An Integrated Approach to the Collection and Analysis of Network Data*

Kathleen M. Carley, Jana Diesner, Jeffrey Reminga, Maksim Tsvetovat

Kathleen M. Carley, Jana Diesner, Jeffrey Reminga, Maksim Tsvetovat
Carnegie Mellon University
kathleen.carley@cmu.edu
diesner@cs.cmu.edu
jreminga@cs.cmu.edu
maksim@cs.cmu.edu

Abstract

To facilitate the analysis of real and simulated data on groups, organizations and societies tools and measures are needed that can handle relational or network data that is multi-mode, multi-link, and multi-time period, in which both nodes and edges have attributes and there are possible errors in the data. The integrated CMU dynamic network analysis toolset described in this paper enables the coding, analysis, and visualization of such data. Herein we present these tools and illustrate there interoperability and capabilities using data collected from a series of 247 texts on an organizational system in the Mideast.

Contact:
Kathleen M. Carley, Jana Diesner, Jeffrey Reminga, Maksim Tsvetovat
Center for Computational Analysis of Social and Organizational Systems (CASOS)
Institute for Software Research International
Carnegie Mellon University
School of Computer Science
5000 Forbes Avenue
Pittsburgh, PA, 15213

Tel: 1-412-268-6016 (Carley)
1-412-268-5866 (Diesner, Reminga, Tsvetovat)
Fax: 1-412-268-1744
Email: kathleen.carley@cmu.edu
diesner@cs.cmu.edu
jreminga@cs.cmu.edu
maksim@cs.cmu.edu

Key Words: interoperability of network analysis software, meta-matrix, network text analysis, AutoMap, DyNetML

* This work was supported in part by the Office of Naval Research (ONR), United States Navy Grant No. N00014-02-10973 on Dynamic Network Analysis, Grant No. N00014-97-1-0037 on Adaptive Architecture, the DOD, the NSF MKIDS program, and the NSF IGERT program in CASOS. Additional support was provided by CASOS and ISRI at Carnegie Mellon University. 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 Office of Naval Research, the DOD, the National Science Foundation, or the U.S. government. We thank Dan Wood for his help with pre-processing the data.

1 Introduction

Herein we present a novel approach for the interoperability of a series of scalable tools that facilitate the dynamic extraction, analysis, visualization, and reasoning about empiric and simulated network data. The integrated CMU dynamic network analysis toolset described in this paper takes into account multi-mode, multi-link, and multi-time period data including attributes of nodes and edges. This toolset contains the following tools:

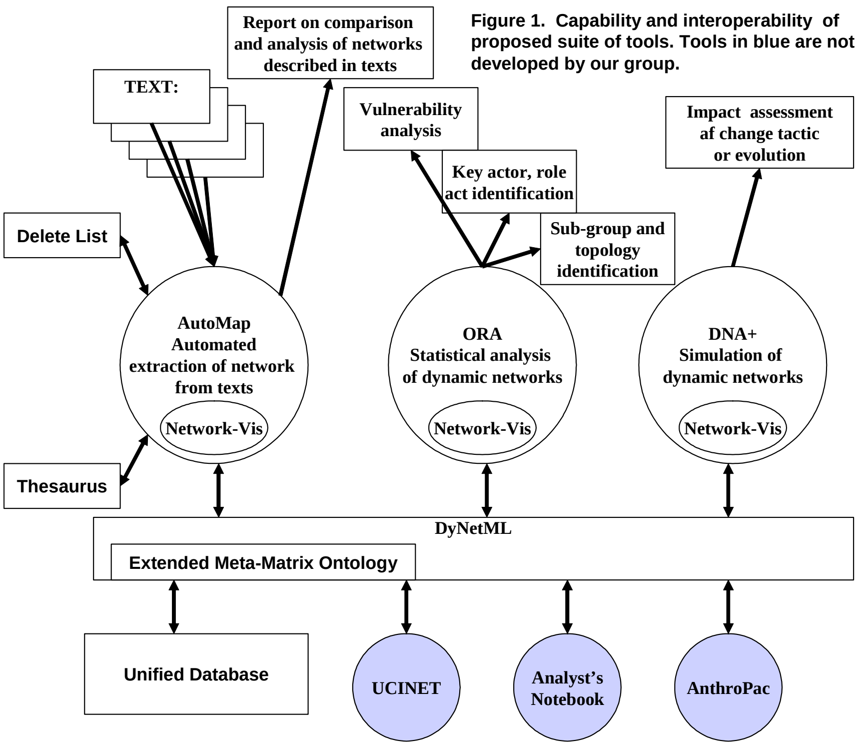


Figure 1. Capability and interoperability of proposed suite of tools. Tools in blue are not developed by our group.

