Toward an Interoperable Dynamic Network Analysis Toolkit*

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

Carnegie Mellon University

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 nodes and edges have attributes with possible data errors and missing data. The integrated CASOS dynamic network analysis toolkit described in this paper is an interoperable set of scalable software tools. These tools form a toolchain that facilitate the dynamic extraction, analysis, visualization and reasoning about key actors, hidden groups, vulnerabilities and changes in such data at varying levels of fidelity. We present these tools and illustrate their capabilities using data collected from a series of 368 texts on an organizational system interfaced with various terrorist groups in the MidEast.

Keywords: interoperability, social network analysis software, dynamic network analysis, meta-matrix model, integrated CASOS toolset, link analysis, counter-terrorism

* 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, the DOD, the CIA, 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 data processing.

† Direct all correspondence to Kathleen M. Carley, Institute for Software Research International, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213; e-mail: kathleen.carley@cmu.edu.

1. Introduction

We live in a complex world made so, in part, by the complexity of the social, legal, technical and other relationships connecting people to each other, and to other entities such as events and groups. Social groups such as a sport clubs, formal organizations such as IBM, and covert networks such as terrorist groups, are all complex socio-technical networks. Such complex systems have been studied by researchers in a number of areas, including Social Network Analysis (SNA) [42], forensic science [33], and link analysis [37]. These areas focus attention on the connections, also referred to as links or edges, between various entities, that are referred to as nodes.

The groups are not only complex, but they are also dynamic. The dynamics result from multiple change processes such as the natural evolutionary processes including learning, birth and aging; but also, intervention processes such as the isolation of key actors. Understanding and evaluating these groups is further complicated by the fact that the data is often incomplete, replete with errors, and difficult to collect. Consequently tools that go beyond the standard SNA and link analysis are needed. In response a new sub-field, dynamic network analysis (DNA), has emerged[6]. DNA combines the methods and techniques of SNA and link analysis along with multi-agent simulation techniques to afford the user with a suite of tools for looking at complex and dynamic socio-technical systems.

Identifying key individuals, locating hidden groups, estimating performance, and so on are only some of the tasks an analyst might want to accomplish when given a suite of DNA tools. DNA tools include those for data collection, analysis, visualization and simulation. Data analysis includes tasks such as identifying relations among individuals

and groups, characterizing the network's structure, locating key actors and points of vulnerability, and contrasting networks. Visualization tools help the analyst to explore systems graphically. Simulations support analysts in learning about possible changes in a system as it evolves naturally or in response to strategic interventions over time or under certain impacts, such as data modification. Herein, we describe a particular suite of DNA tools; those developed at the CMU CASOS lab, and illustrate their use in examining terrorist networks.

In DNA, social systems are represented as relational data. Relational data may reflect a plurality of node types such as people, organizations, resources and tasks (multi-mode), various types of connections among any two nodes (multi-plex), attributes of both, nodes and edges (rich data), and network data over time (dynamic).

Traditionally relational data was collected using labor-intensive survey instruments, participant observer recordings, or laborious hand coding of archival data. This had limited the quantity of data that could be analyzed and the results that could be obtained. Today, the internet, improvements in automated collection technique, and the presence of real time data feeds, have made it possible to collect larger, more detailed relational data bases, faster. Multi-agent simulation techniques make it possible to generate artificial worlds and data sets that approximate observed data and often fill in the gaps in such data. These changes have moved the field from trying to assess networks of less that a thousand nodes to assessing networks as large as 10^6 nodes and from networks at one point in time to networks at multiple points in time. Unfortunately much of the data collected is thrown out, ignored, or never examined, given the lack of good infrastructure tools for handling such data, and scalable processing tools. Many of the

traditional tools and measures for relational analysis (SNA and link analysis) do not scale well and cannot handle such large, dynamic and rich data. Techniques that can be performed by hand by a single analyst for a few texts or a few dozen nodes require many people-hours to be performed for the tera-bytes of data now available. Visualizations adequate for a few dozen nodes appear as incomprehensible yarn balls for large scale networks. Human projections of how networks are likely to change over time are unsystematic, biased, and do not consider all the interactions in the existing data. To be sure there are a growing number of tools that meet these needs; however, they are generally incompatible. For example, they use different nomenclature, database schemes, and ontologies. To meet the challenge of understanding, explaining and predicting complex dynamic socio-technical systems we need to move beyond traditional strategies and techniques [7]. More precisely, we need to further automate the SNA process and concatenate the different parts of it in order to adequately and efficiently represent and investigate relational data.

In this paper we describe and illustrate a novel approach towards the automated extraction, analysis, visualization and simulation of empirical and simulated relational data. The goal is to illustrate a specific DNA toolchain and show the types of advances needed in general to meet analysts' needs. We start by identifying limitations in the current software for doing relational analysis. These limitations, such as lack of interoperability and lack of scalability, inform a set of requirements for a DNA toolkit. Then we present a series of tools that begin to meet these requirements. As part of this presentation, we illustrate the types of results possible when such tools are applied sequentially. In doing so, we employ data collected from a group in the MidEast.

The tools presented form a toolchain for handling data on complex dynamic socio-technical systems. The strength of the approach presented comes, in part, from the fact that the tools are interoperable. Another strength of the approach is that the tools can be, and have been, used in a large number of contexts. Herein, we only show data from a political context that we refer to as the MidEast. However, we have and are applying these same tools in a variety of other contexts including other terrorist groups (al Qa'eda), corporate groups (analysis of the Enron data) and academic groups (analysis of a university department).

Several caveats are worth noting. First, while provocative, the findings *vis-a-vie* the MidEast data should not be viewed as conclusive; rather, they should be viewed as illustrative of the power of this approach. Second, although we stress the need for standards for data inter-change, we do not presume that standard used is the ultimate standard. Ultimately such a standard will need to be jointly and openly developed. We simply present a candidate in the hopes that others will join in the effort to define a more robust standard. Third, the focus here is demonstration. A full validation and usability study across multiple contexts is beyond the scope of this paper. Fourth, although we emphasize the need for scalable tools, we do not, in this paper due to space limitations, present scalability results.

2. Limitations of Current Relational Analysis Software

Software packages for gathering, analyzing, simulating and visualizing relational data exist and are being developed. As noted, these tools derive from three distinct disciplines including social network analysis, forensic science, and linked analysis. Hundreds of new tools have appeared since 9-11. If analysts want to use such tools in a project, they are

likely to face a plethora of problems including, but not limited to, scalability, robustness, interoperability, inconsistent data formats, lack of documentation. On top of this, the entire process may be overly costly in terms of labor and time requirements. For example, even though data exists there may be enormous labor requirements to get it into a form that is usable by the tools. Literally hundreds of analysts may be needed to parse data, such as newspaper articles, corporate reports, HumInt and OsInt reports, into a form usable by the tools. But the time costs do not stop with data entry. For example, sub-group extraction techniques often scale as N3 (where N is the number of nodes).

Monolithic packages can provide great analytic power. This often comes along with high complexity as well as specialization regarding the supported functionalities. This implies a steep learning curve for users. Proficiency in the usage of one tool does not necessarily ease the acquisition of skill with another package. Moreover, most tools perform some but not all of the tasks required of the analyst.

Stand-alone packages need to have communication interfaces with other tools in order to enable flexible research processes and efficient data and project management. But different tools use different, sometimes incompatible data formats. Most data formats are designed to serve their particular tool rather than serve as an interchange format between tools. For example, UCINET's proprietary format stores network data as binary files [5]. Others (DL, Pajek's NET [2]) rely on text files. File import/export options make it possible to use multiple analysis tools within a single project. The problem here is that data transformation steps might be required before data can be migrated from one tool to another. Format conversion can be a laborious and time intensive process. In the past, social network (people by people) and more complex social structure data (e.g.

organizations by resources) had been represented as distinct “datasets”. As network data sets grow in size and complexity, ad hoc approaches to the query, storage, extraction and manipulation of data become obsolete and new strategies are required. Furthermore, as research groups in the field join in large-scale projects, a need for a well-defined data interchange format has arisen. This situation is further exacerbated by the fact that large data sets are often best stored in a relational database so that diverse extractions can be created as needed for different DNA tools.

Another drawback of current tools for relational analysis is that many packages do not provide automation and scripting features. This inhibits batch-processing of data and therefore also inhibits efficient repetitions of analyses with various settings or on different data sets. This disadvantage results in an increase of labor and time required for analyses. This will be a growing concern as the opportunity for and the need to do comparative and over time studies increases,

In order to enhance timely decision making processes that involve the analysis of relational data, we need to provide solutions to drawbacks such as these. We note that analysts are often faced with the need to rapidly analyze complex socio-technical systems and answer diverse and ever changing questions about said systems. This is not just a processing issue, but a data collection, analysis, and reporting issue. Solutions are needed that support relational analysis using DNA tools in a way that is rapid, flexible, and robust. At this point, it is reasonable to conjecture that in the future, analysts will need access to suites of tools, loosely federated, and made interoperable through the use of standardized data formats and exchange languages. Major advances in automated

ontology creation, data-farming, and grid-based computing for large scale analyses are needed before a truly universal tool kit is possible.

3. Requirements for a DNA Toolkit

To overcome the drawbacks in current software for relational analysis we propose the creation scalable, flexible and robust DNA tools that can be linked seamlessly into DNA toolchains. The concept of an analysis toolchain is derived from the software engineering concept of development toolchains [32]. A software development toolchain consists of a number of small self-contained tools such as editors, project management tools, compilers, debuggers and analysis tools. Each of these tools might be developed as a separate product by different people and may vary in complexity, size, and features. In a similar manner, a DNA toolchain needs to consist of a number of self-contained tools that support various steps of the DNA process. The vision is that ultimately there will be a number of tools chains that twine through each other. There are numerous components of this vision, but, it predominantly center on the following points. There will be one or more underlying databases. Transition routines will move the data to/from the database and a common xml interchange language. Tools for data extraction, analysis, visualization, etc. can add, modify, or drop data from these databases by reading/writing to the interchange language, with appropriate locks and passwords. As new problems arise new tools can be created and added. Tools can be run interactively or though xml scripts. Complex analysis can automatically be run using grid computing techniques. Intelligent systems are used to locate appropriate tool chains as needed by the analyst.

We, as a community, are still moving toward this vision. Herein, we propose a DNA toolchain that moves us toward this vision and illustrates the kinds of difficulties

that will need to be addressed. The proposal for toolchains is driven by both a practical and a theoretical goal. From a practical standpoint the effective and efficient collection and analysis of network data across space and time needs to be enabled. A toolchain has the potential of enabling analysts to interactively mix and match tools that facilitate various parts of the SNA process. Script based versions of the tools, which enable batch-mode processing of datasets, would moreover enhance the efficiency of the analysis process. This gain in capabilities has the potential to enhance our understanding of complex dynamic socio-technical systems and provide support for timely decision making, assessment of effects based operations and action planning regarding such systems. From a theoretical standpoint toolchains, since they promote rapid collection and analysis, will increase the opportunities for meta-analysis. Major advances in our understanding of the social actor will be made possible by being able to rapidly and systematically collect, compare and analyze data on these complex and dynamic socio-technical systems.

Based on the limitations of current relational software we specify a set of seven core requirements for a DNA toolkit. These include: 1) toolkit extensibility; 2) tool interoperability; 3) common and extensible ontological framework; 4) XML interchange language; 5) database management; 6) scalable tools; and 7) robust tools. We now describe these requirements in more detail.

3.1 Extensibility

DNA toolkits will be used by analysts to answer a diverse and ever changing set of questions. As such, toolkits are needed. Further, these toolkits need to be easily extensible so that as underlying tools are refined and new measures and techniques are

implemented or new tool are created such changes can be easily accommodated. Building toolkits as suite of independent software tools allows for faster and more distributed software development. It also opens the door to independent developers to create their own tools that can connect to an ever growing common toolkit. Such a system can be facilitated by the use of common visualization and analysis tools, meeting common interoperable requirements, using XML interchange languages, and building individuals programs as web services, or at a minimum specifying their IO in XML. We note that there are currently a number of initiatives in the DOD, under Darpa, and under the NSF to move various types of tools in this toolkit direction. The time is ripe for such a toolkit in DNA.

