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 covert networks in the Middle East.

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 and reviewing.

[†] 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 and technical 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 business corporations, and covert networks such as drug networks, are all complex socio-technical networks. Such systems have been studied by researchers in a number of areas, including Social Network Analysis (SNA) [38], forensic science [30] and link analysis [34]. These areas focus attention on the connections, also referred to as links or edges, between various entities, that are referred to as nodes.

Complex socio-technical systems are furthermore dynamic. The dynamics result from multiple change processes such as natural evolutionary processes including learning, birth and aging as well as intervention processes such as altering the set of individuals who lead a system. Data on these systems is often incomplete, replete with errors, and difficult to collect, which further complicates the understanding and evaluation of these groups. Consequently, tools that go beyond traditional SNA and link analysis are needed. In response to these needs a new sub-field of SNA, Dynamic Network Analysis (DNA), has emerged [6]. DNA combines the methods and techniques of SNA and link analysis with multi-agent simulation techniques to afford analysts with a set of techniques and tools for investigating complex and dynamic socio-technical systems.

Identifying key individuals, locating hidden groups and estimating performance are some of the tasks that analysts might want to accomplish when applying DNA. DNA tools include software packages for data collection, analysis, visualization and

simulation. The process of data analysis includes tasks such as identifying relations among individuals and groups, characterizing the network's structure, locating the network elite or key actors, points of vulnerability, and comparing networks. Visualization tools help analysts to explore networks 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 targeted misinformation or systematic data modification. Herein, we describe a particular suite of DNA tools - those developed at the CMU CASOS lab - and illustrate their use in examining covert 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 data over time (dynamic).

Traditionally, relational data was collected using labor-intensive survey instruments, participant observer recordings, or laborious hand coding of archival data. These data collection procedures limited the quantity of data that could be analyzed. Today, the availability of real time data feeds and improvements in automated data collection techniques enable the faster collection of larger and more exhaustive data sets. Multi-agent simulation techniques made it possible to generate artificial worlds and data sets that approximate observed data and can fill in gaps in such data. These changes have moved the field from trying to assess networks of less than a thousand nodes to assessing networks as large as 10^6 nodes, and from networks associated with one point in time to networks transcending multiple points in time. Unfortunately, currently, much of the data

collected is thrown out, ignored, or never examined, given the lack of adequate and scalable infrastructure and processing tools. Many of the traditional tools and measures for relational analysis (SNA and link analysis) do not scale well and cannot handle large, dynamic and rich data. Techniques that can be performed manually 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, may reflect individual biases, and do not examine all interactions in the data. There is a growing number of tools that meet the current 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 tools [7]. More precisely, we need to further automate SNA and DNA processes and concatenate their different parts 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 goals are to show the types of advances needed to meet the analysts' needs and to illustrate a specific DNA toolchain. We start by identifying limitations in the current software available for analyzing relational data. Building upon this analysis we define a set of requirements for a DNA toolkit. Then we present a series of tools that were designed to 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 on a specific covert network.

The tools presented form a toolchain for handling data on complex dynamic socio-technical systems. The strength of the approach and resulting toolchain presented herein originate from the fact that the tools are interoperable and that they can be, and have been, used in a variety of contexts for the extraction and analysis of social and organizational systems such as covert networks [15][18], corporations [16], military training groups [21] and scientific communities [20].

Several caveats are worth noting. First, while provocative, the findings *vis-a-vie* the covert network used as a sample cannot 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 the standard used is an ultimate, we rather consider it as a starting point. Ultimately such a standard will need to be jointly and openly developed. We present a candidate in the hopes that others will join in the effort to define a more robust one. Third, the focus here is on demonstration. Validation and usability assessment across multiple contexts are beyond the scope of this paper. Fourth, although we emphasize the need for scalable tools, in this paper we do not present scalability results.