- AutoMap, a semi-automated tool for extracting network data from texts (CMU: http://casos.cs.cmu.edu/projects/automap/, Diesner & Carley, 2004a)
- Network-Vis, a network visualization tool (CMU: used in ORA)
- ORA, a statistical toolkit for analyzing dynamic network data composed of multiple entities (CMU: http://www.casos.cs.cmu.edu/projects/ora, Carley & Kamneva 2004, Carley & Reminga 2004)
- DyNetML, an XML based interchange language for relational data (CMU: http://www.casos.cs.cmu.edu/projects/dynetml, Tsvetovat, Reminga & Carley, 2004). By using DyNetML as an unified interchange language other tools such as UCINET (Borgatti, Everett & Freeman 2002) can be linked in and data can be easily exchanged.
- Simulation packages such as Construct dynamically assess network change. The tools are built to degrade gracefully in the face of missing information and to take advantage of multi-mode multi-link data. We are currently applying these tools to the extraction and analysis of covert networks, groups in email archives, and group plans in military planning exercises. To illustrate the integrated suite and it's potential power, we will draw on one of our covert network studies.

2 Approach

The goal in developing this interoperable suite of tools is to enable faster data collection, facilitate analysis, and develop theory. Several principles guide the approach. First, the tools should be interoperable in that all tools should be capable of using (reading/writing) the same set of data. We expect to move over time to interoperability in the form of analytical results from one tool that can be used as input to other tools in terms of not just of data, but also of summary statistics. Second, data can be collected in many ways but stored in a common format. Third, it should be possible to link to other tools. Fourth, the tool set should scale to large data sets and be robust in the face of missing data. Fifth, the approach should be expandable as new entity types and relations need to be considered (e.g., by adding roles, locations and events to the meta-matrix). Sixth, attributes of nodes and relations should also be captured and analyzed. Illustrative attributes for various entities are shown in Table 1. We note that in addition, relations have attributes such as the relation of actor to group (is_a_member_of) might have the attributes 'date of joining' and 'date of leaving'.

Table 1: Illustrative Attributes of Select Entities

Entities	**Actor**	**Knowledge**	**Resource**	**Task/ Event**	**Group/ Organization**
Attribute	age	scientific	economic	organizational	corporate
Attribute	education	engineering	diplomatic	political	political
Attribute	agree-ableness	organizational	social capital	unplanned	religious
Attribute	gender				islamic
Attribute	trust				
Attribute	race				
Attribute	extro-version				
Attribute	need for power				

3 Data and Data-extraction

Our goal here is to illustrate how this integrated approach works and to delineate key difficulties that need to be overcome in creating a fully working set of tools, such that the suite is useful in multiple contexts.

3.1 The Covert Network Data

For this illustration we use a series of 247 texts. These texts were collected through lexis/nexis as well as from web sites, news programs, trial transcripts, research papers, and various books. Search terms were a set of names generated by subject matter experts (SMEs) on the top 116 people and groups that have been of critical importance in this region in the past 25 years. We refer to this data set as the Mideast dataset. Three key caveats regarding data collection: First, the use of textual data alone to extract entity terms (nodes) often requires the use of a large corpus of texts drawn from various sources. For example, for HAMAS, we used a corpus of 306 texts drawn from the same range of sources like the Mideast data set. Second, to increase the credibility of the coding samples, the coded data is checked by independent human coders. Generally 1 to 10 per cent are sufficient for checking. Third translations and aliases need to be considered. For example, for the actors and organizations in the unified database, a set of aliases and alternative spellings were kept that enabled information on a single actor to be fused. We note that in the future automatic alias indicators might be able to be used here.

In general the types of textual data available will vary with respect to the source of the data itself. These sources include actual network members' descriptions of one another, non-network source descriptions of network actor's roles, actions, participation, education and so on, observed accounts of network actor's roles and behaviors (e.g., ethnographic summaries), and various media sources for related items (e.g., Al-Jazeera TV). Under the circumstances it is difficult to speculate about the first three sources of data, but certainly the first may be the most important in terms of gaining an actual understanding of 'native' accounts of actor's roles within the network, as demonstrated in the interview example above. Thus, in general, the analyst might want to augment textual data with other data. For example, in other studies we have augmented the networks extracted from texts on covert networks with data sets such as the Krebs 9-11 Hijacker data, the Tanzania Embassy bombing data (CMU-Alphatech Network project), and the Johnson-Kremple JI data.