3.2 Interoperability

Within such a toolkit, subsets of tools should form interoperable toolchains. All tools embedded in a toolchain need to be capable of using (reading/ writing) the same data format and data sets. This means that the output from one tool can be used as input to other tools. This does not mean that each tool needs to use all the data from another tool. Each tool needs to be able to operate on relevant subsets of data but without altering the basic data format. While not a hard requirement for a toolchain, if toolchains are to be built out of existing tools developed by a diverse set of developers, the use of an XML interchange language for IO would be beneficial. Other key aspects of interoperability are the use of a common ontology for describing data elements, and the ability for tools to be called by other tools through scripts.

3.3 Ontology

Socio-technical systems need to be represented by a model that captures the entities that such systems are typically composed of, the relations between those entities, and attributes of the entities and relations. Such a model and its implementation should be readily expandable in order to handle new types of entities and relations as they become relevant. The set of entity classes, relation classes, and attribute classes form an ontology for representing socio-technical system data. Ultimately, such ontologies should be automatically derived from the data and the analysts' needs.

Currently, for relational data, most of the focus has been on simple social networks. This there has been little effort to create an ontology with multiple entity classes. Secondly, only recently have multi-link data sets become the norm. Most tools to date, simply assume there is only a single relation type at a time. Third, few data sets have attributes, and so until very recently there has been little attention to what are appropriate attribute classes. In other words, there is not a wealth of candidate ontologies to choose from.

3.4 Data XML Interchange Language

Network data collected with various techniques or by various people, stored or maintained at different sites, and used as input and output of various tools needs to be represented in a common format. A common data interchange language ensures the consistent and compatible representation of various networks or identical networks in various states and facilitates data sharing and fusion. A common format will enable different groups to run the same tools and share results even when the input data cannot be shared. Tools need to be able to access data from diverse database with different database structures. These challenges require a data format that is engineered for

compatibility and flexibility and can serve a variety of tools. We define the following requirements for a DNA data interchange format:

- The format has to be able to represent rich multi-mode, multi-link, network data with multiple time points and multiple attributes of nodes and edges.
- The format has to be flexible enough to be used as both input and output of analysis tools.
- The format needs to be a human-readable file that can be parsed by computers.
- The format needs to allow one or many datasets - including computed measures on the networks to be stored in one file.
- Translations and aliases need to be considered. For example, for agents and organizations, a set of aliases and alternative spellings need to be kept to enable fusion of information on a single node. In the future, however, automatic alias detectors might be used here.
- The format has to allow developers to extend it in a fashion that will not break existing software.

This suggests the need for an XML interchange language. In the area of relational analysis there are a few XML schemes [37]. All existing schemes however, are geared for a certain type of network, such as a Markov network, or cannot handle attributes, such as graph-ml. A more flexible language is needed.

3.5 Data Storage and Management

If multiple DNA tools, which receive and return data, are used in a research project, a data storage and management system is needed. For this purpose, databases are typically used. The usage of a database for adding information to networks that are

already stored in the same database and the analysis of such extended datasets can lead to a larger and more complete picture of social systems. Moreover, the use of SQL type databases affords the analyst the luxury of using integral database tools for data search, selection, and refinement. The core difficulty at the moment, particularly in the intelligence area, is that there is not a common database structure. Hence, there is a need for translation and management tools to combine data across datasets, convert data from database into an interchange language used by tools, and so on. A key concern here is the need for a common ontology (as previously noted) so that diverse structures can be utilized but data rapidly fused together as needed. A second difficulty is that most existing relational data currently sits in flat files, in text files, in excel and is stored in a way that it is difficult to augment with other information. Utilization of SQL databases instead of these other formats will facilitate handling multiple instances of multi-mode, multi-link relational data with attributes.

3.6 Scalability

Traditional SNA tools have been designed and tested with small data sets (less than a thousand nodes). Link analysis tools have been designed and tested with relatively small data sets of relations. All of these tools need to be scaled to handle the large and complex datasets coming to be available. Scalability to at least 10^6 nodes and 10^7 links appears critical.

3.7 Robustness

DNA tools need to be robust in the face of missing data or common data errors. There are two aspects of robustness. First, measures should be relatively insensitive to slight

modifications of the data. Second, the tools should be able to be run even on data sets with diverse types of errors and varying levels of missing data.

4. Illustrative DNA Toolkit

Based on the needs of the analysts we have interviewed a number of core capabilities have emerged as being necessary for a DNA toolkit. These include, but are not limited to: 1) tools for populating the database, 2) tools for visualizing multi-mode, multi-link relational data-sets, 3) tools for identifying key actors, locating hidden groups, and identifying points of influence in a socio-technical system, and 4) tools for assessing change in that socio-technical system. Based on the requirements specified above and the capabilities identified by analysts as being important we have developed a number of tools that begin to meet the specified requirement and provide, at least at an elementary level, the capabilities identified as critical. As we present these tools, we focus on the methodology and leave out technical details on how to use the specific tools. The goal is not to provide the reader with a step by step guide to using these tools, but rather, to show the general power afforded to the analyst of using a toolkit that meets the identified requirements and capabilities, particularly a toolkit with embedded toolchains.

Our basic goal was to provide a set of tools that enable the analyst given a new context to rapidly gather data, analyze the data, and make predictions about that socio-technical system. We wanted a generic toolkit that could be used in diverse contexts to facilitate the gathering and investigation of multi-mode, multi-plex, multi-time period and rich relational data loosely coupled by data interchange and communication standards. In order to achieve this goal we modified existing tools (AutoMap [21], ORA [17], Construct [33], DyNet[6]) and implemented new tools (NetIntel [39], SocialInsight,

NetWatch [37]). We have applied the integrated CASOS toolset to the extraction and analysis of the structure of social and organizational systems such as covert networks [19][22], analysis of email collections [20], military exercises [24] and networks among scientific communities [13]. In the following we give an overview on the components of the integrated CASOS toolset and its features relative to the criteria previously listed.

4.1 Meeting the Toolkit Criteria

First, we describe how we met each of the criteria for a DNA toolkit.

4.1.1 Extensibility

Extensibility is made possible through the use of a common ontology and the XML interchange language. To test extensibility, after link three CMU CASOS tools together to analyze a data set, we then linked in several non CMU tools such as UCINET and NetDraw to ensure that such tools could also be used. This extensibility affords the analyst the ability to use whatever tool best meets the current need. This summer, at the CASOS summer institute, students will be using the extended DNA toolkit including tools from multiple developers such as CASOS and Steve Borgatti (creator of UCINET).

4.1.2 Interoperability

Interoperability is made possible through the use of a common ontology (the meta-matrix) as operationalized in the XML interchange language – DyNetML. In addition, to facilitate interoperability, all tools can be run under diverse platforms including windows, apple, and UNIX. The tools can also be run using scripts. Each tool is gradually moving to accept not just data but all instruction in XML.

Since each tool reads/writes relational data using DyNetML any other tool, regardless of where it was developed, can interface with these tools if they also read/write

DyNetML. In addition, the core statistical tool can read/write relational data in other formats such as those used by UCINET and Pajek thus increasing inter-operability with those tools. The CASOS tools also provide a network data converter package that enables transformation between the most widely used relational formats (none of which are XML based).

4.1.3 Ontology – The Meta-Matrix

At this point in time, automated ontology tools are still in their infancy. Thus, we decided to use an ontology derived from organizational theory that has been referred to as the meta-matrix [7][15][27]. The meta-matrix is a multi-mode, multi-plex approach to organizational design. Each socio-technical system is represented using the entity classes: actors, knowledge, resources, tasks, organizations and locations. Any two entity classes, which can be the same or different, and the relations among the elements in each entity class form a network. For example, there can be social networks (agent by agent), membership networks (agent by organization), or knowledge networks (agent by knowledge), among others. Between any two entity classes multiple relations can exist. For example between agents and agents there can be multiple relations such as "is_related_to" and "was_seen_with". Properties of the organization as a whole can be analyzed in terms of one or more of the networks contained in the meta-matrix. We have found this ontological scheme to be sufficient for assessing issues of power, vulnerability, and change in diverse contexts.

4.1.4 Data XML Interchange Language – DyNetML

We have developed DyNetML [40][41], a XML based data interchange language. DyNetML enables the exchange of rich social network data and improves the

compatibility of SNA tools. Figure 1 shows the hierarchical structure of a DyNetML document.

[Insert Figure 1 here]

DyNetML supports the representation of networks that consists of an arbitrary number of meta-matrix elements, each describing a complete multi-modal network. Each network consists of an arbitrary number of node sets and graphs. Node sets group together nodes of the same type, complete with any rich data (Properties and Measures). Each graph consists of a set of edges that connect nodes as described in the Node sets section, complete with any rich data attached to the graph itself or any of its edges. Arbitrary numbers of graphs and node-sets and the ability to add rich data to any object within the hierarchy provides users with high flexibility and enables the representation of complex datasets within one file.

DyNetML is open source and has evolved as users within and beyond CMU have augmented the language. Graph-ml is a subset of DyNetML.

Note, the simulation tools are designed to generate missing segments of data using statistical profiles and output simulated data in the same DyNetML format as the original data. This is crucial for performance evaluation of simulations, as it enables side-by-side comparisons of results of a simulation to the real datasets. By handling all data in the DyNetML format and using the same ontological model for real and simulated data new capabilities are facilitated such as: validation, model tuning, creation of partially artificial datasets for war-gaming, and so on.

4.1.5 Data Storage and Management

We have developed a database referred to as the NetIntel database [39]. The fundamental goal when designing the database was to provide a common SQL database structure that could be used with any data collected in using the meta-matrix ontology. The database employs the meta-matrix model and uses DyNetML as an input and output format. The database uses the PostgreSQL database engine. Extraction capabilities were written in the languages of PL-SQL, which is used to create extractions on the fly and execute complex SQL queries, and C++, which supports the import of data from data gathering tools and the export of network data into analysis and simulation tools. Since SQL does not allow graph-theoretic computations as most of them require recursion, which is expressly forbidden in SQL semantics, we used PL-SQL to enable graph traversals within the database. The databases' structure is designed in an extensible manner, allowing for the easy addition of new attributes, node and edge types as they are added in DyNetML.

When designing the database we considered the fact that the data may come from various sources and include (foreign) named entities such as names of people and places, which may differ in spellings. To consistently control alternative spellings by converting them into canonical terms we integrated a Thesaurus table that associates various spellings with a unique canonical form. When a node or edge is inserted, queried or updated, a Trigger Function checks spelling of the entity or ID and makes sure that it is spelled in a canonical way. A drawback of our system is that the data populating the Thesaurus table had to be compiled by hand. However, with a simple conversion tool that also was integrated, NetIntel can make use of thesauri written for data collection tools, and can therefore capitalize on the manual work that was invested in the creation of

thesauri. Utilizing the same thesauri for both data coding and storage minimizes potential errors in interpretation.

Data management in NetIntel is enhanced by denoting the source of network data: for each node and each relation its source, entity classes, and set of associated edges and nodes are stored. This facilitates creating large-scale multi-source datasets while preserving the original data sources. We made this design choice in order to enable a future functionality for weighting the confidence in the source.

As datasets grow it is often necessary to extract subsets from them; for example, when analysts want to look at ego-networks or a certain set of nodes. NetIntel as a SQL type database supports the extraction and deletion of a node (set) and its related edges. Subsets of the network can be extracted based on graph-theoretic properties of the network such as graph distance (e.g., “Find all nodes at a graph distance of 2 or less from a given node”) and graph density (e.g., “Find all nodes embedded in subgraphs with given density”). Building upon our experience with network data we tuned the database to support extractions that are based on the source of data (e.g. “Find all social structure data that came from New York Times“ or “Find all articles from New York Times from 10/10/2003”) or attributes of nodes and edges (e.g. “What is the network of people who were born in Syria?”). Another form of subsetting network data is to create time slices. Those can be created in NetIntel from the complete dataset or any subset. The key issue in dealing with time is distinguishing the data of the source from the dates of the events, actions, and so on mentioned in the actual text. NetIntel handles this by treating the dates of events as attributes of nodes and edges and the date of the source as an attribute of the source.

We note that the use of an SQL database and associated tools affords the analyst with the ability to select and analyze only those data of interest, which reduces processing time. It also makes possible the comparison of alternative data sets which enables cross-cultural comparisons and facilitates more systematic learning from the past. Finally another feature of this approach is storing the aliases along with the ids, storing the attributes with the nodes, and so on. This facilitates data fusion and promotes more in-depth analyses such as linking psychological and structural information. This makes possible new understandings, such as a more detailed understanding of how leaders can be influenced.

