkage of meta-matrix entity classes to the same or other entity classes, our results reveal that concepts in the entity class Task-Event are more likely to be connected to concepts in classes other than Task-Event. In contrast to Task-Event, concepts in the entity class Location are most likely to link to other Location concepts (Table 7).

Table 7: Linkage of meta-matrix entity classes

Meta-Matrix entity class	With same entity class (%)	With other entity classes (%)
Task-Event	13	87
Agent	20	80
Knowledge	25	75
Organization	33	67
Resource	35	65
Location	39	61

Furthermore, the results indicate that within the networks that we extracted from the texts most information refers to membership networks (13.8% of all statements, Figure 1). Although, there is also substantial information on inter-organizational networks (11.1%), and organizational location networks (10.4%). The least information is provided on precedence networks (1.2%) and knowledge requirement networks (1.2%). This suggests that more is known, or at least presented in the news, about who the terrorists are and where they are, than about what they do when and what they need to know in order to engage in such actions or why.

Figure 1: Total number of links between meta-matrix categories



The analysis of the distribution of statements formed from meta-matrix entity classes across the text set reveals that all entities are covered in at least one third of the texts. In addition, Organization, Location, and Agent classes appear in more than half of the texts (Table 8). Again, this suggests that more is reported about who and where than about what, how and why. We note, that a human reading of these texts may pick up a little more about what and how, although, such information does appear to be less common in general in the texts used for this purely illustrative analysis.

Table 8: Number of texts in that links appears

Meta-Matrix	Agent	Knowled	Resource	Task-	Organiza	Location	Sum

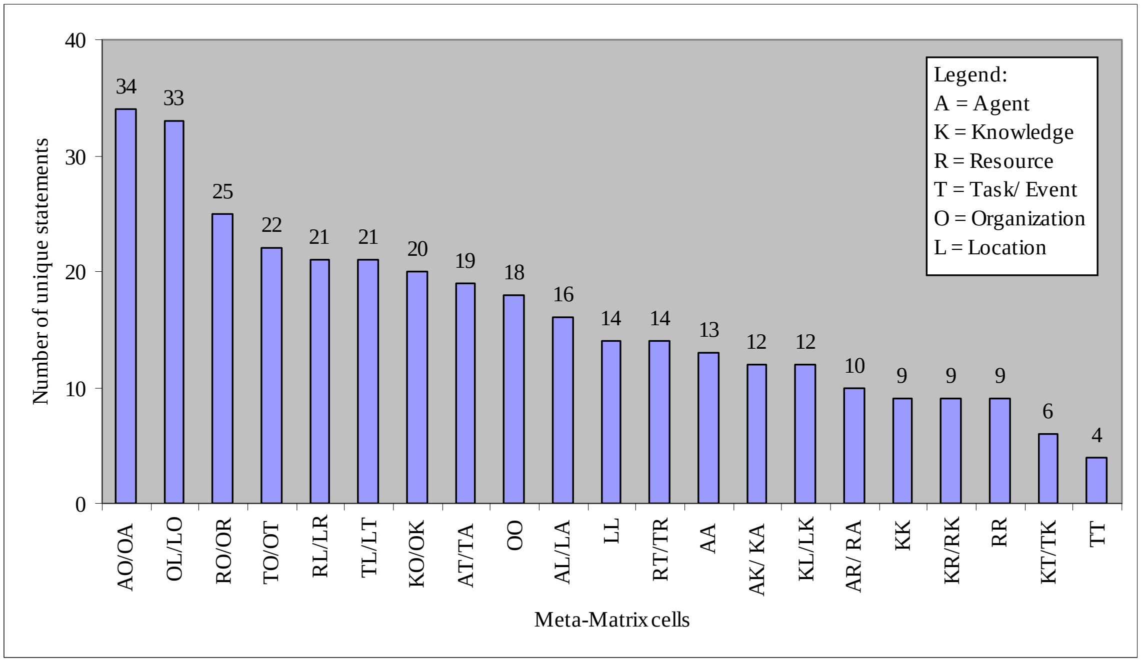
		ge		Event	tion		
Agent	13	5	6	10	17	9	**10.0**
Knowledge	7	9	5	3	9	5	**6.3**
Resource	4	4	9	7	12	11	**7.8**
Task-Event	9	3	7	4	11	10	**7.3**
Organization	17	11	13	11	18	16	**14.3**
Location	7	7	10	11	17	14	**11.0**
Sum	**9.5**	**6.5**	**8.3**	**7.7**	**14.0**	**10.8**	**9.5**