3.1.1 Extraction of Network Data from Texts Using AutoMap

We used AutoMap to extract meta-matrix data from a set of texts and then to fuse these extracted "maps" into a combined network. Various combined networks can furthermore be consolidated into one network. AutoMap is a software for Network Text Analysis (NTA). NTA is a specific text analysis method that supports the encoding of the relationships between terms (words) in texts and the construction of a network of the linked words (Popping, 2000). NTA is based on the assumption that language and knowledge can be modeled as networks of words and the relations between them (Sowa, 1984). The extraction to semantic networks from texts NTA also implies the analysis of the existence, frequencies, and covariance of terms and themes; thus subsuming the analytic spectrum of classical content analysis (Alexa, 1997). Several NTA methods exist, 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), and Word Network Analysis (Danowksi, 1982).

AutoMap uses map analysis for the extraction and analysis of semantic nets. Map analysis focuses on the retrieval of meaning from texts. The method systematically extracts and analyzes the links between words in a text in order to model the authors "mental map" as a network of links between concepts (Carley, 1997; Carley & Palmquist, 1992).

We operationalized and formalized the map analysis methodology and implemented the technique in a software package called AutoMap (Diesner & Carley, 2004a). AutoMap's workflow is the following: As an input AutoMap takes a set of free-flowing, unmarked texts. Thus, articles published in mass media and texts available online can be loaded into AutoMap. As an output AutoMap represents mental maps as verbal networks, and statistical and network analytic measures. The tool also supports the comparison of maps generated with AutoMap. The automation of map analysis as provided in AutoMap enables the effective and efficient analysis of large collections of texts, the multi-level access to the meaning of textual data, the preservation and concise representation of the network structure of multiple texts, and the detection of the structure of social systems that are reflected in texts. In the following we give more detailed information about coding texts as maps in AutoMap.

AutoMap enables the user to extract semantic nets from texts in a computer-assisted manner. Computer-assisted coding means that the software applies a set of coding rules that were defined by a human in order to index the input texts and code them as networks. The coding rules that the user needs to specify in AutoMap are about text pre-processing and statement formation:

Text pre-processing condenses the data to the concepts that capture the features of the texts that are relevant to the researcher. This set of concepts defines the domain knowledge for a particular context. For example, the names of terrorist groups, terrorists, acts of hostility, and so forth are part of the "language" or domain knowledge peculiar to the discussion of terrorism; whereas, words such as "a", "and", "the", "Star Trek", and "Johnny Bravo" are not. Thus, pre-processing simplifies the task of finding meaningful interpretations of texts. In AutoMap, pre-processing is a semi-automated process that involves four techniques: named-entity recognition, stemming, deletion, and thesaurus application (for more information see Diesner & Carley, 2004a). While it is fairly mechanical to create a delete list, creating a thesaurus requires significant knowledge of the context. Once developed, however, the thesaurus enables aliases and various (mis-)spellings to be converted into core concepts, synonyms to be cross classified by the id they equate to in the context being researched. The result of this type of thesaurus application is a conceptual network in which a set of concepts and the linkages from them are extracted from the text. In the resulting network, all concepts are at the same ontological level. To move to the next layer of analysis it is key that these concepts are categorized by the type of entity that they represent. For social and organizational systems the meta-matrix provides such an ontology. We adapted AutoMap to enable the use of the meta-matrix ontology (Diesner & Carley, 2004b). Our results show that AutoMap can be used to extract meta-matrix data.

3.1.2 Illustration of the extraction of meta-matrix data

To illustrate this approach we look at an article from The Wall Street Journal from July of 2003, p. A1 on the Jemaah Islamiyah (JI) terrorist group in Indonesia. A series of quotes from the article will be provided that help illustrate the kinds of data that can be extracted.