4.1.6 Scalability

Currently we are in the middle of a major effort to make sure that all of the tools scale. At the moment all tools can handle network with up to 106 nodes and 107 relations. However, in the case of the simulation tools, networks of this size take days to be simulated. The data coding tools can code an unlimited number of texts; however, each text needs to be relatively short (a few megabytes). All measures in the statistical toolkit (other than the grouping algorithms and the optimizer) run in under 30 minutes for networks of this size, the vast majority of the measures take less than 10 minutes per measure.

4.1.7 Robustness

The CASOS tools degrade gracefully in the face of missing information. That is, when various entity classes are not available, the tools, make use of what is available. The robustness of the measures is still being assessed. Recent work suggests that even with

30% missing data (in terms of links) most DNA measures will still tend to get approximately the same rankings. However, much more work is called for in this area.

4.2 Toolchain for Covert Network Analysis

We now describe a toolchain developed and used to support the assessment of covert networks such as those described by Berry [3]. To orient the reader, Figure 2 illustrates the workflow used in a typical relational analysis process. As can be seen, this tool chain enables the analyst to move from raw texts to network to the identification of patterns in those networks to an analysis of possible effects of alternative interventions. This is accomplished using three types of tools: data coding, statistical network analysis, and computer simulation. In all cases, secondary tools such as those for visualization and data editing provide support functionality.

[Insert Figure 2 here]

4.2.1 Data Coding

Information about covert networks that is relevant in the context of homeland security is often conveyed in textual sources such as analysts' reports, transcripts of communication among people, or news coverage. Those data collections can entail hundreds of thousands of files. In order to enable efficient decision-making based on the information given in those sources, automated ways of extracting social structure from machine-read texts are necessary.

In principle, there are three steps here. First, collect the texts that you want to analyze using web-scraping or other tools. Second, convert these files into a format that an automated coding tool can use (such as converting all formats to .txt formats). And, third, run the coding tool.

AutoMap, a component of the CASOS toolset, supports this functionality [21]. AutoMap [23] is a statistical network text analysis system [9][36][1] that can be used to systematically convert texts into semantic networks that can then be cross-indexed using the meta-matrix ontology. Consequently, using AutoMap one can extract the structure of socio-technical systems such as covert networks from texts [19]. The software furthermore facilitates the fusion of the networks from diverse texts into a single meta-network. Thus AutoMap takes in raw texts and outputs relational data in DyNetML.

Texts are converted into networks using a distance based approach also referred to as windowing [18]. Windowing basically slides a fictitious window over the text and concepts within the size of that window are linked together if they match the coding rules specified by the analysts. Note, a concept is a single idea represented by a single word, e.g. mastermind, or a phrase, e.g. training camp. As concepts are linked together forming statements knowledge is extracted from the text in the form of a semantic network or map [10]. This process is illustrated and detailed information on coding rules is provided in the section “From Texts to Meta-Matrix Data”.

The key here is the term concept. The quality of the date coding and the speed with which data from a new context can be coded depend on the extent to which the analyst wants to specify specific concepts, recoded concepts into more general terms, or customize the tool to extract particular concepts. Once such customization is done texts can be rapidly converted to networks and raw data converted to a form that can be

[1] Network text analysis is based on the assumption that language and knowledge can be modeled as networks of words and the relations between them [36]. Several NTA methods exist (for an overview see [31], one of them being map analysis, which we have operationalized, formalized and implemented in AutoMap.

analyzed by various statistical network analysis toolkits or processed by various data-mining or machine learning tools.

4.2.2 Statistical Network Analysis

Given a set of relational data, such as a DyNetML file, the analyst needs to process it. Common analyses would be the location of key actors, hidden groups, or points of vulnerability. To do this, statistical and machine learning tools are needed to extract and interpret patterns in the relational data.

ORA is a statistical analysis toolkit that enables such analyses. ORA, unlike the vast majority of social network analysis tools, makes use of the meta-matrix ontology and facilitates the analysis of multi-mode, multi-link relational data. Thus, in contrast to other tools for analyzing relational data, ORA enables the user to calculate both traditional social network measures (like degree centrality) as well as measures that come out of other traditions but are calculable on meta-matrix data (like cognitive demand) [17][16]. ORA contains a number of sub-tools for pattern identification and analysis, creation of modified relational data, and analysis and comparison of diverse socio-technical systems. ORA takes DyNetML as input and generates DyNetML as output.

There are a number of advantages to a statistical network analysis tool like ORA that considers the entire meta-matrix. First, the analyses possible are more comprehensive and provide greater insight into the factors that drive behavior. The types of analyses that can be done in ORA include:

- Identification of the weak and strong actors or organizations in a network, points of influence, hidden sub-structure, organization's capabilities.

- Optimization of an organization's structure for various outcomes including general high performance and adaptivity.
- Comparison of the current organization with other organizations such as the same organization after a particular intervention, the same organization at a previous time point, equivalently sized random network, or other known organizations.
- Identification of the sphere of influence surrounding specific actors or organizations.

One of the key uses of such statistical network analyses is to identify possible courses of action and their immediate impact. An example would be the identification of the emergent leaders in a group as possible targets for conversion or isolation. The immediate impact of, e.g., isolating such targets, can be assessed by seeing the instantaneous change in the underlying networks. However, since these are dynamic systems, simulation is needed to move beyond the immediate impact and take into account the ability of people and groups to learn, evolve, and change.

4.2.3 Computer Simulation

Multi-agent dynamic-network computer simulation systems (MADN) can serve as effective tools for reasoning about the behavior of individuals and groups and the networks that constrain and enable their behavior. In traditional social network analysis, link analysis and forensic science, behavioral interpretations are drawn from a representation of the network at a particular point in time. The ability to look at how the network might evolve depends on the analyst's ability to think in multiple complex dimensions. In general, analysts have trouble doing this in their heads for more than two dimensions and can typically only look ahead a few time periods. But, MADN systems,

are able to assess the dynamics of complex non-linear systems and so make systematic forecasts of change in these systems [12][11][8] for many time periods. Simulation tools provide the analyst with a decision aid for thinking through the complexities of change in networks in response to various interventions. Analysts can use computer simulations to engage in various “what-if” analyses to begin to address questions of the probable. Note to predict what will happen, but reduce surprise and suggest the realm of what is likely to happen.

Unlike traditional economic models, the agents in MADN simulations act in a boundedly rational fashion [35] on the basis of their mental models, emulating what people might do. As part of MADN systems the actions of individual agents lead to changes in the underlying networks that then effect what actions agents take in the future. For example, typically agents obtain information via interaction with other agents. Some of that information might be “views” about a third agent. Acquiring such information changes the knowledge network (people to knowledge) which in turn leads to changes in the social network (who interacts with whom).

MADN systems are more valid and effective when the input is empirical data, when parameters in the models are based on empirical data, and when the generated outputs can be comparable to empirical data. An example of a parameter that can be set is rational for interaction. Empirical results suggest that people spend 60% of their time interacting with those to whom they are similar. In models, such as DyNet, multiple mechanisms for choice of interaction partner exist and these can be prioritized using this and other findings. Simulation results can be validated by comparing against known facts. For example, studies show that people's knowledge of each other decreases exponentially

with the increases in social distance between them [26]. In DyNet, the cognitive accuracy of each actors model of other decreases with social distance even when agents are initialized with perfect knowledge.

However, the key to using MADN models by an analyst is the ability to ask "what-if questions" based on the real relational data. For simulation tools developed at CASOS such as Construct [33], DyNet [6], and NetWatch [37], the integration of the meta-matrix ontology into the models facilitates the development and tracking of models in which different entity classes of the meta-matrix to evolve differently. For example, while actors can learn a piece of knowledge and so create a connection from actor to knowledge, knowledge cannot learn about actors. Further, all of these tools can read in a DyNetML file for a real socio-technical system and then "evolve" it over time subject to various constraints and then write it out in DyNetML. The "evolved" network can then be read into ORA and compared with the original data (showing predicted changes) or compared with a known later state of the original network (for purposes of validation). This interoperability dramatically increases the usability of all tools and provides the analysts with a powerful system for evaluating the potential effects of diverse operations. Given the current state of simulation, the results are most useful for the "relative" evaluation of different operations. That is, if a MADN computer simulation predicts that a particular operation will reduce conservativism in the MidEast by 10%, one cannot necessarily count on the "10 %", although one can count on the fact that conservativism is likely to decrease. Moreover, if the same model predicts that for operation A conservativism decreases by 10% and for operation B it decreases by 20% then one can count on the fact that the second operation will be more effective in decreasing

conservativism. That is, the key use of these models is to suggest the relative impact of different operations.

In this paper, to illustrate the value of MADN simulations we will use the model DyNet. DyNet is a complex system simulation model in which the social and knowledge networks co-evolve as agents interact, communicate, and engage in tasks. The tool captures the variability in human and organizational factors of groups under diverse socio-technical and cultural conditions. DyNet places the constructural model [8][11][12] [33] in an information awareness context and enables the user to explore alternative destabilization strategies, both isolation and information operations [1], under varying levels of information availability.

From an interoperability angle, the workflow between ORA and DyNet proceeds as follows. The analyst assesses the network and identifies alternative intervention strategies. These might include, do nothing, isolate the most conservative leader, or isolate the top five conservative leaders. Then the same DyNetML file is input into DyNet to initialize the system and the three strategies entered as alternative interventions. Then a virtual experiment is run and the resultant evolved networks saved. These new networks are generated in DyNetML by DyNet and can then be read in to and evaluated with ORA.

4.2.4 Support Tools – Network Visualization

Reasoning about and interpreting the results of statistical and simulated analyses of dynamic social network data is facilitated by the ability to visualize the data and metrics on the data. Visualization tools need to be integrated in and/or be usable with all tools in the tool chain. Many visualization tools are available and more are being developed (see

e.g. www.casos.cs.cmu.edu/computation_tools/tools.html). Key limitations applicable to many of these tools include the inability to handle data over time, the inability to meaningfully visualize networks over 1000 nodes, and the inability to visualize large meta-matrix datasets. We do not purport to resolve these issues, just to recognize that visualization is a critical aspect of analysis and interpretation. From a toolchain perspective, since different visualization tools have different strengths, it is important to facilitate interoperability between tools. As such, tools such as ORA, that need to visualize data both call up the SocialInsight visualizer internally and export data in another formats so that other visualizers, such as NetDraw (by Borgatti) can be used.

5. Illustrative Application of CASOS Toolkit

We now illustrate the potential of a toolkit in which sets of tools can be linked into toolchains. In this example, we assume that the analyst has been told to evaluate various courses of action for an area that is likely to become problematic in the near future. There may be little processed data and certainly a relational dataset does not exist. Thus the analyst needs to move from a set of raw texts to a what-if analysis in a relatively short period of time.

5.1 Network Data

The analyst is faced with a covert network dataset consists of 368 texts. We refer to this dataset as MidEastIV. Of the texts, 158 were collected through LexisNexis Academia via an exact matching Boolean keyword search. The media searched with LexisNexis included major papers, magazines and journals. Sources for the 210 other texts were web sites, trial transcripts, scientific articles and excerpts from books. The search terms were the names of 109 top people and groups identified by subject matter experts (SMEs) to

have been of critical importance in the MidEast region over the last 25 years. For each of the 109 individuals and groups the articles that were most relevant according to the LexisNexis sorting function are selected, or the most relevant articles from other sources according to independent researchers. The time frame of MidEastIV ranges from articles published in 1977 to 2004. MidEastIV contains 17,792 unique terms and 179,702 total terms. The number of unique concepts considers each word only once per corpus, whereas the number of total concepts also considers repetitions of words per text.

In general, the credibility of the coding samples can be increased by using a large corpus that integrates texts from a variety of sources. Hence, in this sample, the sources include texts generated by the network or agent(s) under consideration, such as manifestos, web pages, announcements of attacks, (transcribed) video messages; non-network sources and observed accounts of the network such as ethnographic summaries and scientific reports; and media sources. The first text type may be the most important one in terms of gaining an actual understanding of 'native' accounts of this system. In general, analysts might want to use a variety of text types and sources and also augment their networks extracted or texts with other types of network data sets such as socio-matrices. For example, in other studies we have augmented covert network data revealed from texts with the Krebs 9-11 Hijacker data [28] and the Tanzania Embassy bombing data [13].