In Figure 1 and Tables 6 and 8 we have been discussing the total links or statements. Looking at the total links provides information about the overall structure of the networks and the elements of the structure (agents, knowledge, etc.) that are considered critical by the authors or for which they have a wealth of information. It is often useful to ask about unique links, however, if we want to understand the structure itself. In Figure 2, we display the number of links per sub-matrix that are unique. That is, a link or statement is only counted once per text regardless of how often it occurs in one text.

Comparison of Figures 1 and 2 shows that a great deal of information – particularly in the Agent to Agent sub-matrix is repeated across texts. This suggests that either, many of the texts were discussing the same information (repetition), or they got their information from the same source. Note, if we knew that each source was unique, then the difference between the total (Figure 1) and the unique (Figure 2) would be an indicator of the reliability of the information.

The overall structure for this covert network is very sparse. In some sense, based on these texts, more is known about the Agents and Organizations affiliations, locations, resources, and knowledge than is known about with the interrelations of knowledge, resources and tasks. It is interesting to note which sub-matrices have more unique links than texts – e.g., the Agent by Knowledge and the Organization by Knowledge sub-matrices. This indicates, that texts that discuss the knowledge network tend to do so by discussing multiple linkages; e.g., all of these people know item z. Whereas, texts that discuss, e.g., the social network (Agent by Agent) are more likely to simply talk about a single pair of actors and the nature of their relationship. Whether this pattern of reporting would hold in other cultures is debatable.

Figure 2: Number of unique links between meta-matrix categories



Beyond learning about the network structure of the meta-matrices and the distribution of concepts and connections between them across the sample data, analysts might be interested in investigating in more detail the concepts and links contained in the meta-matrix. In order to gain this knowledge, sub-matrix text analysis[5] can be run. For illustrating the results of this procedure, we show a map from the same text in Tables 9 to 11. A map contains one coded statement per line and its frequency.

Table 9: Who has what means? Organizational capability network (organization by resource)

Statements formed from Higher-Level Concepts (Sub-Matrix Analysis)	
Sample text 1:	**Sample text 2:**
1 Al-Qaeda - training camp	1 Al-Aksa - assets
1 network- Hawala	1 Al-Aksa - money
1 Hawala - money	1 Hamas - sponsoring
1 finance - network	1 aid - Hamas
1 camp - US-Government	1 aid - Treasury Department
	1 money - Hamas
	1 support - Hamas
	1 Treasury - assistance
	1 US-Government - assistance
	1 assets - Treasury Department

Table 10: Who knows what? Knowledge network (agent by knowledge)

Statements formed from Higher-Level Concepts (Sub-Matrix Analysis)

Sample text 1:	Sample text 2:
1 chairman - monitoring	1 FBI - Analyst
1 evidence - Saddam Hussein	

Table 12: Who is located where and does what? (Localized assignment network: agent by task-event by location)

Statements formed from Higher-Level Concepts (Sub-Matrix Analysis)	
Sample text 1:	**Sample text 2:**
1 Saddam Hussein - Iraq	1 arrest - Leader
	1 Leader - Germany

These various sub-matrix networks enable a more clear understanding of what attributes of the meta-matrix link to other attributes, and with what strength. All three sub-matrices together enable a broader view on the situation. Figures 3 and 4 illustrates this is broader picture. Comparison of figures 3 and 4 illustrates that text 1 presents a more disconnected story than does text 2. Further, even were the two stories combined; the overall map would tell us little about the structure of the two terrorist groups – Al Qaeda and Hamas.