The following are the names of JI members mentioned in the article and the quotes from the article relevant to extracting concepts that can be identified as **actors**, *roles*, knowledge, *attributes*, *tasks*, *locations*, *groups, resources* and *actions*. Although this is a simple example it provides the grounds for discussing the extraction of meta-matrix data from textual sources:

Amrozi: "Often referred to as the *"smiling bomber,"* he worked as a *mechanic* in *Indonesia* and is *accused* of *acquiring bomb-making material* for several of *Jemaah Islamiyah's terrorist strikes*." (p. A6) "The *investigators* says that when Mr. **Amrozi** *woke up*, "He was very *easygoing*. He *said* 'What's going on?—and then he *laughed*. He *laughs* all the time." (p.A6)

Abu Bakar Baasyir: "Among them was **Abu Baakar Baasyir**, a white-haired *Islamic cleric* alleged to be the *leader* of *Jemaah Islamiyah*, the *al Qaeda*-linked group to which Mr. **Amrozi** belonged." (p. A1) "*Indonesian cleric believed* to be the *leader*, or *Emir,* of the *JI*. He *operated* a famous religious *school* known as Pesantren al-Mukmin outside *Solo*, in central *Java*." (p. A6)

Iman Samudra: "The *alleged mastermind* of the *Bali bombing*." (p. A6)

Azahari Husin and Rusdi: "They would later *find* during the *interrogation* of Mr. **Rusdi** that *walking* several meters behind Mr. **Rusdi** when he was *arrested* was **Azahari Husin**, a *Malaysian lecturer* and one of two key *JI bomb-makers* who *remains at large*" (p. A6)

Nasir Abbas: "The *lieutenant,* **Nasir Abbas**, *said* JI members were again *warned* to *limit the use* of their *mobile phones*." (PA6)

Riduan Isamuddin: "He says *key leaders* of the group, such as **Riduan Isamuddin**, who is better known as **Hambali**, *remain at large*," (p. A6)

Abu Rusdan: "**Abu Rusdan**, who had *taken over* the *leadership* of the group after the *arrest* of Mr. **Baasyir** within weeks of the *bombing*." (p. A6)

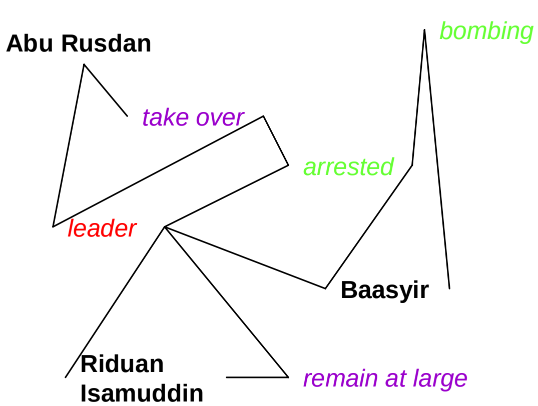
From these various quotes we can broadly identify a set of specific instances of each of the entities of the extended meta-matrix using AutoMap. Note, the stemmer in AutoMap will convert terms like Indonesia in to Indones, laughed and laughs into laugh, and leadership in to leader. A delete list could eliminate concepts like investigators. The thesaurus in AutoMap can be used to convert aliases, such as Hambali into Riduan Isamuddin. The specific entities that AutoMap would extract after the stemmer, delete list and thesauri are applied are shown in Table 2. Using the relative proximity of terms in sentences, AutoMap will create links between Abu Rusdan and leader, between arrest and Baasyir, and so on. Further, AutoMap will put links between terms based on proximity such that the user can control distance and the "sense" of proximity. Assume that all concepts not in the lists above are deleted. Then using a distance of 3 with breaks at the ends of sentence, the last two quotes would become the network in Figure 2.