5.2 From Texts to Meta-Matrix Data for Dynamic Social Networks

To extract the social structure of the system under investigation from the MidEastIV corpus we used AutoMap. When coding texts as networks in AutoMap the analysts have to make decisions about the coding rules regarding text pre-processing and statement

formation. Text pre-processing condenses the data to the concepts (in network terms nodes) that are considered to be relevant in a certain context or corpus. 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 "and", "Star Trek", and "Johnny Bravo" are not. Thus, pre-processing simplifies the task of finding meaningful interpretations of texts. Statement formation rules determine how the relevant concepts will be linked into statements (in network terms edges).

In AutoMap, pre-processing is a semi-automated process that involves four techniques (for details see [21]): Named-Entity Recognition, which retrieves proper names such as names of people and places, numerals, and abbreviations from texts [29]; stemming, which detects inflections and derivations of concepts in order to convert each concept into the related morpheme [24]; deletion, which removes non-content bearing concepts denoted in a delete list such as conjunctions and articles from texts [10]; and thesaurus creation and application, which associates concepts with more abstract concepts (generalization thesaurus) or meta-matrix entities (meta-matrix thesaurus). Meta-matrix thesauri allow analysts to associate text terms with meta-matrix entities, thus enabling the extraction of the structure of social and organizational systems from textual data. The result of the application of a meta-matrix thesaurus is a network in which all concepts are at the same ontological level.

For pre-processing we developed a delete list with 170 entries. Applying the delete list to the data reduced the number of unique concepts by 13.8 percent and the number of total concepts by 43.5 percent. Next we created a generalization thesaurus that

associates several instances of relevant named entities or ideas, aliases and misspellings into a canonical form. For example, Al-Mohsen, Abd Al-Mohsen and Abu Hajjer (another name under that Al-Mohsen is known) were all translated into the single-worded core concepts Abd_Al-Mohsen. The thesaurus was developed incrementally. This means that after each phase of extension and refinement we applied the thesaurus to the data and checked if further additions or modifications needed to be made in order to cover relevant terms. The resulting generalization thesaurus contained 3150 associations of text level concepts with higher level concepts. After applying the delete list and generalization thesaurus to the data we associated the remaining concepts that were relevant for analyzing covert networks with entities of the meta-matrix ontology by building and applying a meta-matrix thesaurus. The first column of Table 3 provides quantitative information on of the meta-matrix thesaurus. Further entities can easily be added to the meta-matrix ontology in AutoMap by adding categories to the meta-matrix thesaurus and associating them with terms. Thus, analysts can use their own ontology for text coding in AutoMap. While it is fairly mechanical to create a delete list, thesaurus creation requires significant domain knowledge.

To illustrate the meta-matrix text analysis technique we code a portion of a sample article [30] from MidEastIV as a meta-matrix network. The following are the names of agents in the network under investigation as they appear in the article and the information provided on them. Underlined are the relevant concepts that can be cross-linked with the meta-matrix entities. This simple example aims to provide the grounds for discussing the extraction of meta-matrix data from texts.

Abdul Rahman Yasin:

… Abdul Rahman Yasin, the Al Qaeda operative indicted who federal prosecutors indicted for mixing the chemicals in the bomb that rocked the World Trade Center, killed six, and injured 1,042 people on February 26, 1993.

Abu Abbas:
… Palestinian terrorist Abu Abbas made news March 9 by dying of natural causes in U.S. military custody in Iraq. Green Berets captured him last April 14 in Baghdad, where he had lived under Hussein's protection since 2000. After masterminding the 1985 Achille Lauro cruise ship hijacking, in which U.S. retiree Leon Klinghoffer was murdered, Abbas slipped Italian custody.

Hisham Al Hussein:
… the Philippine government booted the second secretary at Iraq's Manila embassy, Hisham Al Hussein, on February 13, 2003, after discovering that the same mobile phone that reached his number on October 3, 2002, six days later rang another cell phone strapped to a bomb at the San Roque Elementary School in Zamboanga.

Abu Madja and Hamsiraji Ali:
That mobile phone also registered calls to Abu Madja and Hamsiraji Ali, leaders of Abu Sayyaf, Al Qaeda's Philippine branch.

Abdurajak Janjalani:
It was launched in the late 1980s by the late Abdurajak Janjalani, with the help of Jamal Mohammad Khalifa, Osama bin Laden's brother-in-law.

Hamsiraji Ali
… Hamsiraji Ali, an Abu Sayyaf commander on the southern island of Basilan, bragged that his group received almost $20,000 annually from Iraqis close to Saddam Hussein.

Muwafak al-Ani:
Iraqi diplomat Muwafak al-Ani also was expelled from the Philippines… . In 1991, an Iraqi embassy car took two terrorists near America's Thomas Jefferson Cultural Center in Manila. As they hid a bomb there, it exploded, killing one fanatic. Al-Ani's business card was found in the survivor's pocket, triggering al-Ani's ouster.

From these quotes we can identify a set of specific instances of each meta-matrix entity (see Table 1).

[Insert Table 1 here]

After pre-processing the data the analyst specifies the statement formation rules that determine the proximity of terms in texts that will be linked into statements if they match the pre-processing scheme (for detailed information about coding choices in AutoMap and their impact on map analysis results see [21]). AutoMap will put links between terms based on proximity so the user can control distance and the “sense” of proximity. Assume that all concepts not underlined in the sample above are deleted while

original distances are maintained. The meta-matrix network shown in Figure 3 results when a distance of 6 with breaks at the end of paragraphs is used.

[Insert Figure 3 here]

In addition to the extraction of social structures analysts might be interested in investigating the properties of specific entities in the network. Table 2 shows the properties (roles and attributes) of each node in the sample network.

[Insert Table 2 here]

The report was generated based on a Sub Matrix Text Analysis run in AutoMap. A Sub Matrix Text Analysis distils sub-networks from the meta-matrix. This technique is typically used for analyzing particular sections of the meta-matrix in detail [19]. Sub-networks are, for example, membership networks (agent by organization) or organizational assignment networks (organization by task-event).

Next we analyzed the entire MidEastIV corpus using a window of size 4 and rhetorical adjacency.[2] With this setting we best covered the statements that hand coders were finding. The results from applying AutoMap are shown in Table 3.

[Insert Table 3 here]

Statements between meta-matrix entities were formed from eight classes of meta-matrix entities. These entities on average linked into 22.8 unique statements per text, raging from 2 to 60, and 53.5 total statements, ranging from 2 to 688. The number of

[2] In order to find the statement formation setting that most closely resembled the links that a human coder would find we ran several pre-tests where we randomly picked an input text, had two independent human coders code a portion of this text, and compared the hand coding results against the machine generated results. The human coders were using the same pre-processing material that was used by AutoMap. Based on the insights we gained from the pre-tests we decided to form links across each document using a window size of four. We applied the meta-matrix thesaurus in AutoMap in such a way that only concepts that text level concepts had been translated into were maintained in the pre-processed texts. All other concepts were disregarded and replaced with imaginary placeholders that ensure the maintenance of the original distance of the translated terms (rhetorical adjacency).

unique statements considers each statement only once per text, whereas the number of total statements also takes into account repetitions of statements. Note that maps generated with AutoMap are digraphs in order to adequately represent the inherently directed structure of texts, thus the lower and the upper triangle of the meta-matrix are not necessarily symmetric. Across the dataset 8394 unique ties and 19701 total ties were identified. Figure 4 provides an overview on the distribution of the total edges across the meta-matrix. Note that connections between and among roles and attributes were not considered for Figure 4 so that the Figure contains only a total of 13465 edges.

[Insert Figure 4 here]

5.3 Analyzing Dynamic Social Networks

The network extracted with AutoMap was output in DyNetML format, stored in the NetIntel database, and combined with other data such as SME information whether an actor was a conservative or a reformist. Then a new DyNetML file was extracted and then loaded as input into ORA and statistically analyzed. Note, it would have been equally fine to have gone AutoMap to DyNetML to ORA. ORA can be used to visualize the network and/or generate a number of reports.

Often, the first thing that an analyst does after loading the data is to visualize it. This can be done using SocialInsight from within ORA or the data can be output in DL format and visualized with other tools, such as NetDraw, as was done to produce Figure 5. Several features of the network standout. First, the nodes on the left are isolates – actors who are not directly linked to other actors. The circular sub-graph at the left side of the inner circle that is not connected to agents out of the sub-graph represents the people who were charged with the Khobar Tower Bombing in Saudi Arabia in 1994.

[Insert Figure 5 here]

Next, an analyst might ask “who is critical?”. One report that is particularly germane is the Intel report, which provides network analytic measures relevant to the Intel domain. The Intel reports identifies key actors and organizations including limited information on interpretation of the results. Table 4 shows the part of the Intel report that contains the top five individuals in the given socio-technical system with respect to measures that determine an individual’s prominence or importance in that system. Table 4 is annotated with the meaning and a potential interpretation of each measure.

[Insert Table 4 here]

In Table 4 we see that Mohammad Khatami and Ali Khamenei stand out in almost every dimension. Were these two individuals to be isolated the two individuals most likely to emerge as leaders are Hashemi Rafsanjani and Kamal Kharazi. Of these two, Rafsanjani is likely to have more support (degree centrality) but is likely to bring as many or as large of disjoint groups together as Kharazi. Were these two to work in opposition, the system as a whole could become slightly unstable.

The Intel report also contains information on the key organizations in the MidEastIV dataset. Some of this data is shown in Table 5. Here we see that the Islamic Revolutionary Guard Corps and the Guardian Council are the most dominant organizations due to their high degree centrality and large number of members. The Islamic Coalition Society however is the group most likely to connect other groups, but is only slightly more that way than the guardian council.

[Insert Table 5 here]

The Intel report also provides information on the overall nature of the socio-technical system – such as its complexity (0.0048 for the MidEast IV dataset). The integration of the system can be assessed by examining the number of components (94 for the MidEastIV dataset). This suggests a fairly integrated system. Note, components are maximally connected subset of nodes, also referred to as subgraph. Weak components do not consider directionality of a link, whereas strong components take a link's directionality into account. The existence of components indicates that a graph is disconnected. In order to support the analyst in putting the measures from the Intel report in a broader context and so support reasoning about the results, ORA's context report compares values for the system being examined with numbers computed on a directed uniform random graph of identical size and density as the given network (shown in Table 6), as well as making comparisons to values on other social networks stored in the NetIntel database.

[Insert Table 6 here]

The data in Table 6 indicate that the densities for the MidEastIV dataset are very low compared to what we are likely to see in other datasets. This suggests that either there is significant missing data or the system is structured extremely differently from other systems. The value of context information is that helps the analyst answer questions like "how strong is a .2?". As more and more data sets are added to the underlying corpus the value of such comparisons will increase.

Recall that Mohammad Khatami stood out as key on almost all dimensions. In Figure 5, Khtami appears in the lower right area of the graph where we can see that he is highly connected to other people. A question that an analyst might want to ask is how can

he be influenced or who does he influence. To examine this, we look at the sphere of influence around Khatami. The sphere of influence is the meta-matrix extension of the ego-net. In a standard social network (agent to agent) the ego-net is the set of others that ego is connected to and the connections among them. In the meta-matrix this concept has been generalized to the sphere of influence – the sets of other nodes (in all entity classes) that are directly connected to ego and the connections among them. The sphere of influence for Khatami is shown in Figure 6. As can be seen, this sphere is quite complex suggesting that there are no simple ways to influence Khatami.

[Insert Figure 6 here]

The analyst might also ask, "what is the immediate impact of a change in the network?" This question can be addressed by picking an action, such as isolating an actor, and then comparing the meta-matrix before and after this action has been applied. For example, let's assume that the top five individuals in cognitive demand (the emergent leaders) are isolated. The immediate impact can be seen in changes to items such as the new set of emergent leaders, or overall network metrics such as the estimated performance or speed of information diffusion. What changes are looked at depend on the analyst's question.

For the MidEastIV data, when the top five emergent leaders are isolated the new emergent leaders include Said Mortazavi, Kamal Kharazi, Reza Asefi, Morteza Sarmadi, and Hashemi Shahroudi. However none of these individuals are anywhere near as strong in the emergent leadership quotient as the original. This suggests that the system may enter a fragile state where major changes are possible. Further, by comparing the meta-matrix before and after the isolation of the original emergent leaders we find that this

change should drop the systems performance by 4% and increase the rate of information diffusion by 67%. This suggests that it might be fruitful to follow up such an isolation strategy with some form of information operation.