Figure 3: Visualization of sub-matrices from sample text 1

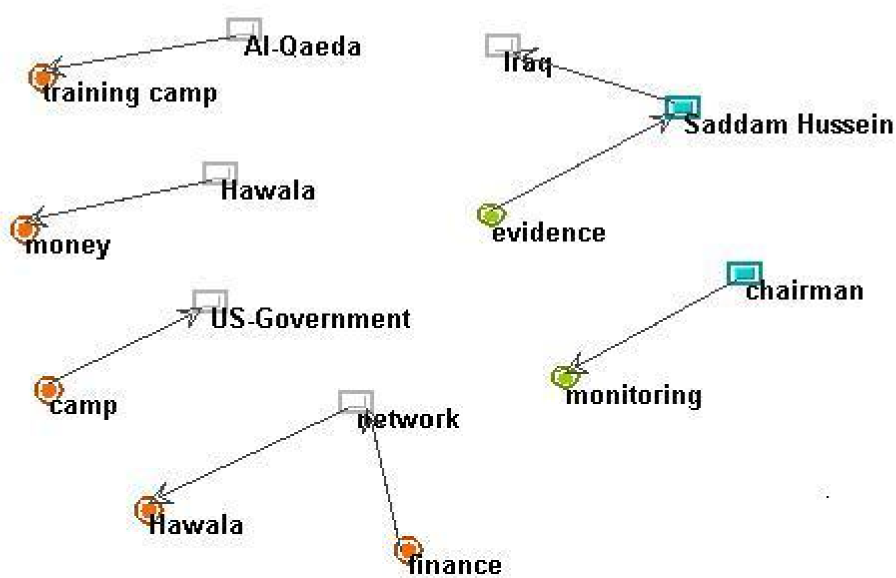
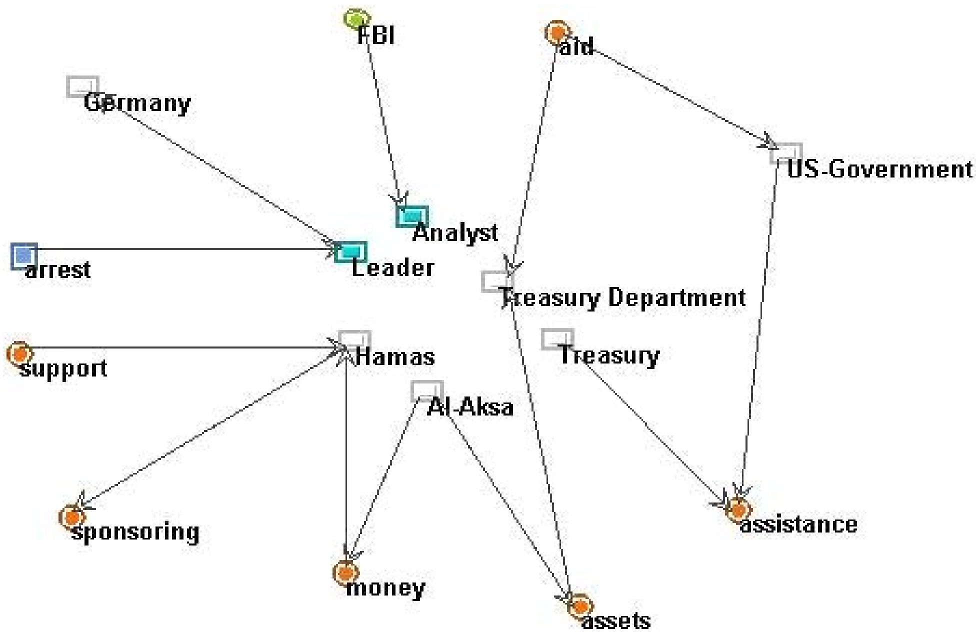