2. Limitations of Current Software for the Investigation of Relational Data

Software packages for investigating relational data exist and are being developed based on research in fields as diverse as computer-science, bio-informatics, and sociology. Many new tools have appeared since the 9-11 attacks in 2001 in the USA. If analysts want to use such tools in a project, they are likely to face difficulties such as limited

scalability, robustness and interoperability, incompatible data formats and lack of documentation. The entire process may be overly costly with respect to labor and time requirements. For example, even though reports from newspapers, corporations, human intelligence and open source intelligence exist, there may be heavy labor requirements to parse the data into formats required by the tools that analysts want to apply.

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

Stand-alone packages need to have communication interfaces with other tools in order to enable flexible research processes and efficient management of data and projects. However, different tools often require different, sometimes incompatible data formats. Most formats are designed to serve their particular tool rather than being usable 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 (agent by agent) and more complex social structure data (e.g. organizations by agents and by resources) had been represented as distinct data sets. As network data sets grow in size and complexity, ad hoc approaches to the integration, query, storage, extraction and manipulation of data become obsolete and

new strategies are required. As research groups join in large-scale and multi-disciplinary projects, a need for a well-defined data interchange format has arisen. Finally, large data sets are often best stored in a relational database so that diverse extractions can be created as needed for different analyses; however, none of the extant tools operates off a database nor is there a commonly shared ontology for creating such a database structure for relational data.

Another drawback of current tools is that many packages do not provide scripting features. This inhibits batch-processing of data, which in turn hinders 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 opportunities and need for comparative and over time studies increase.

In order to enhance timely decision-making processes that involve the analysis of relational data, we need to provide solutions to drawbacks outlined in this paper. 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 implies data collection, analysis, visualization, simulation and reporting issues. Solutions are needed that support relational data 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 that are loosely federated and are made interoperable through the use of standardized data formats and exchange languages. Advances in data-farming, automated ontology creation 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 of 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 [29]. A software development toolchain consists of a number of small self-contained tools such as editors, project management tools, compilers, debuggers and analysis software. 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 centers on the following points: There will be underlying databases. DNA tools can extract, add, modify or drop data from the databases by reading and/or writing to a common interchange language with appropriate locks and passwords. Transition routines will move the data between databases and tools. As new problems arise, new tools are developed and added. Tools can be run interactively and via scripts, which enable fast batch mode processing. Complex analysis can automatically be run using grid computing techniques. Intelligent systems support analysts in locating appropriate tool chains and components.

The DNA community is moving toward this vision. Herein, we propose a DNA toolchain that moves us toward this vision and illustrates the kinds of difficulties that need to be addressed. The proposal for toolchains is driven by both a practical and a theoretical goal: From a practical standpoint, analysts need to be supported in the

interactive mix and match of tools that facilitate various parts of the analysis process across space and time effectively and efficiently. From a theoretical standpoint toolchains - since they promote rapid, systematic and large-scale DNA - increase the opportunities for meta-analysis of complex and dynamic socio-technical systems. The gains in capabilities have the potential of enhancing the understanding of complex, dynamic socio-technical systems and of providing support for timely decision making, assessing effects based operations, and action planning regarding such systems.

Based on the limitations of current tools we specify a set of seven core requirements for a DNA toolkit: extensibility, interoperability, common and extensible ontological framework, common interchange language, data management, scalability and robustness. In the following we describe these requirements in detail.

3.1 Extensibility

DNA toolkits need to support analysts in answering a diverse and changing set of questions. Therefore, toolkits need to be easily extensible so that already integrated tools can be refined and new measures and techniques can be added. Building toolkits as modular suites of independent packages enables more distributed software development and allows independent developers to create their own tools and then connect them to a growing federation of toolkits. Such an integrated system can be facilitated by the use of common visualization and analysis tools, meeting common interoperable requirements, using XML interchange languages, and providing individual 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 Department of Defense (DOD), under the Defense Advanced Research Projects Agency (DARPA), and under the National Science Foundation (NSF)

to move various types of tools in a toolkit direction. Herein we focus on such a toolkit for DNA.