5.5 Simulating Dynamic Social Networks

To illustrate the use of a DNA model as part of the toolchain we will use DyNet. The DyNetML file for the MidEastIV data is used to instantiate DyNet. A simple virtual experiment is run – do nothing, isolate Khatmi, isolate Khamenia, isolate the five most central reformists, isolate the five most central conservatives. Note, this virtual experiment was informed by the results in the Intel report. Each of these conditions was run multiple times in typical Monte-Carlo fashion.

Within DyNet the social and knowledge networks co-evolve as individuals interact. This process enables the social network to recover from attacks. Imagine that agent A and B only interact through agent C. Imagine further that agent C is isolated. Over time, through a processes of introduction and learning agent A might come to interact with B, or another agent D, might come to act as the go between. In this way, the network dynamically heals itself. The results from the simulation indicate the probable ways in which networks will evolve. Given that four of the conditions are interventions designed to break the networks, the results can be thought of as the near term impact of these operations. By using the simulation we can begin to estimate the extent to which an intervention will result in a significant as opposed to an immediate change.

At the end of the virtual experiment DyNet outputs a DyNetML file for each of the five conditions. This DyNetML file contains a complete snapshot of the evolved MidEast system after 50 time periods. These files can be read in to ORA and compared

with the original organization. Upon doing so an analyst might compare the original and the evolved network to see what new relations are likely to emerge. Thus near term change can be assessed using any of the ORA measures.

In addition, DyNet directly outputs change in information diffusion, performance, and change in a "belief" of interest. In Figure 7, the results of this virtual experiment for information diffusion are shown. As can be seen, any isolation strategy results in information taking longer to diffuse. The largest impact, however, is when Khatami is isolated. This is due in part to the intricate ways in which he is connected and his importance as an information conduit. His role is so unique that it is difficult to replace him. However, when he along with other reformists are isolated the impact on diffusion is less. This is due to a couple of reasons. First there are now five fewer individuals for information to diffuse to. Second, and more critically, there are so few reformists that removing them effectively enables new paths to form among the conservatives as well as among the conservatives and the neutrals. In other words, these top reformists had been blocking information flow.

[Insert Figure 7]

As noted, DyNet can also be used to assess change in a belief or the extent to which a belief is shared across the population. Here we treated reformism and conservativism as beliefs and examined how the various interventions effected the overall level of conservativism (Figure 8). The results indicate that in general, regardless of the intervention this population is becoming more conservative. Without Khamenei or the top conservatives the move to conservativism is slowed but not halted. And, not surprisingly, as more reformists are isolated the trend toward conservativism is exacerbated.

[Insert Figure 8]

6. Discussion

Independently the set of tools presented are valuable at addressing core analytical questions. Collectively, however, they provide the analyst with the ability to move back and forth between data in various forms, texts and processed, to compare real and virtual data, and to think about complex dynamic socio-technical systems in new and exciting ways. The strength of the overall approach will increase as new tools are linked in.

Each of the individual components has strengths and weaknesses. For extracting networks from texts we used a network text analysis approach embodied in AutoMap. The current operationalization is limited in that it does not make use of all of the information in the text. As such, the following improvements could be made in the text coding procedure. First, associated with each specific entity are a number of possible attributes such as roles. Several role terms represent informal roles (terrorist, leader) and formal roles (second secretary, diplomat). Different roles might be instrumental or expressive. The leader roles of Abu Madja and Hamsiraji Ali, for example, may be more symbolic or expressive serving as a reference point for other members. On the other hand, the operative role of Abdul Rahman Yasin may reflect more of an instrumental role in the network. Coding could be improved if it was possible to infer attributes of some specific entities – in this example 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 could be improved if it was possible to infer connections from a given node to implied nodes, such as a connections from mixing chemicals in the bomb to knowledge of bombs. Third, certain task-resource pairings,

such as acquisition of chemical materials for bomb making, imply roles such as buyer. Here, coding would be improved if it was possible to infer the presence of other specific entities from the "fact" that 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. Coding could be improved if variable window sizes could be used for different pairs of entities. Fifth, pronouns such as he or she are not coded automatically and numerous potential ties are thus lost. In this case, coding could be improved by anaphora resolution.

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. As the number of data sources increases there will be some decrease in the scarcity of the data, up to a limit. We also expect the use of multiple sources to aid in sorting out the types and numbers of entities present in a given social system. 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 anaphora resolution and the need to infer relations and entities. To improve the

extraction of networks from texts we plan to use an expert system to reason about relations given an organizational based ontology.

AutoMap uses a statistical approach to textual processing coupled with a secondary mapping of the semantic network using a simple ontology. This is valuable for dynamic network analysis. As the analysis of the MidEast data illustrates, computer-assisted text coding facilitates systematic analysis and rapid coding of a large corpora. We note that most of the changes identified above could be done by augmenting the statistical approach with an intelligent reasoning system operated at the social knowledge and linguistic level. However, such an approach increases the language dependency of the tool.

One final note on text processing is that even with AutoMap, building thesauri is a person and time intensive task. For example, it took three days to construct the thesauri used in this study. We note that, over time, fewer and fewer items need to be added to existing thesauri when new input texts are added and that, even with the substantial effort involved in constructing these items, the resultant coding is substantially faster and more accurate than manual coding. Although intensive, we note that three days is much less than the well over 100 person-days that would be required were all of the texts to be hand coded.

The statistical network analysis was done with ORA. As previously noted, one of the core advantages of ORA is that it enables the analyst to examine multiple networks, multi-mode and multi-link networks. Currently, we find that from a management and intervention perspective measures that utilize multiple cells in the meta-matrix have more predictive power than measures that consider connections among one class of entities,

such as people. For example, we used in this paper the measure cognitive demand. The estimation of cognitive demand utilizes most cells in the meta-matrix. Other studies have used this measure to successfully identify emergent leaders and potential chains of succession. Whereas, similar predictions using just the social network have failed to correctly predict succession.

Key limitations to the use of ORA are the lack of wizards to facilitate ease of use by the analyst, better integrated visualization, and more scalable grouping algorithms. In addition, as new entity classes are added, such as location, new measures need to be developed and added that make use of those entity classes. In this study we used a single attribute – reformist or conservative. Currently ORA cannot alter its behavior based on attributes. Ideally one should be able to select nodes with a certain attribute and analyze only those, contrast the behavior of entities with different attributes, display networks using the attribute information, and so on. Generally, most studies using relational data have ignored attributes. However, as move to providing a more rich understanding of complex systems we will need to account for differences due not only to relations but also to individual differences (attributes). Finally, major breakthroughs will require expanding ORA to facilitate more over time analysis. In that way, it can be used for examining historical trends and contrasting the trends in the real data with those predicted by the simulations.

Moving from ORA to a simulation system is another point where substantial time and person effort are requited. At the moment, although ORA can be used to identify potential interventions there is no way to store those interventions and automatically test them in the simulation tools. The DNA simulation tools, which use the same data as

ORA, enable the user to see the impact of altering the system by exploiting one or more of the vulnerabilities identified by ORA. But building the input file for the virtual experiment is still time consuming. Moving the output from the simulation tools back into ORA is straight forward as the output networks can simply be saved in DyNetML. Ultimately, it would be ideal to have the movement of data between ORA and the simulation engine be more automated and, for the most part, invisible to the analyst. This would facilitate seamless operation and enable analysts to assess the possible impact of change more rapidly, freeing their time for reasoning.

As to simulation tools, the one used herein, DyNet, enables the evolution of networks to be examined. Clearly there are many changes that could be made to the tool so that it could address a wider range of outcomes or generate more realistic results. The point we wish to make is that, an alternate approach is to provide multiple simulations for multiple types of problems and simply link in which ever is most relevant into the toolchain. In the long run this is likely to be more advantageous than creating a single monolithic engine.

This being said, there are some generic things that need to be done to the simulation engines. First MADN simulations tend to be relatively slow due to the network dynamics. Substantial research is needed into how to make such systems scale. Secondly, these systems can be used to evaluate a large number of virtual experiments. But to run the experiments necessary to completely characterize the response surface of these models is not only overly time consuming but would generate way more data than can be reasonably analyzed with existing tools. Here research is needed into data farming

environments and new statistical techniques for use with massive data sets. Ultimately, such tools will need to be linked into tool chains like the one described herein.

7. Conclusion

We have introduced a toolchain for extracting, processing, analyzing, and reasoning about social network data in general and covert networks in particular. Such toolchains are critical for the future analysis of covert networks because they admit flexible and efficient analysis. The proposed tool chain can be expanded as new tools and methods become available. Critical to the approach is the use of an ontology to provide a secondary classification on the underlying data, thus facilitating coding, analysis and simulation. Other critical features include the use of DyNetML, an XML interchange language, and an underlying extensible database using a common structure. Whether future work uses the presented ontology, interchange language, or database structure will depend on their completeness for the task at hand. All three items are open and will evolve as they are employed by diverse researchers and analysts. The key is that such features need to be common to enable the rapid development, integration and ease of operations in a tool chain.

Toolchains such as the one described here in can facilitate better analysis by reducing the time spent in repetitive tasks where little analyst insight is needed. Linking automated data collection tools to analysis tools makes it possible to rapidly assess new contexts. This is critical as new areas become “hot” in terms of terrorist, drug, or other illegal activity. Even if it is thought that text analysis is a poor substitute for the detailed reading an analyst provides (which is a debatable point), utilization of an automated system to jump start the analysis enables rapid early assessment. With this in mind, it is

critical that future generations of DNA tools take into account factors such as the confidence in the data, automated estimates of robustness, and tools for user in the loop data testing.

Moving beyond methodological issues, we note that from the perspective of understanding large, dynamic and complex socio-technical systems the approach used affords the analyst with greater analytical power. By taking into account not just the web of relations among people and organizations, but also their relations with resources, knowledge, etc., key insights into diverse behaviors can be garnered. In a sense, if we look only at the social network (people to people) than the focus of attention is on lines or authority, communication and other social relations. The addition of resources makes it possible to consider issues of economics. The addition of knowledge makes it possible to consider issues of training, learning, education, creativity and so on. In other words, by moving beyond the social, this inherently relational approach now has the promise of enabling effects based operation in areas as diverse as diplomatic, information, military and economic to be assessed in a relational context. Such analyses will provide greater insight into multiple aspects of the human condition.

References

[1] D. Alberts, J. Gartska and F. Stein, *Network Centric Warfare: Developing and Leveraging Information Superiority*, CCRP Publication Series (1999).

[2] V. Batagelj and A. Mrvar, Pajek - analysis and visualization of large networks. In: M. Juenger and P. Mutzel, Eds., *Graph Drawing Software*, Springer, Berlin (2003), pp. 77-103.

[3] N. Berry, The International Islamic Terrorist Network. CDI Terrorism Project, http://www.cdi.org\/terrorism\/terrorist-network-pr.cfm (Sept., 2001).

[4] S.P. Borgatti, *NetDraw1.0* (2002).

[5] S.P. Borgatti, M.G. Everett and L.C. Freeman, *UCINET for Windows* (2002).

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

[7] K.M. Carley, Smart Agents and Organizations of the Future. In: L. Lievrouw and S. Livingstone, Eds., *The Handbook of New Media* Ch 12, Sage, Thousand Oaks, CA (2002), pp. 206-220.

[8] K.M. Carley, On the Evolution of Social and Organizational Networks. In: S.B. Andrews and D. Knoke, Eds., *Research in the Sociology of Organizations* **16**, JAI Press, Greenwhich, CT (1999), pp. 3-30.

[9] K.M. Carley, Network Text Analysis: the network position of concepts. In: Carl W. Roberts, Ed., *Text analysis for the Social Sciences*, Mahwah, NJ (1997), pp. 79-102.

[10] K.M. Carley, Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis. In: P.V. Marsden, Ed., *Sociological Methodology* **23**, Cambridge, MA (1993), pp. 75-126.

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

[12] K.M. Carley, Group Stability: A Socio-Cognitive Approach. In: E. Lawler, B. Markovsky, C. Ridgeway and H. Walker, Eds., *Advances in Group Processes: Theory and Research* **7**, Greenwhich, CN (1990), pp. 1-44.