In a sense, this is the easy part of coding. Implicitly the sources (texts) contain more information even vis-a-vie the extended meta-matrix that we are extracting. Consider the following illustrative limitations of the proposed coding. First, associated with each specific entity are a number of possible attributes. For example, several of the possible role terms represent informal roles (mastermind, bomb-maker, leader) and several formal roles (Emir, cleric, mechanic, lecturer). For example, different roles might be instrumental or expressive. The leader role of Bakar Baasyir, for example, may be more symbolic or expressive serving as a reference point for other members. On the other hand, the mastermind role of Iman Samudra may reflect more of an instrumental role in the network. In other words, coding could be improved if it were possible to infer attributes of some specific entities – in this case whether the role is formal or informal, expressive or instrumental. Second, many of these roles imply specific knowledge/skills, resources, and task assignments. Here, coding would be improved if it were possible to infer the presence of other specific entities and connections (ties) from specific entities in the text to those implied, such as a connection from bomb-maker to knowledge of bombs. Third, certain task-resource pairings, such as acquisition of bomb making materials, imply roles such as buyer. Here, coding would be improved if it were possible to infer the presence of other specific entities from a "fact" two concepts and the connection between them, and links from those concepts to the inferred entity. Fourth, the syntactic reach of different types of entities might vary. Thus, for the quotes examined, it might make sense to infer a tie between Abu Rusdan and Baasyir which are within 5 concepts of each other, and it make less sense to place a tie between take over and arrest which are within 3 concepts of each other. Potentially, coding could be improved if variable window sizes could be used for different pairs of entities. Fifth, pronouns are not coded automatically and numerous potential ties are thus lost. In this case, coding could be improved by pro-noun resolution. Sixth, there are cases where the use of adjectives may provide information about attributes of multiple entities. For example, the comment "**Azahari Husin**, a *Malaysian lecturer*" can be taken explicitly as an attribute that the lecturer is Malaysian, but it can also be taken to infer that Azahari Husin is Malaysian. In this case coding could be improved by enabling multiple inferences regardless of distance. Finally, many of the various specific entities and the connections among them are difficult to assess due to the use of a single data source.

Table 1: Specific instances of entities derived from a Wall Street Journal (Wed. July 2 2003:1) article on the Jemaah Islamiyah terrorist group

Actors	Know-ledge	Resource	Task/ Event	Group/ Or-ganization	Roles	Actions	Location
Abu Bakar Baasyir	school	*bomb-making material*	*acquire*	*Jemaah Islamiyah*	*bomber*	*accuse*	*Indonesia*
Abu Rusdan		*mobile phones*	*operate school*	*al Qaeda*	*mechanic*	*remain at large*	*Solo*
Amrozi			*bombing*		*cleric*	*laugh*	*Java*
Azahari Husin			*arrested*		*leader*	*limit the use*	*Bali*
Rusdi			*strikes*		*mastermind*	*allege*	
Nasir Abbas					*lecturer*	*believe*	
Riduan Isamuddin					*bombmaker*	*find*	
Iman Samudra					*lieutenant*	*interrogate*	
					members	*walking*	
					terrorist	*woke up*	
						warned	
						say	
						take over	

Figure 2. Illustrative network



Of course, extraction of the meta-matrix for each of the terrorist groups to be considered will be done using multiple sources and the resultant database will be much less sparse with the addition of these other data sources. The use of multiple sources will aid in sorting out the types and numbers of entities. For example, the use of multiple texts will enable a better coding of leadership roles, identify further roles, and eliminate roles that are irrelevant (e.g., media conventions). An example of a possible irrelevant role is that of Lieutenant, which is a term used by the media and probably has no relevance here except that it may reflect a midlevel 'follower' role in the network (an extremely important one for group function, however, given that a group needs a proper mix of roles including leaders, followers and lower status actors). We note that there are thousands of sources, many of which draw on each other, from which we will extract the relevant meta-matrix data. However, the utilization of multiple sources will not resolve other key difficulties such as pronoun resolution and the need to infer relations and entities.

4 Initial Analysis

Despite the coding issues discussed above, the coding of texts with AutoMap is extremely valuable for dynamic network analysis. Computer-assisted text coding, as done with AutoMap, ensures systematic analysis, and enables the more rapid coding of a larger corpus of texts. However, building thesauri is person and time intensive. For example, it took three days to construct the thesauri for these texts. A generalization thesaurus is necessary to capture problems such as mis- or alternative-spellings of names. The cross-classification or meta-matrix thesaurus is necessary to extract the relational data; i.e., to convert a semantic or concept network into meta-matrix data. We note that, over time, fewer and fewer items need to be added to the thesauri and that, even with the substantial effort involved in constructing these items, the resultant coding is substantially faster and more accurate than when done by hand.

To date we test out the feasibility of the integrated suite of tools. We have applied it to various text sets on terrorists (see for example Diesner & Carley, 2004b). The coding of the Mideast data set resulted in information shown in Table 4. We did not code actions in this analysis, but we can go back and re-code these texts trivially with AutoMap. For example, this data was initially coded without roles and then later with roles.