3.2 Interoperability

All tools embedded in a toolchain need to be capable of reading and writing the same data formats and data sets. As a result, output from one tool will be usable as input to other tools. This does not mean that each tool needs to be able to use all the data from the other tools, but each tool needs to be capable of operating on relevant subsets of data without altering the data format. While not a hard requirement for a toolchain, it would be beneficial to agree upon a common interchange language, preferably in XML, for input and output (IO). This would ease the concatenation of existing tools created by various developers. Further 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 Ontologies

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 expandable in order to handle new types of entities and relations as they become relevant. The sets of classes representing entities, relations and attributes should form an ontology for representing socio-technical system data. Ultimately, we will need tools that automatically derive such ontologies from the data or stem from theory.

For relational data, most of the focus has been on connections between agents (social networks). This approach has to be expanded for three reasons: First, multi-link

data sets became available recently, but there has been little success in defining ontologies with multiple entity classes. Second, most tools to date are built upon the assumption that there is only a single relation type at a time. Third, few data sets contain attributes, and so there has been little attention to what are appropriate attribute classes. Because of these classical approaches, there is not a wealth of candidate ontologies from which to choose.

3.4 XML Interchange Language

A common data interchange language ensures the consistent and compatible representation of various networks or identical networks at various states and facilitates data sharing and fusion. Therefore, network data collected with various techniques or 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 format or data interchange language can enable different groups to run the same tools and share results even when the input data cannot be shared and ensures consistent and compatible representations of relational data. Tools need to be able to access data from different databases with different 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 language:

- 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 represented as human-readable files that can be parsed by computers.
- The format needs to allow one or many data sets - 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 needs to be provided to enable fusion of information on a single node. In the future, however, automatic alias detection might be used.
- 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 [35]. All existing schemes, however, are geared for a certain type of network, such as a Markov network, or cannot handle attributes. 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 single research project, a data storage and management system is needed. For this purpose, databases are typically used. A database function that enables the adding of information to networks that are stored in a database and the analysis of such extended data sets can lead to a more complete picture of social systems. Moreover, the use of SQL type databases affords analysts the advantage of using integral database tools for data search, selection and refinement. The current core difficulty in the intelligence domain is that there is not a common database structure. Hence, there is a need for translation and management tools

that combine data across data sets and convert between interchange languages. A key requirement here is a common ontology (as previously noted) so that diverse structures can be utilized and data can be rapidly fused together. A second difficulty is that most of the currently available relational data is contained in raw text files or stored in excel files, which complicates augmenting relational data with attributes, such as a persons' age or gender. Utilization of SQL databases instead of these other formats will facilitate the handling of relational data.

3.6 Scalability

Traditional SNA tools have been designed and tested with small data sets. Small in this context means less than a thousand nodes. Link analysis tools have been designed and tested with relatively small sets of relations. All of these tools need to be scaled to handle large and complex data sets. Scalability to at least 10^6 nodes and 10^7 links appears critical.

3.7 Robustness of Tools

DNA tools need to be robust in the face of missing data and 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 on data sets with diverse types of errors and varying levels of missing data.

4. Illustration of DNA Toolkit

Based on the requirements specified above and discussions with over two dozen analysts from various DOD and Department of Justice (DOJ) analysis units[1] we have developed

[1] The interviews with analysts reflect an opportunistic sample of analysts who are interested in using, or have used, various toolkits for relational analysis. Many analysts were made available to the first author as subject matter experts in association with the particular contracts that supported this work.

tools that begin to meet the identified requirement and support, at least at an elementary level, the various steps of a DNA process. We further note that based on discussions with these analysts a number of core capabilities for DNA tools have emerged. 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. Our motivation behind designing and implementing the tools was to provide analysts with a set of tools that enable them to perform tasks such as rapid data collection, analysis and forecasting of processes in socio-technical systems as required for today's decision making processes. We were aiming for the power of a general purpose toolkit with the capability of having a customizable user-interface 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 [17], ORA [13], Construct [30], DyNet [6]) and implemented new tools (NetIntel [36], a network visualizer, a network editor, and NetWatch [35]). We refer to the resulting suite of tools as the *integrated CASOS toolset*. In the following, we give an overview on the components of the integrated CASOS toolset, focusing on how its components overcome drawbacks of DNA tools.