[13] K.M. Carley, T. Franz, G. Davis and J. Diesner, Surface Structure – Deep Structure. Paper presented at the *INSNA Conference*, San Diego, CA (2005).

[14] K.M. Carley, M. Dombroski, M. Tsvetovat, J. Reminga and N. Kamneva, Destabilizing Dynamic Covert Networks. *Proceedings of the 8th International Command and Control Research and Technology Symposium*, Vienna, VA (2003).

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

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

[17] K.M. Carley and J. Reminga, *ORA: Organization Risk Analyzer*. Technical Report, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, http://www.casos.cs.cmu.edu/projects/ora/publications.html (2004).

[18] J.A. Danowski, Network analysis of Message Content. In: W.D. Richards and G.A. Barnett, Eds., *Progress in Communication Sciences* **12**, Norwood, NJ (1993).

[19] J. Diesner and K.M. Carley, Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. In V.K. Narayanan and D.J. Armstrong (Eds.), *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations* Harrisburg, PA: Idea Group Publishing (2005), pp.81-108.

[20] J. Diesner and K.M. Carley, Exploration of Communication Networks from the Enron Email Corpus. *Proc. of Workshop on Link Analysis, Counterterrorism and Security at SIAM International Conference on Data Mining 2005*. Newport Beach, CA (2005).

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

[22] J. Diesner and K.M. Carley, Using Network Text Analysis to Detect the Organizational Structure of Covert Networks. *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference*, Pittsburgh, PA (2004).

[23] J. Diesner, E.T. Lewis and K.M. Carley, Using Automated Text Analysis to Study Self-Presentation Strategies. *Computational Analysis of Social and Organizational Systems (CASOS) conference*, Pittsburgh, PA (2001).

[24] J. Graham, Dynamic Network Analysis Estimation of Shared Situation Awareness, Ph.D. Dissertation (Draft), ISRI, Carnegie Mellon University, Pittsburgh, PA (2005).

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

[26] D. Krackhardt, Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. *Administrative Science Quarterly* 35 (1990), pp. 342-369.

[27] D. Krackhardt and K.M. Carley, A PCANS Model of Structure in Organization. *Proceedings of the 1998 International Symposium on Command and Control, Research and Technology*, Monterrey, CA (June 1998), pp. 113-119.

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

[29] B. Magnini, M. Negri, R. Prevete and H. Tanev, A Wordnet-based Approach to Named-Entites Recognition. *Proceedings of SemaNet02, COLING Workshop on Building and Using Semantic Networks* (Aug. 31, 2002).

[30] D. Murdock, Clarke's Not Blind. *National Review* (March 26, 2004).

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

[32] L. Ritter, Tools and methods for embedded system design using Ada. *Proceedings of the conference on TRI-Ada '88* Charleston, WV, CM Press, New York, NY (1989), pp. 416-425.

[33] R.E. Saferstein, *Forensic Science Handbook*, Prentice Hall, NJ (2001).

[34] C. Schreiber and K.M. Carley, *Construct - A Multi-agent Network Model for the Co-evolution of Agents and Socio-cultural Environments*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-109 http://reports-archive.adm.cs.cmu.edu/anon/isri2004/abstracts/04-109.html (2004).

[35] H. Simon, A behavioral model of rational choice. *Quarterly Journal of Economics* **69** (1955), pp. 99-118.

[36] J.F. Sowa, Conceptual *Structures: Information Processing in Mind and Machine*, Reading, MA (1984).

[37] M. Thelwall, *Link Analysis: An Information Science Approach*. Academic Press (2004).

[38] M. Tsvetovat and K.M. Carley, Modeling Complex Socio-Technical Systems Using Multi-Agent Simulation Methods. *Kuenstliche Intelligenz* **2**, Mannheim, Germany (May 2004), pp. 23-28.

[39] M. Tsvetovat, M., J. Diesner and K.M. Carley, *NetIntel: A Database for Manipulation of Rich Social Network Data*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-135. http://reports-archive.adm.cs.cmu.edu/anon/isri2004/abstracts/04-135.html (2005).

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

[41] M. Tsvetovat, J. Reminga and K.M. Carley, DyNetML: Interchange Format for Rich Social Network Data. *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference*, Pittsburgh, PA (2003).

[42] S. Wasserman and K. Faust, *Social Network Analysis.* New York: Cambridge University Press (1994).

URL's for CASOS software pages

AutoMap: http://www.casos.cs.cmu.edu/projects/automap
Construct: http://www.casos.cs.cmu.edu/projects/construct
DyNetML: http://www.casos.cs.cmu.edu/projects/dynetml
DyNet: http://www.casos.cs.cmu.edu/projects/dynet
NetWatch: http://www.casos.cs.cmu.edu/projects/NetWatch
ORA: http://www.casos.cs.cmu.edu/projects/ora/