Figure 4: Visualization of sub-matrices from sample text 2



Meta-matrix data and sub-matrix data generated with AutoMap can be saved and then re-analyzed outside of AutoMap using standard social network analysis tools. AutoMap can code these networks and then output them in two useful exchange formats for use with other network analysis tools – DL for UCINET and DyNetML for ORA (www.casos.cs.cmu.edu/projects/ORA, Carley & Reminga, 2004). For this paper, we used ORA as it enables the analysis of all the cells in the meta-matrix at once. In either case, the combination of text and network analysis enables the analyst to readily combine rich textual data with organizational data connected through other methods; thus, enhancing the analysis process.

8. Discussion – Features and Limitations

The techniques of meta-matrix text analysis and sub-matrix text analysis described herein can support analysts in investigating the network structure of social and organizational systems that are represented in textual data. Furthermore, these novel and integrative methods enable analysts to classify words in texts into entity classes (node types) associated with networks common to organizational structures according to a theoretically and empirically validated ontology – the meta-matrix.

The validity of the method and the results presented in this paper are constrained by the little experience we gained so far with these novel techniques, the small number of texts analyzed, and the implementation of the techniques into one software. The tool should also be applied to multiple and larger data sets.

8.1. Lessons learned

In general we find that the Named Entity Recognizer greatly enhances the ability to locate concepts associated with the meta-matrix ontology. In particular, it facilitates locating Agents, Organizations, and Locations. For entity classes that are less associated with proper nouns, the Named Entity Recognizer is of less value.

Clearly, coding texts using AutoMap is not a completely automated process. However, AutoMap does provide a high degree of automation that assists the user and increases the efficiency and effectiveness of meta-matrix text analysis in comparison to manual coding. As with most text analysis techniques that seek to extract meaning, significant manual effort needs to be expended on constructing the delete list and thesauri, even though these methods are computer supported. For example, the delete list used in this study took 30 minutes to construct; however, the thesauri (and there are 3) took 4 days to construct. Thesauri enable the minimization of miscoding, as in missed relations, due to aliases and misspellings, and differences due to the underlying languages. Analysts have to decide on an optimal trade-off between speed of the computer-supported research process and enhancement of the quality of automated coding caused by the manual creation and refinement of pre-processing tools according to their goals and resources.

It is worth noting that significant improvement over straight manual coding can be achieved by building thesauri and delete lists based on only a fraction of texts. As more texts in this domain are coded, we will have to expend relatively little additional effort to expand the delete list and thesauri. For example, we suspect that hundreds of additional texts will be codable with maybe only a day more attention to the thesauri. The reason is that, when in the same domain, construction of thesauri is like building a sample via the snowball method; i.e., with each iteration fewer and fewer novel concepts are found. How large that fraction should be is a point for future work. However, preliminary studies suggest 10% is probably sufficient. Future work should explore whether intelligent data mining and machine learning techniques can be combined with social network analysis and text analysis to provide a more automated approach to constructing thesauri on the fly.

We also find that the higher the level of generalization used in the generalization thesaurus the greater the ability to compare two diverse texts. Not counting typographical errors, often the translation of 2 to 10 text-level concepts per high-level concept seems sufficient to generate a "language" for the domain being studied.

We note that when forming thesauri, it is often critical to keep track of why certain types of concepts are generalized in to others. At the moment there is no way to keep that rationalization within AutoMap. In general, the user should keep a lab-note book or read-me file for keeping such rationalizations.

Finally, we note that for extracting social or organizational structure from texts a large corpus is needed. The point here is comprehensiveness not necessarily specific number of texts. Thus, one might use the entire contents of a book that describes and discusses an organization or a large set of newspaper articles. In building this corpus, not all texts have to be of the same type. Thus, the analyst can combine newspaper reports, books, board-of-directors reports, web-pages etc. Once the networks are extracted via AutoMap they can be combined into a comprehensive description of the organization being examined. Further, the analyst needs to pre-define what the basic criteria are for including a text in the corpus – e.g., it might be publication venue, time frame, geographic area, specific people, organizations or locations mentioned.