4.1 Meeting the Toolkit Criteria

4.1.1 Extensibility

Extensibility is made possible through the use of a common, extensible ontology (see 4.1.3) and an XML interchange language developed for relational data (see 4.1.4). To test extensibility, we first linked three of the tools from the integrated CASOS toolset together to analyze a data set (see 4.2). Then we linked to UCINET and NetDraw, two non CMU tools, to verify interplay with external tools (see 4.2).[2] This extensibility affords analysts the ability to use the tools that best meets their needs.

4.1.2 Interoperability

Interoperability is made possible through the use of a common ontology. Each tool reads and writes relational data using a common XML language. Any other tool, regardless of where it was developed, can interface with the CASOS tools if it uses the same XML language. Each CASOS tool is gradually moving to accept not just data but also instruction in XML. In addition, some of the tools can read and write relational data in other formats such as those used by UCINET [5] and Pajek [2], thus increasing inter-operability with external packages. The CASOS tools also provide a network data converter package that enables transformation between the most widely used relational formats. To further facilitate interoperability the tools were developed to be executed under the Windows, Linux and Unix platforms. For each tool a script version that runs in batch mode has been created.

4.1.3 Ontology – The Meta-Matrix

Automated ontology identification from text data is still in its infancy. Thus, we decided to use an ontology derived from organizational theory that is referred to as the meta-matrix [7][12][24]. The meta-matrix is a multi-mode, multi-plex approach to

[2] Although not demonstrated in this paper, ORA can also import and export data to analyst notebook, a commonly used tool in forensics.

organizational design. Socio-technical systems are represented using the entity classes: Agent, knowledge, resource, task, organization and location. Any two entity classes and the relations among the elements in each entity class form a network, e.g. a social network, membership network (agent by organization), or knowledge network (agent by knowledge), among others. Between any two entity classes multiple relations can exist. For example, among agents there can be relations such as “is related to” and “receives instructions from”. 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 organizational change in diverse contexts [15][21][11].

4.1.4 XML Interchange Language – DyNetML

We have developed DyNetML [37], an 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 an arbitrary number of complete, multi-modal meta-matrix networks in one file. Each network consists of an arbitrary number of node sets, which group together nodes of the same entity class, and a graph, which consists of a set of edges that connect nodes from the respective node set. Nodes and graphs can be enriched with properties, such as attributes and measures. Arbitrary numbers of node sets and graphs as well as the ability to add rich data to any object

within the hierarchy affords analysts with high flexibility and enables the representation of complex data sets within one file.

DyNetML is an open source format for the interchange of relational data that has evolved as users within and beyond CMU have augmented the language.

By handling all data in DyNetML format and using the same ontological model for real and simulated data, the integrated CASOS toolset eases validation, model tuning and creation of partially artificial data sets.

4.1.5 Data Storage and Management

For storing and managing relational data we have developed a database referred to as NetIntel database [36]. The fundamental goal when designing the database was to provide a common SQL database that handles meta-matrix networks and uses DyNetML as input and output format. Extraction capabilities were designed and implemented to create data extractions on the fly, execute complex SQL queries, support data import from data gathering tools and data export to other tools.

The database allows for graph-theoretic computations including recursion, which enables graph traversals within the database. The databases' structure is designed in an extensible manner, allowing for the easy addition of new types of nodes, edges and attributes.

When designing the database we considered the fact that the data may come from various sources and therefore might show variations in the spelling of, for example, the names of people and places. To consistently normalize 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, spelling is checked and normalized if applicable. A drawback of our system is that the data populating the thesaurus 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 such as AutoMap, and can therefore capitalize on the manual work that was invested a priori in the creation of thesauri. Utilizing the same thesauri for both data coding and storage, minimizes potential errors in the results.