Figures

Figure 1: Hierarchical structure of DyNetML

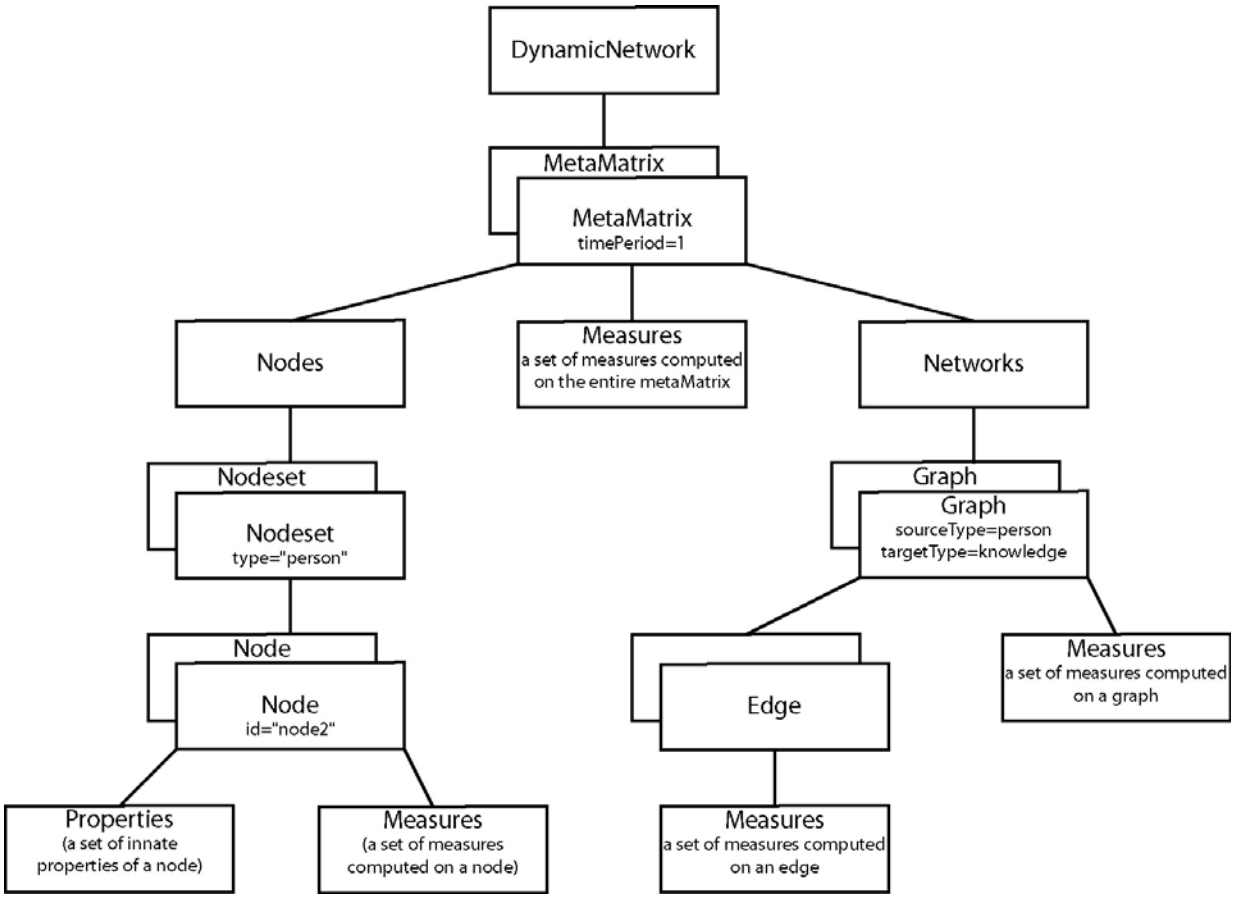


Figure 2: Workflow of integrated CASOS toolset

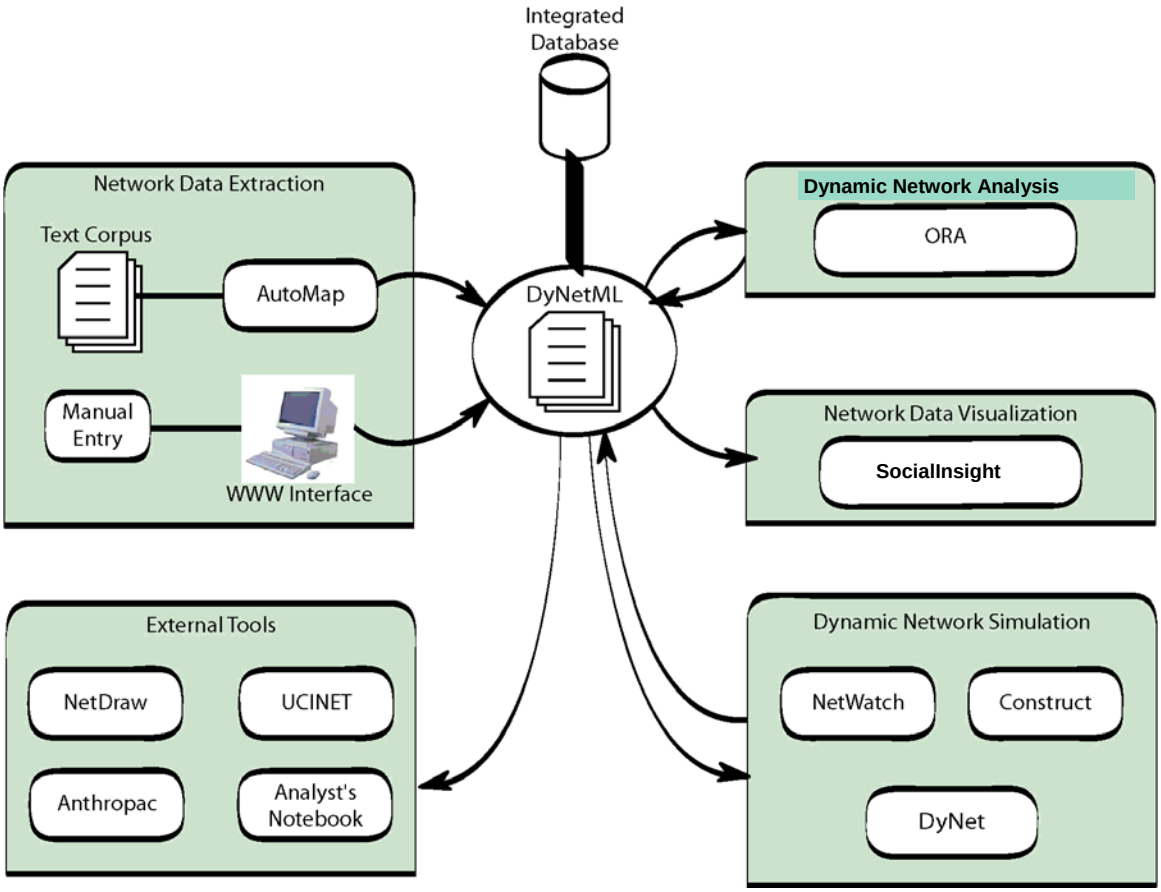


Figure 3: Visualization of sample meta-matrix network

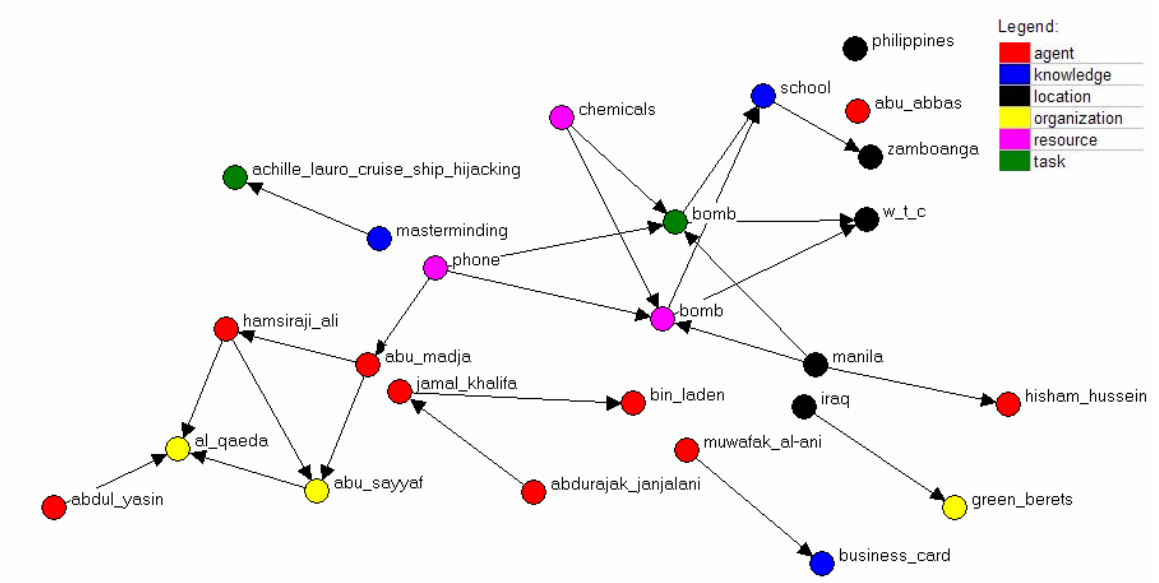


Figure 4: Distribution of edges across extracted meta-matrix

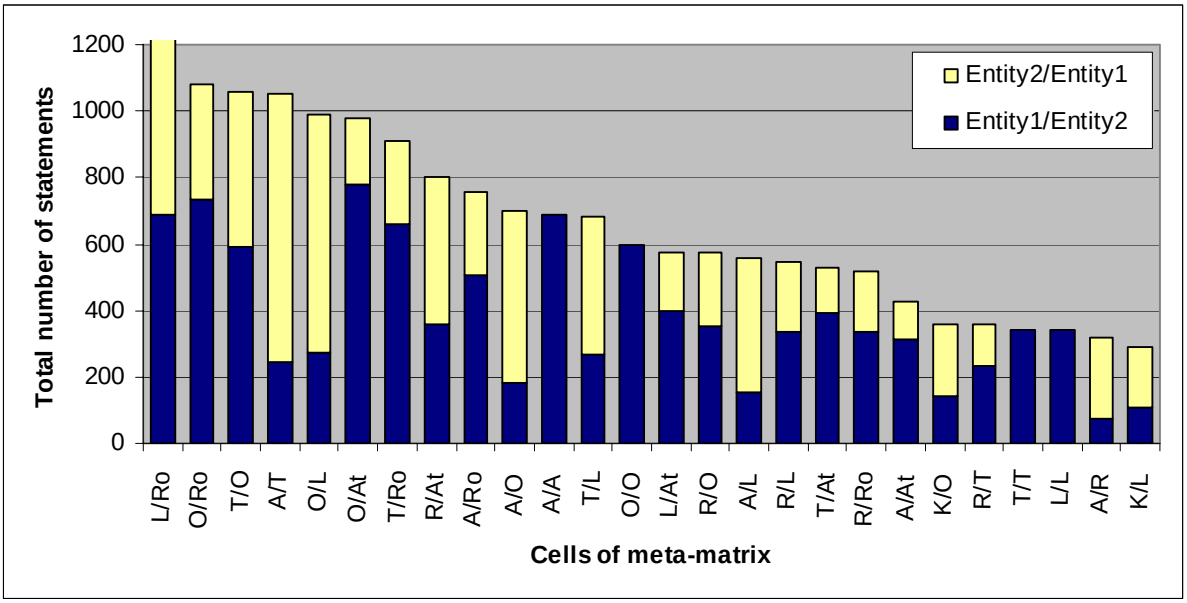


Figure 5: Social Network in MidEastIV

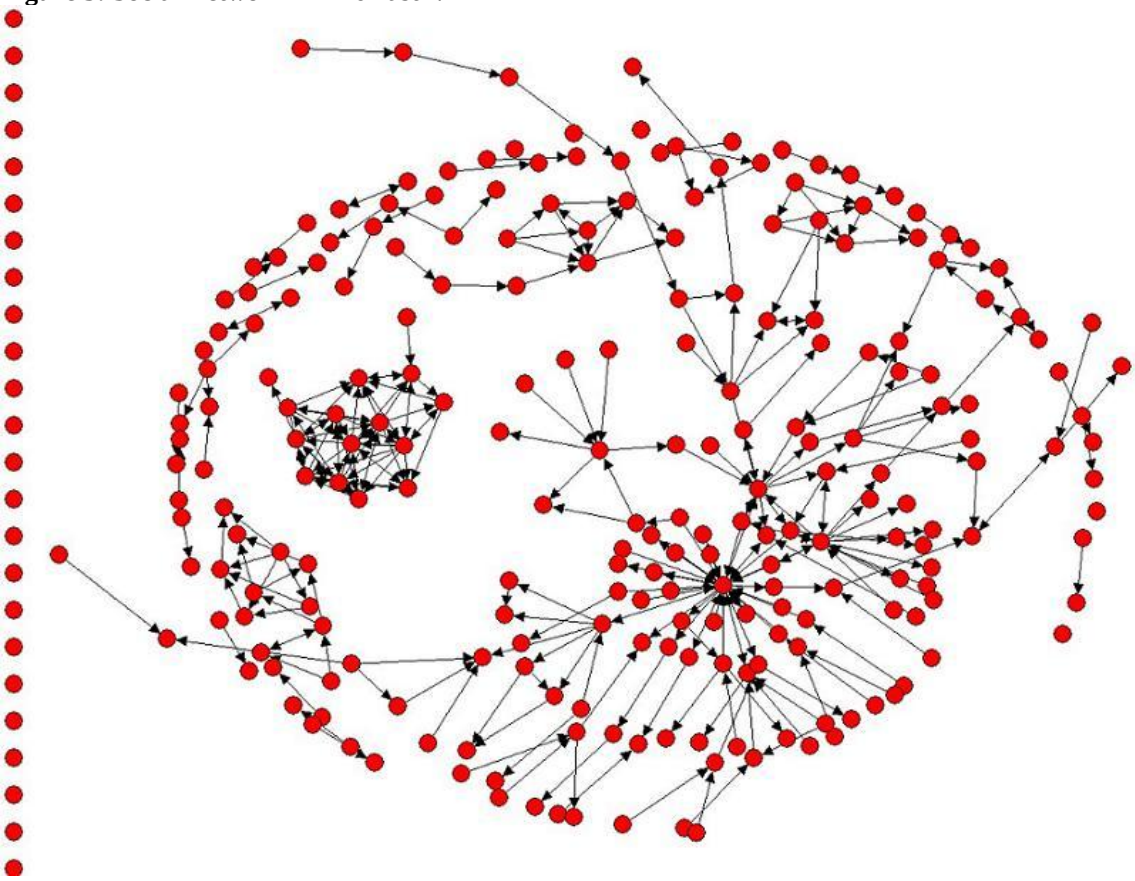


Figure 6: Sphere of Influence for Mohammad Khatami

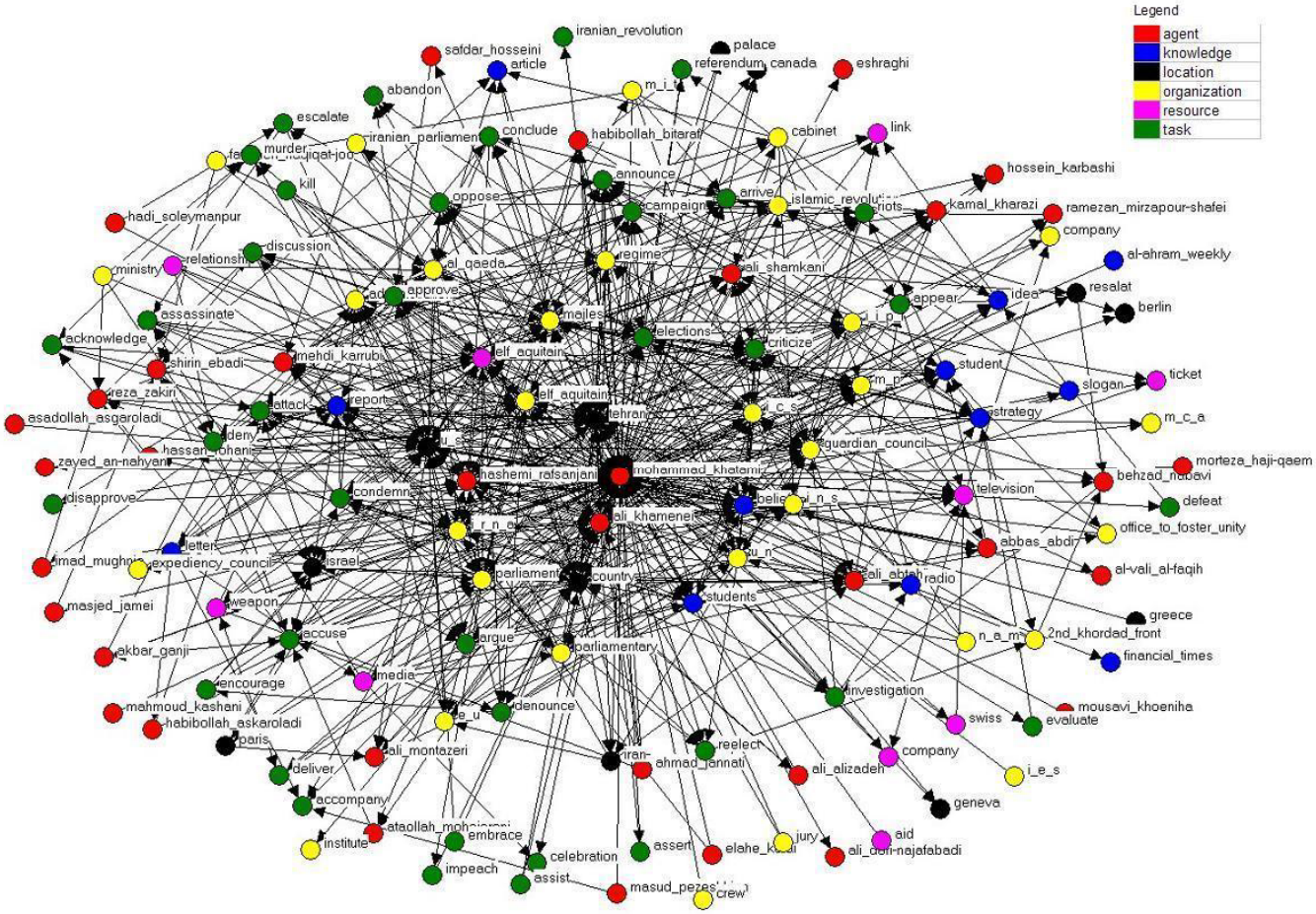


Figure 7: Diffusion Results from Virtual Experiment

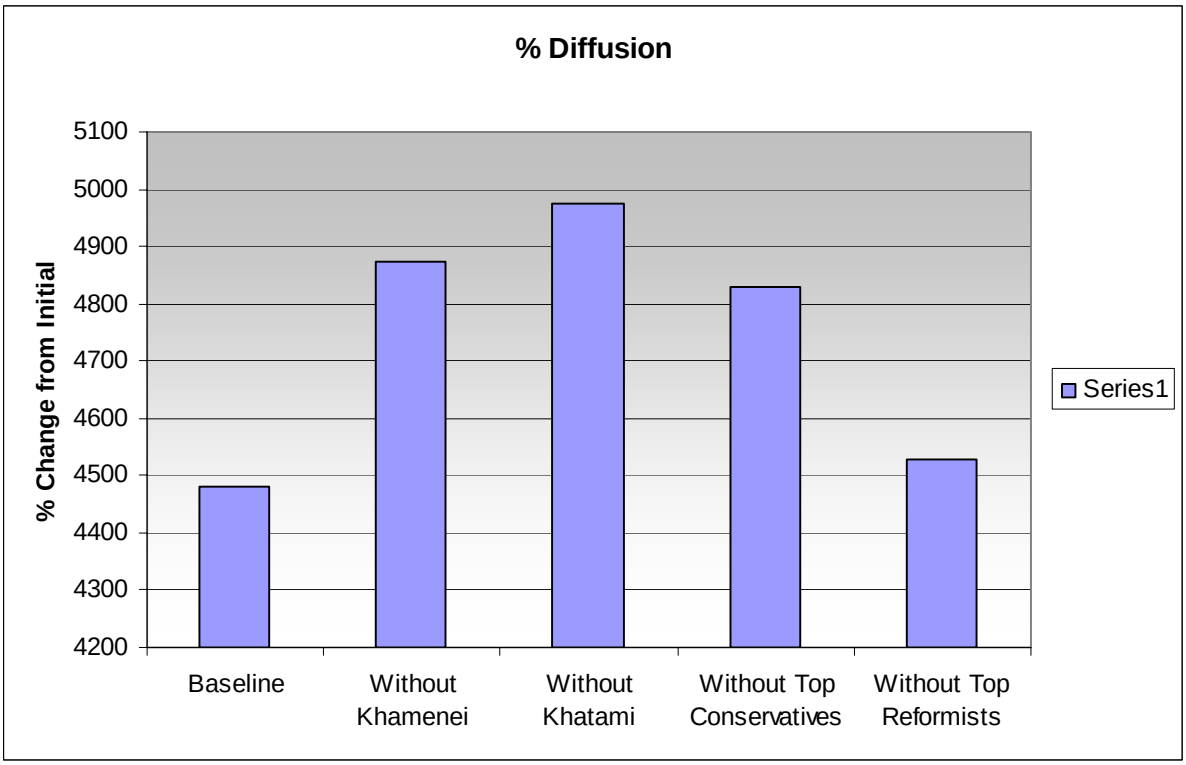
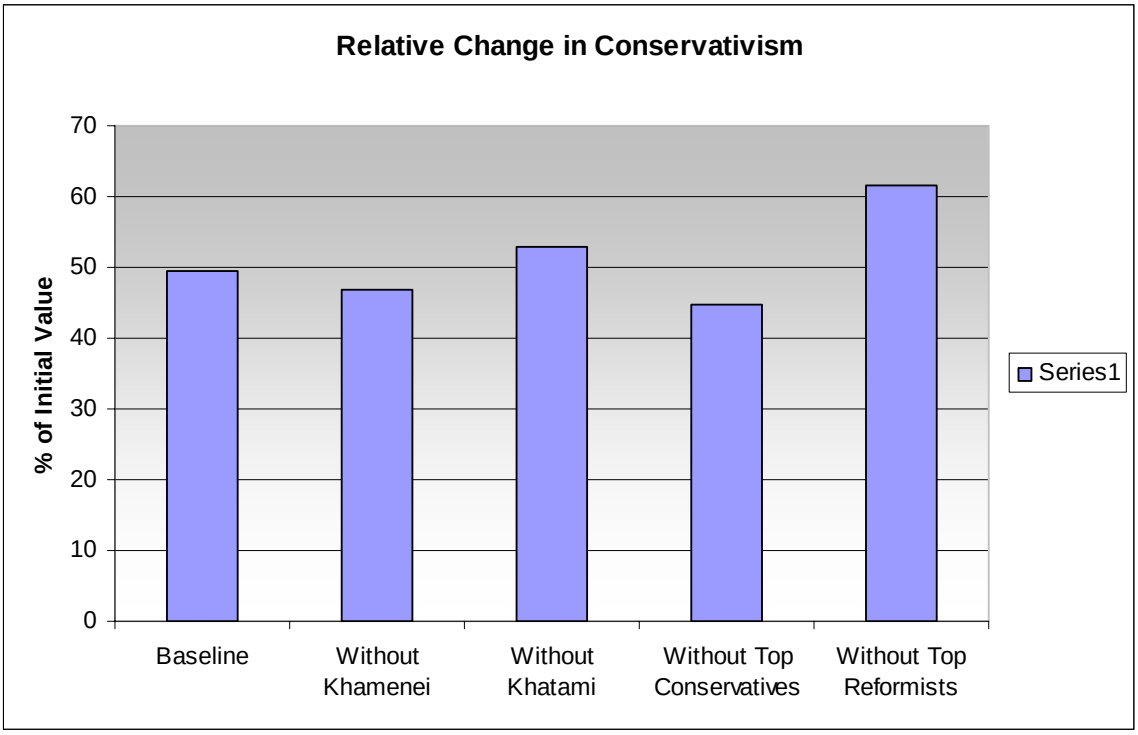


Figure 8: Conservativism Results from Virtual Experiment



Tables

Table 1: Exemplary instances of meta-matrix entities

Name of Individual	Meta-matrix Entity							
	Agent	**Know-ledge**	**Resource**	**Task-Event**	**Organization**	**Location**	**Role**	**Attribute**
Abdul Rahman Yasin		chemicals	chemicals	bomb, World Trade Center	Al Qaeda		operative	February 26, 1993
Abu Abbas	Hussein	masterminding		Dying, Achille Lauro cruise ship hijacking	Green Berets	Iraq Baghdad	terrorist	palestinian 1985 2000
Hisham Al Hussein		school	phone, bomb			Manila, Zamboanga	second secretary	February 13, 2003, October 3, 2002
Abu Madja			phone		Abu Sayyaf, Al Qaeda	Philippine	leader	
Hamsiraji Ali			phone		Abu Sayyaf, Al Qaeda	Philippine	leader	
Abdurajak Janjalani	Jamal Mohammad Khalifa, Osama bin Laden							1980s brother-in-law
Hamsiraji Ali	Saddam Hussein		$20,000		Abu Sayyaf, Iraqis	Basilan	commander	
Muwafak al-Ani		business card	bomb			Philippines, Manila	terrorists, diplomat	Iraqi 1991

Table 2: Properties of entities

Meta Matrix Entity	Name of Meta Matrix Entity	Attribute	Role
Agent	Abdurajak Janajalani	1980s	
	Abdul Yasin		operative
	Abu Madja		leader
	Muwafak Al-Ani	Iraqi	diplomat
	Hamsiraji-Ali	Philippine	commander, leader, second secretary
	Hisham Hussein	2003	
	Jamal Khalifa	brother-in-law	
	Abu Abbas	Palestinian	terrorist
Knowledge	masterminding	2000, 1985	
Task-Event	Achille Lauro Cruise Ship Hijacking	2000, 1985	
Organization	Abu Sayyaf	Philippine	commander, leader
	Al Qaeda	Philippine	leader
Location	Philippines	1991, Iraqi	
	Manila		second secretary

Table 3: Quantitative information on meta-matrix thesaurus and data pre-processing

Meta-matrix entity	Number of occurrence of entity in meta-matrix thesaurus	Total number of entity analyzed in corpus	Percentage of texts analyzed entity occurs in	Total number of entity linked into edges	Percentage of text in that linked entity occurs in
agent	577	3599	95.7%	5387	95.1%
knowledge	188	1849	81.8%	2005	72.3%
resource	301	2584	84.8%	2899	76.4%
task-event	264	3994	95.7%	4347	91.0%
organization	314	5463	96.2%	6483	94.8%
location	336	4113	93.8%	4802	89.1%
role	444	5319	98.9%	6814	97.3%
attribute	596	5239	98.9%	6665	97.0%

Table 4: Key Actors Located by Intel report

Measure	Rank	Value	Name of Agent	Meaning	Interpretation
Cognitive Demand	1	0.0591	Mohammad Khatami	Measures the total cognitive effort expended by each agent to do its tasks.	Individual most likely to be an emergent leader. Isolation of this person will be moderately crippling for a medium time.
	2	0.0579	Ali Khamenei		
	3	0.0356	Hashemi Rafsanjani		
	4	0.0244	Kamal Kharazi		
	5	0.0203	Ali Montazeri		
Degree Centrality	1	0.1618	Mohammad Khatami	A node has high degree centrality if it is directly connected to a larger number of other nodes.	Individual most likely to diffuse new information, most likely to know information,. Isolation of this person will be slightly crippling for a short time.
	2	0.0956	Ali Khamenei		
	3	0.0662	Hashemi Rafsanjani		
	4	0.0441	Hashemi Shahroudi		
	5	0.0368	Ali Montazeri		