8.2. Considerations for future work

We also note that, the higher the level of generalization the more ideas are being inferred from, rather than extracted from, the texts. Research needs to be done on the appropriate levels of generalization. Note, that the level of generalization can be measures as the average number of text-level concepts associated with each higher level concept.

One of the strengths of NTA is that the networks extracted from the texts can be combined in a set theoretic fashion. So we can talk about the network formed by the union or intersection of the set of networks drawn from the set of text. When combining these networks we can, for each statement, track the number of texts that contained that statement. Since a statement is a relation connecting two concepts – this approach effectively provides a weight for that relation. Alternatively, the analyst can compute whether any text contained that statement. In this case, there are no weights and the links in the network are simply present or not (binary). If these texts represent diverse sources of information then the weights are indicative of the certainty or verifiability of a relation. Future work might also explore utilizing Bayesian learning techniques for estimating the overall confidence in a relation rather than just summing up the number of texts in which the statement was present.

We also note that when people read texts there is a process of automatic inference. For example, when people read about a child talking to a parent they infer based on social experience that the child is younger. Similarly, it appears that such inferences are common between the entity classes. For example, if Agent x has Resource Y and Knowledge K is needed to use Resource Y then in general Agent X will have Knowledge K. Future work needs to investigate whether a simple inference engine at the entity class level would facilitate coding. We note that previous work found that using expert systems to assist coding in terms of adding general social knowledge was quite effective (Carley, 1988). Thus, we expect this to be a promising avenue for future research.

Finally, we note that the use of an ontology adds a hierarchical level to the coding. This is invaluable from an interpretative perspective. There is no reason, conceptually, why multiple hierarchical levels could not be added – denoting finer and finer levels of detail. We suspect however, based on the use of hierarchical coding schemes in various scientific fields (e.g., biology and organization theory) that a) such hierarchies are likely to not be infinitely deep, b) a certain level of theoretical maturity and consensus in a field is needed for such a hierarchy to be generally useful, and c) eventually we will need to move beyond such a "flat" scheme for extracting meaning. As to this last point, by flat what we are referring to is the fact that a hierarchy can be completely represented in two dimensions. We found, even when doing this limited coding that some text-level concepts and higher-level concepts needed to be cross-classified into two or more entity classes. As more levels are added in an ontological hierarchy, such cross classification is likely to occur at each level resulting in a network of inference not a simple hierarchy and so a non-flat structure. Future work should examine how to code, represent, and reason about such networks.

9. Conclusions

One of the key advantages of classic content analysis was that macro social change could be tracked by changes in content and by over or under-representation of various words. For example, movements toward war might be signaled by an increasing usage of words describing hostile acts, foreign powers, and weapons. One of the key advantages of Network Text Analysis (NTA) over standard text analysis is that it enables the extraction of meaning and enables interpretation by signaling not just what words are used but how they are used. This enables differences and similarities in viewpoints to be examined; and it enables the tracking of micro social change as evidenced by changes in meaning. Differences and similarities in viewpoints about a meta-structure described or discussed in the text can be examined by adding an ontology to NTA.

In this paper, we used the meta-matrix ontology as we were interested in the underlying social/organizational structure described in the texts. Several points are critical to note. First, the mere fact that we used an ontology to define a set of meta-concepts enables the extraction of a hierarchy of meaning thus affording the analyst with greater interpretive ability. Second, any ontology could be used, and the analysts need to consider the appropriate ontology for their work. In creating this ontology the analyst wants to think in terms of the set of entity classes and the relations among them that define the second level network of interest. For us, these entity classes and relations were those relevant to defining the organizational structure of a group.