Data management in NetIntel is enhanced by denoting the source of the 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 data sets while preserving the original data sources. This design choice allows for a future functionality for weighting the confidence in the source.

It is often necessary to extract subsets from data sets; for example, when analysts want to look at ego-networks or a certain set of nodes. NetIntel supports the extraction and deletion of a node (set) and its directly related edges. Subsets of the network can be extracted based on graph-theoretic properties of the network such as distance (e.g., “Find all nodes at a graph distance of two or less from a given node”) and density (e.g., “Find all nodes embedded in subgraphs with a 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 data set or any subset. The key issue in dealing with time is distinguishing the

data of the source from the dates mentioned in the actual text such as the dates of events and actions. 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.

The use of an SQL database and associated tools can afford analysts with the ability to select and analyze data of interest only, which reduces processing time. It also facilitates the comparison of alternative data sets which enables cross-cultural comparisons and facilitates a more systematic learning from the past. Storing aliases along with the respective identifiers and storing attributes of nodes facilitates data fusion and promotes more in-depth analyses such as linking psychological and structural information. This makes possible new understandings, such as possibilities for influencing selected individuals.

4.1.6 Scalability

Currently we are in the middle of a major effort to enable the tools to scale. At the moment all tools can handle networks with up to 10^6 nodes and 10^7 relations. However, in the case of the simulation tools, networks of this size take days to be simulated. The data coding tool can code an unlimited number of texts; however, each text needs to be relatively short (a few megabytes). All features and measures in the statistical toolkit, except for 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. When various entity classes are not available, the tools make use of what is available. Initial

assessments of the robustness of the measures suggest that even with 30% of missing links, most DNA measures still tend to get approximately the same rankings [19]. Much more work is called for in this area.

4.2 Toolchain for Covert Network Analysis

In the following we present a toolchain - the integrated CASOS toolset - developed and used to support the assessment of covert networks such as those described by Berry [3]. In the presentation we focus on methodological aspects rather than providing the reader with a step by step guide to using the tools. The goal is to show the general power of using a toolkit that meets the identified requirements and capabilities, particularly a toolkit with embedded toolchains. To orient the reader, Figure 2 illustrates the workflow in a typical DNA process. As can be seen, the toolset enables analysts to move from raw texts to networks to the identification of patterns in networks to analysis of possible effects of alternative interventions. This is accomplished by using tools for data coding, statistical network analysis and computer simulation. Secondary tools such as those for visualization and data editing provide supportive functionalities.

[Insert Figure 2 here]

4.2.1 Data Coding

Information about covert networks that is relevant in the context of homeland security is often contained 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 electronic texts are necessary.

In principle, data coding entails three steps. First, collect texts; second, convert the texts into a format that an automated coding tool can read; and third, run the coding tool.

AutoMap, a component of the integrated CASOS toolset, supports the third functionality [17]. AutoMap is a statistical network text analysis system [9][28][3] 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 [15]. The software facilitates the fusion of the networks from diverse texts into a single meta-network. AutoMap takes in raw texts as input and outputs relational data in DyNetML and other formats such as DL.

Texts are converted into networks using a distance based approach also referred to as windowing [14]. Windowing 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 analyst. 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 texts in the form of semantic networks or maps [9]. The process is illustrated with detailed information on coding rules provided in section 5.2.

The quality of the text coding and the speed with which data from new contexts can be coded depend on the extent to which analysts performs pre-processing or customize the tool to extract particular concepts. Once such customization is done, texts

[3] Network text analysis is based on the assumption that language and knowledge can be modeled as networks of words and the relations between them [33]. Several NTA methods exist (for an overview see [28]), one of them being map analysis, which we have operationalized, formalized and implemented in AutoMap.

can be rapidly converted to networks that can be further analyzed with various statistical network analysis toolkits or processed by various data-mining or machine learning tools.