Boundary Spanner	1	1	Mohammad Khatami	A node is a boundary spanner if it is between otherwise predominantly disconnected groups of nodes.	Individual most likely to connect otherwise disconnected groups. Isolation of this person might increase instability.
	2	0.8876	Ali Khamenei		
	3	0.8742	Mohammad Reza Aref		
	4	0.8305	Kamal Kharazi		
	5	0.5682	Hashemi Rafsanjani		
Eigenvector Centrality	1	1	Mohammad Khatami	A node has a high eigenvector centrality if the person is connected to many agents that are themselves well-connected	Individual who is most connected to most other critical people. Isolation of this person is likely to have little effect.
	2	0.7709	Ali Khamenei		
	3	0.578	Hashemi Rafsanjani		
	4	0.4104	Ali Montazeri		
	5	0.4023	Ahmad Jannati		
Task Exclusivity	1	0.0313	Ali Khamenei	An agent node has high task exclusivity if for one or more of the tasks performed there are a dearth of others who perform the same task.	Critical individual, if the tasks are mission critical, isolation of this person is likely to be crippling.
	2	0.0143	Kamal Kharazi		
	3	0.0099	Mohammad Khatami		
	4	0.0098	Reza Asefi		
	5	0.0072	Hashemi Rafsanjani		

Table 5: Table 4: ORA Intel report for central groups in the network

Measure	Rank	Value	Name of Agent
Degree Centrality	1	0.2083	Islamic Revolutionary Guard Corps
	2	0.2083	Guardian Council
	3	0.1667	Majles-e-Shura-ye-Eslami, Islamic Consultative Assembly
	4	0.125	Islamic Coalition Society
	5	0.125	Islamic Republic of Iran Broadcasting
Boundary Spanner	1	1	Islamic Coalition Society
	2	0.9692	Guardian Council
	3	0.8462	Mojahedin-e Khalq
	4	0.7487	Islamic Revolutionary Guard Corps
	5	0.6699	Majles-e-Shura-ye-Eslami, Islamic Consultative Assembly
Membership	1	0.0735	Majles-e-Shura-ye-Eslami, Islamic Consultative Assembly
	2	0.0588	Islamic Republic of Iran Broadcasting
	3	0.0441	Islamic Revolutionary Guard Corps
	4	0.0441	Guardian Council
	5	0.0294	Atomic Energy Organization of Iran

Table 6: ORA context report

Measure	Type	MidEast IV	Other Networks	Interpretation
Centrality-Betweenness	Mean	0.000	0.047	On average there are fewer paths by which information can get from any one person to any other person in this. group than in other groups.
Centrality-Closeness	Mean	0.002	0.380	On average it takes more steps for information to get from any person in this group to any other person in this group compared to other groups.
Centrality-Eigenvector	Mean	0.007	0.165	On average this group is less cohesive than other groups.
Centrality-In Degree	Mean	0.001	0.284	On average each person in this group is connected to fewer others than is typical for other groups.
Centrality-Information	Mean	0.002	0.060	On average each person in this group has less access to information than is typical for other groups.
Centrality-Inverse Closeness	Mean	0.005	0.473	On average each person in this group is closer to all others (takes fewer steps to send a message) than is typical for people in other groups.
Centrality-Out Degree	Mean	0.001	0.284	On average each person in this group sends information (messages/goods/advice) to fewer others than is typical for people in other groups.
Centrality-Total Degree	Mean	0.001	0.284	On average each person in this group has fewer connections to others than is typical for people in other groups.
Clustering Coefficient-	Mean	0.045	0.377	This group is less cohesive than other groups.
Component Count-Strong	Mean	511	8.5	On average there are more components in this group than in other groups: i.e. it is more disconnected. You might consider treating it as multiple groups.
Connectedness	Mean	0.056	0.798	On average people in this group are less connected to others than is typical in other groups.
Constraint-Burt	Mean	0.183	0.320	On average people in this group are less constrained in their action than is typical in other groups.
Diameter	Mean	545	22	On average information will take more time to flow through this group than other groups.