The proposed meta-matrix approach to text analysis makes it possible to track more micro social change in terms of changes not just in meaning, but in the social and organizational structures. Using techniques such as this facilitates a more systematic analysis of groups, broadens the types of questions that can be effectively answered using texts, and brings the richness of textual information to bear in defining and understanding the structure of the organizations and society in which we live.

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.), *Proceedings of the 5th International Multi-Conference Information Society - Language Technologies:* 143-148. Ljubljana, October 2002. Jezikovne tehnologije / Language Technologies, Ljubljana.

Burkart, M. 1997. 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*: 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 CMU-ISRI-04-101.

Carley, K.M. 2003. Dynamic Network Analysis. In R. Breiger, K.M. Carley & P. Pattison (Eds.), *Summary of the NRC workshop on social network modeling and analysis*: 133-145. Committee on Human Factors, 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. 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 C.W. Roberts (Ed.), Text *Analysis for the Social Sciences:* 79-102. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Carley, K.M. 1986. An Approach for Relating Social Structure to Cognitive Structure. *Journal of Mathematical Sociology*, 12: 137-189.

Carley, K.M., Dombrowski, M., Tsvetovat, M., Reminga, J., & Kamneva, N. 2003. Destabilizing Dynamic Covert Networks. In *Proceedings of the 8th International Command and Control Research and Technology Symposium.* Washington, D.C. Evidence Based Research. Vienna, V.A.

Carley, K. M., & Hill, V. 2001. Structural Change and Learning Within Organizations. In A. Lomi & E.R. Larsen (Eds.), *Dynamics of Organizations: Computational Modeling and Organizational Theories*: 63-92. Live Oak, CA: MIT Press/AAAI Press.

Carley, K.M., & Krackhardt, D. 1999. A Typology for C2 Measures. In *Proceedings of the 1999 International Symposium on Command and Control Research and Technology.* Newport, RI, June, 1999.

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

Carley, K.M., & Reminga, J. 2004. *ORA: Organizational Risk Analyzer*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-106.

Carley, K.M, Ren, Y., & Krackhardt, D. 2000. Measuring and Modeling Change in C3I Architectures. In *Proceedings of the 2000 Command and Control Research and Technology Symposium*. Naval Postgraduate School, Monterrey, CA, June, 2000.

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. 1993. Network Analysis of Message Content. In W.D. Richards & G.A. Barnett (Eds.), *Progress in Communication Science, XII:* 197-222. Norwood, NJ: Ablex Publishing.

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, 6:* 904-925. New Brunswick, NJ: Transaction Books.

Diesner, J., & Carley, K.M. 2004. *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

Galbraith, J. 1977. *Organizational Design*. Reading, MA: Addison-Wesley.

Hill, V., & Carley, K.M. 1999. An Approach to Identifying Consensus in a Subfield: The Case of Organizational Culture. *Poetics*, 27: 1-30.

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

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

Kelle, U. 1997. Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data. *Sociological Research Online*, 2 (2). URL: http://www.socresonline.org.uk/2/2/1.html

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.*: 255-261, University of Freiburg, Berlin, New York: Springer.

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:* 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.

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:* 38-44. Taipei, Taiwan, August 2002.

March, J.G., & Simon, H.A. 1958. *Organizations*. New York: Wiley.

Monge, P.R., & Contractor, N.S. 2003. *Theories of Communication Networks.* Oxford University Press.

Monge, P.R., & Contractor, N.S. 2001. Emergence of Communication Networks. In F.M. Jablin, & L.L. Putnam (Eds.), *The new Handbook of Organizational Communication: Advances in Theory, Research and Methods*: 440-502. Thousand Oaks, CA: Sage.

Popping, R. 2003. Knowledge graphs and network text analysis. *Social Science Information* 42 (1): 91-106.

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

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 e.V.*: 381-898, University of Freiburg, Berlin, New York: Springer.

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

Reimer, U. (1997). Neue Formen der Wissensrepräsentation. In M. Buder, W. Rehfeld, T. Seeger & D. Strauch (Eds.), *Grundlagen der praktischen Information und Dokumentation: Ein Handbuch zur Einführung in die fachliche Informationsarbeit*: 180 – 207. 4th edition. München: Saur.