Table 3: Characteristics of coded texts

Entity	**Unique # of entity analyzed**	**Total # of entity analyzed**	**# of texts analzyed entity occurs in**	**Total # of entity linked into edges**	**# of text linked entity occurs in**
agent	415	2703	236	4243	234
knowledge	186	1611	221	1937	207
resource	274	2228	215	2603	199
task-event	74	1024	188	1184	163
organization	309	3746	238	4274	234
location	282	3888	242	4545	238
role	257	3584	246	4960	240
attribute	230	2878	226	3184	221

In addition, AutoMap automatically codes the relations among these entities. The number of unique connections between entity classes is shown in Table 5. Again, since there are no actions coded there are no relations to or from such entities.

Table 4: Based only on 247 texts - number of connections from nodes of entity type (row) to entity type (column)

Entity by Entity	**agent**	**Know-ledge**	**resource**	**task-event**	**Organi-zation**	**location**	**role**	**attribute**
agent	605	122	79	35	174	191	389	129
knowledge	124	201	52	47	121	139	142	76
resource	117	77	415	74	144	245	131	98
task-even	82	28	47	46	95	177	94	89
organizat	277	166	189	104	459	336	469	223
location	243	121	243	76	267	762	331	155
role	839	148	81	73	404	284	467	192
attribute	232	172	196	71	387	213	449	251

4.1 Data Management

One of the key points is that additional data from other sources can be added to the unified database. For example, we have integrated data donated by researchers such as Renfrow. These additions led to a larger and more complete picture. Another key feature is that it is often necessary to extract only a portion of the data. For example, in this case we wanted to eliminate connections between actors in our group and other world leaders such as non mid-eastern politicians. Both of these functions can be relatively easily managed in an SQL type database.

4.2 Locating criticalities

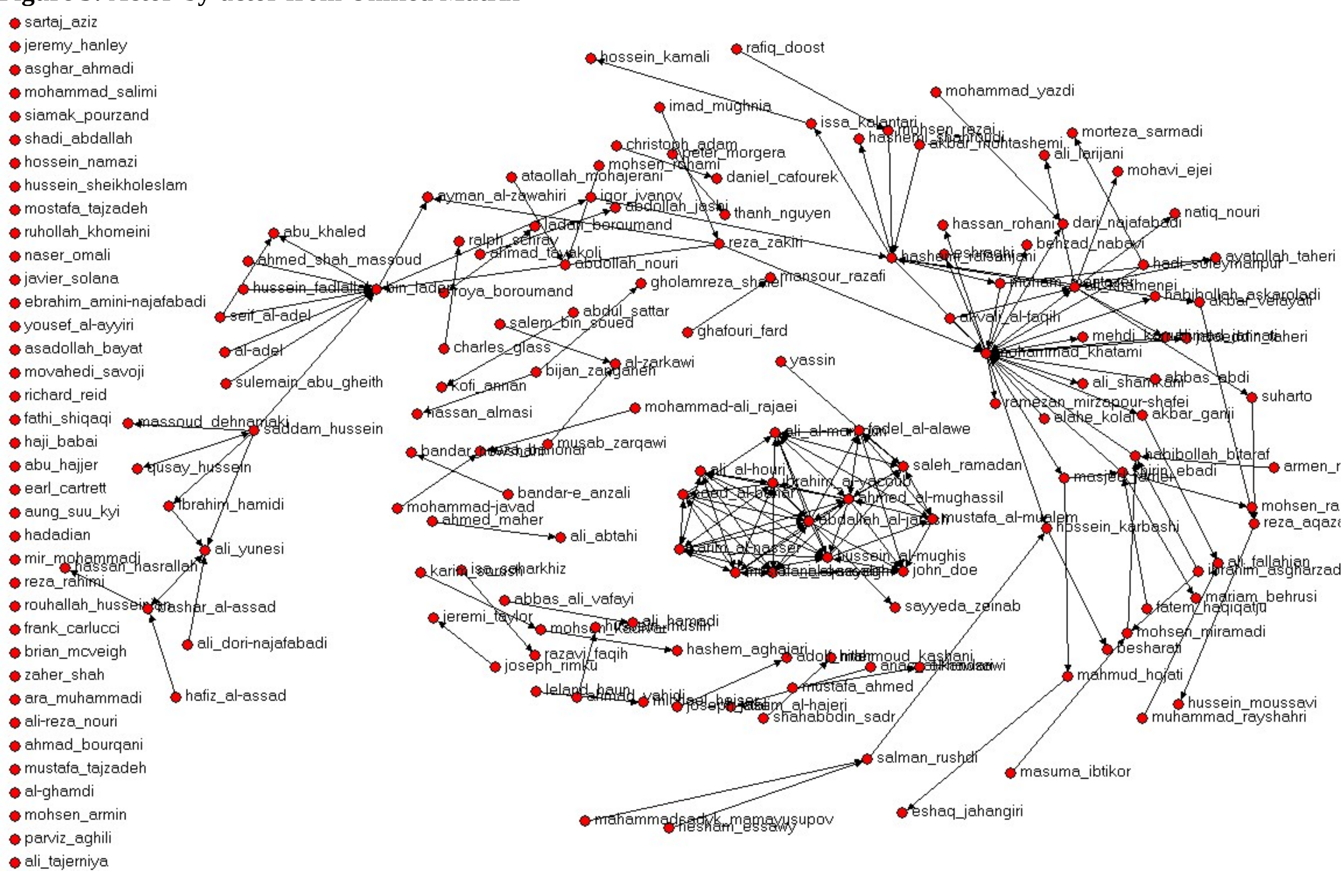
The resultant data is then analyzed with ORA. ORA can be used to locate vulnerabilities, optimize the network, etc. Illustrative measures critical for Intel have been added in ORA's intel report. This report is shown in Table 5 (next page) and has been annotated with the meaning of each variable. Note, a potential measures of coverage is the rate of change in these measures. That is, the slower the rate of change, the higher the likelihood that the appropriate information has been more completely extracted.

Table 5. Intel Report from ORA for Unified Database			
Measure	**Value**	**Definition**	**Meaning**
Cognitive Demand	0.0191 Mohammad Khatami	Actor with highest Cognitive Demand	Individual most likely to be an emergent leader, isolation of this person will be moderately crippling for a medium time
Total Degree Centrality	0.034 Mohammad Khatami	Actor with highest Total Degree Centrality	Individual most likely to diffuse new information, isolation of this person will be slightly crippling for a short time
Betweenness Centrality	0.0064 Mohammad Khatami	Actor with highest Betweenness Centrality	Individual most likely to be able to find out information, also good at spreading information, isolation of this person will be slightly crippling for a short time
Task Exclusivity	1.3679	Actor with highest task exclusivity	Critical individual, if the tasks are mission critical, isolation of this person is likely to be crippling
Clique Count	12 Abdallah Al-Jarash	Actor in most cliques	Individual embedded in multiple groups, most likely to get groups to work together.
Simmelian Ties	4 Mustafa Al-Qassab	Actor with most simmelian ties	Possible leader of the most cohesive sub-group.
Overall Complexity	0.0027	Density of the entire Meta-Matrix	Very low density. Probably major amounts of missing data, possibly cells are self directed.
Component Count	70	Number of undirected components in the entire Meta-Matrix	Possible indication that cells are self directed, possibly competitive factions exist.

4.3 Visualization

The next step is visualization. For example, in Figure 3 (next page) the actor-by-actor cell from the meta-matrix extracted from the Mideast data is shown. Note: the nodes on the left are isolates – actors who are not directly linked to other actors. We note that the networks extracted are large and improved visualization is necessary. In Figure 3, the actor-by-actor sub-matrix is shown and the isolates suppressed.

Figure 3: Actor-by-actor from Unified Matrix



5 Future Directions

We intend to code the Mideast dataset also for actions. Currently, this entity can be extracted using AutoMap. Now appropriate measures need to be added to ORA. Further, we are in the process of building a tool for generating input files for the simulation engines, like Construct, automatically from the data processed with ORA.

The expanded meta-matrix and the attribute based change mechanisms will be employed in DNA+, a new multi-agent network simulation model of network dynamics. In developing DNA+ we will draw on the lessons learned in the construction of both Construct (Carley, 1990; 1991b) and BioWar (Carley et al., 2004; Carley et al., forthcoming). Construct, as noted, is a model of the co-evolution of social and knowledge networks at the structural and cultural levels. From Construct we will draw on the learning, assignment, change mechanism models and the performance, diffusion and cohesion metrics. BioWar is a multi-agent model of the impact of weaponized biological attacks on cities in which agent actions depend on agent, role, task, and location attributes. From BioWar we will draw on the procedure for linking attributes to action and the scaling and optimization techniques developed for handling large scale (e.g., 1 million agents) network models. A significant portion of the proposed research will be to determine the change mechanisms and developing the appropriate computational algorithms for using them to alter the extended meta-matrix. A second portion of the proposed research will be to ensure the DyNetML interchange language can handle attribute data.