Ryan, G.W., & H.R Bernard. 2000. Data Management and Analysis Methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research*: 769-802. 2nd ed., Thousand Oaks, CA: Sage Publications.

Simon, H.A. 1973. Applying Information Technology to Organizational Design. *Public Administration Review*, 33: 268-78.

Scott, J.P. 2000. *Social Network Analysis: A Handbook*. 2nd edition, Sage Publications, London.

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

Wasserman, S., & Faust, K. 1994. *Social Network Analysis. Methods and Applications*. Cambridge: Cambridge University Press.

Zuell, C., & Alexa, M. 2001. Automatisches Codieren von Textdaten. Ein Ueberblick ueber neue Entwicklungen. In W. Wirth & E. Lauf (Eds.), *Inhaltsanalyse – Perspektiven, Probleme, Potenziale:* 303-317. Koeln: Herbert von Halem.

Appendix

Software:

AutoMap: Diesner, J. & Carley, K.M. (2004). *AutoMap1.2: Software for Network Text Analysis.*

AutoMap is a network text analysis tool that extracts, analyzes, represents, and compares mental models from texts. The software package performs map analysis, meta-matrix text analysis, and sub-matrix text analysis. As an input AutoMap takes raw, free flowing, and unmarked texts with ASCII characters. When performing analysis, AutoMap encodes the links between concepts in a text and builds a network of the linked concepts. As an output AutoMap generates representations of the extracted mental models as a map file and a stat file per text, various term distribution lists and matrices in comma separated value (csv) format, and outputs in DL format for UCINET and DyNetML format. The scope of functionalities and outputs supported by AutoMap enables one way of analyzing complex, large-scale systems and provide multi-level access to the meaning of textual data.

Limitations: Coding in AutoMap is computer-assisted. Computer-assisted coding means that the machine applies a set of coding rules that were defined by a human (Ryan & Bernard, 2000: 786; Kelle, 1997: 6; Klein, 1997: 256). Coding rules in AutoMap imply text pre-processing. Text pre-processing condenses the data to the concepts that capture the features of the texts that are relevant to the user. Pre-processing techniques provided in AutoMap are Named-Entity Recognition, Stemming, Deletion, and Thesaurus application. The creation of delete lists and thesauri requires some manual effort (see Discussion section for details).

Hardware and software requirements: AutoMap1.2 has been implemented in Java 1.4. The system has been validated for Windows. The installer for AutoMap1.2 for Windows and a help file that includes examples of all AutoMap1.2 functionalities are available online under http://www.casos.cs.cmu.edu/projects/automap/software.html at no charge. More information about AutoMap, such as publications, sponsors, and contact information is provided under http://www.casos.cs.cmu.edu/projects/automap/index.html.

AutoMap has been written such that the only limit on the number of texts that can be analyzed, the number of concepts that can be extracted, etc. are determined by the processing power and storage space of the user's machine.

Acknowledgement

We want to thank Maksim Tsvetovat and Jeffrey Reminga from CASOS, CMU for helping with generating the visualizations.

---

[1] The delete list was applied with the rhetorical adjacency option. Rhetorical adjacency means that text-level concepts matching entries in the delete list are replaced by imaginary placeholders. Those place holders ensure that only concepts, which occurred within a window before pre-processing, can form statements (Diesner & Carley, 2004).

[2] We did not choose the Thesaurus content only option. Thus, adjacency does not apply.

[3] We used the thesaurus content only option in combination with the rhetorical adjacency. Thus, the meta-matrix categories are the unique concepts.
[4] We used the following statement formation settings: Directionality: uni-directional, Window Size: 4, Text Unit: Text (for detailed information about analysis settings in AutoMap see Diesner & Carley, 2004).
[5] Sub-Matrix selection was performed with the rhetorical adjacency option.