To improve the extraction of this data several things are needed. We can improve the coding using pronoun identification ad tagging words by their grammatical position. This is referred to as POS tagging. POS tagging (Abney, 1996a; Allen, 1995; Jurafsky and Martin, 2000), which requires parsing grammars. The advantage of POS based text analysis is that it nables the disambiguation of words with identical spelling, but different grammatical or syntactic function. The disadvantage of the POS approach is that is that is makes the extraction software, such as AutoMap, language dependent, non universal (today's parsers are domain dependent, language dependent, and corpus dependent), and slower. For each language a separate grammar and parser is required, and those are also domain-restricted. Thus, POS based text analysis can enhance the quality of coding of small datasets from pre-defined, restricted domains, but is not suited for general analysis. To date, AutoMap's data structure and architecture are based on an alternative approach referred to as statistical language processing (reference). Many researchers working on language technology suggest that statistical language processing (Abney, 1996b; Baeza-Yates and Ribeiro-Neto 2000; Church, Gale, Hanks and Hindle 1991; Manning and Schuetze, 2002) is the appropriate approach for: texts from different and multiple domains and corpora, texts in different languages, ad hoc information retrieval and data mining, and fast analysis of large-scale text collections. Statistical language processing techniques are applied to a variety of application-driven text analysis tasks, such as information retrieval, summarization, clustering, categorization, question answering, and machine translation. Furthermore, statistical language processing helps to replace ad-hoc solutions by theories that are identified during the process of practically employing statistical language processing techniques and to evaluate and validate yet untested or emerging theories. In short, statistical language processing seeks for unique and consistent patterns in and quantitative properties of texts while POS based analysis try to understand written language. Reclassification of the semantic or concept net into the meta-matrix requires a combination of understanding and pattern recognition. Hence, we are looking at appropriate mechanisms for blending the approaches yet retaining the ability to rapidly extract data from a large number of texts. We expect that using an expert system to reason about relations given an organizational based ontology, such as that in the meta-matrix will be a key first step.

References

Abney, S. (1996a). Part-of-Speech Tagging and Partial Parsing. In K. Church, S. Young, & G. Bloothooft (Eds.), *Corpus-Based Methods in Language and Speech* (pp. 118-136). Dordrecht: Kluwer Academic Publishers.

Abney, S. (1996b). Statistical Methods and Linguistics. In J. Klavans, & P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* (pp. 1-26). Cambridge, MA: The MIT Press.

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

Allen, J. (1995). *Natural Language Understanding*. 2nd edition. Redwood City, CA: Benjamin/Cummings.

Baeza-Yates, R., & Ribeiro-Neto, B. (2000). *Modern Information Retrieval*. Boston, MA: Addison-Wesley.

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

Borgatti, S.P., Everett, M.G., & Freeman, L.C. (c) (2002). *UCINET. for Windows. Software for Social Network Analysis.* Analytic Technologies, Inc.

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

Carley, K.M., & Kamneva, N.Y. (2004). *A Network Optimization Approach for Improving Organizational Design*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-102. URL: http://reports-archive.adm.cs.cmu.edu/anon/isri2004/abstracts/04-102.html.

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. (1990). Group Stability: A Socio-Cognitive Approach. In E. Lawler, B. Markovsky, C. Ridgeway, & H. Walker (Eds.), *Advances in Group Processes: Theory and Research* . Vol. VII. (pp. 1-44). Greenwhich, CN: JAI Press.

Carley, K.M. (1991). A Theory of Group Stability. *American Sociological Review*, 56(3), 331-354.

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. (1998). Organizational Adaptation. *Annals of Operations Research.* 75, 25-47.

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

Carley, K.M., Altman, N., Kaminsky, B., Nave, D., & Yahja, A. (2004). *BioWar: A City-Scale Multi-Agent Network Model of Weaponized Biological Attacks*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-101. URL: http://reports-archive.adm.cs.cmu.edu/isri2004.html.

Church, K.M.., Hanks, P., Hindle, D., Gale, W. (1991). Using Statistics in Lexical Analysis. In Zernik (Ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, (pp. 115-164). Hillsdale, NJ: Lawrence Erlbaum.

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/.

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

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.

Manning, D., & Schuetze, H. (2002). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT 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.