4.2.2 Statistical Network Analysis

Common statistical network analyses include the location of key actors, hidden groups and points of vulnerability. These features are provided by statistical and machine learning tools that extract and interpret patterns in relational data. ORA is a statistical analysis toolkit that, unlike the majority of social network analysis tools, makes use of the meta-matrix ontology and facilitates the analysis of rich, multi-mode, multi-link relational data [13]. Thus, 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) [13]. 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 uses DyNetML as its IO format.

There are a number of advantages to a statistical network analysis tool like ORA that originate from implementing the meta-matrix. The analyses possible are more comprehensive and provide greater insight into factors that drive behavior in comparison to tools that are restricted to operate on the agent level. The types of analyses supported with ORA include:

- Identification of weak and strong agents 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 an organization with other organizations, random networks of equivalent size, or the same organization after an intervention or at a different time point.
- Identification of the sphere of influence surrounding specific agents or organizations.

One of the key uses of statistical network analyses is to identify possible courses of action and their immediate impact. An example would be the identification of emergent leaders in a group as possible targets for misinformation or diplomatic efforts. The immediate impact of such actions can be assessed by analyzing 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. Traditionally behavioral interpretations were drawn from a representation of the network at a particular point in time. The ability to look at how the network might evolve depended on the analyst's ability to think in multiple complex dimensions. In general, analysts can perform this task in their heads for not more than two dimensions and only a few time periods. MADN systems in contrast are able to assess the dynamics of complex, non-linear systems for many time periods and therefore can make systematic forecasts of changes in these systems [10][8]. Simulation tools provide analysts with a decision aid for thinking through the

complexities of changes in networks in response to various interventions. Analysts use computer simulations to engage in various “what-if” scenarios to predict what will happen, reduce surprise and understand the scope of the realm of what is likely to happen.

Unlike in traditional economic models, the agents in MADN simulations act in a boundedly rational fashion [32]. Based on their mental models they emulate what other people might do. In MADN systems the actions performed by individual agents lead to changes in the underlying networks that then affect 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 an agent’s knowledge network which in turn leads to changes in its social network.

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 compared to empirical data. An example of a parameter that can be set is the rational for interaction. Empirical results suggest that people spend 60% of their time interacting with people to whom they are similar. In simulation models multiple mechanisms for choosing an interaction partner exist and these can be prioritized using empiric findings. Studies also show that people's knowledge of each other decreases exponentially with the increase in social distance between them [23]. This finding has been used for DyNet, a simulation model used in the integrated CASOS toolset: The cognitive accuracy of each agent’s model of others decreases with increasing social distance, even when agents are initialized with perfect knowledge.

However, analyst's key purpose in using MADN is to ask "what-if questions" based on empiric relational data. For simulation tools developed at CASOS such as Construct [31], DyNet [6], and NetWatch [35], the integration of the meta-matrix model into the simulation models facilitates the development and tracking of models in which different entity classes of the meta-matrix evolve differently. For example, while agents can learn a piece of knowledge and so create a connection from agent to knowledge, knowledge cannot learn about agents. Furthermore, all of these tools can read DyNetML files that represent real socio-technical systems and then evolve these systems over time and write the results out in DyNetML. The evolved networks can then be read into ORA to compare them with the original data in order to evaluate changes, or with a known later state of the original network for validation purposes. The simulation tools are designed to generate missing segments of data using statistical profiles. This is crucial for performance evaluation of simulations because it enables side-by-side comparisons of results from simulated and real data. The interoperability of simulation and analysis tools increases the usability of the involved tools and provides analysts with a powerful system for evaluating potential effects of diverse operations.

Given the current state of simulation, the results are most appropriate for the "relative" evaluation of different operations: If a MADN computer simulation predicts that a particular operation will reduce suicide bombings by 10%, one cannot necessarily count on the 10 %, but one can count on the fact that the number of suicide bombings is likely to decrease. If the same model predicts that for operation A suicide bombings decrease by 10% and for operation B they decrease by 20%, then one can count on the fact that the second operation will be more effective than the first one.

In this paper, to illustrate the value of MADN simulations, we will use the DyNet model. 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 respective tool captures the variability in human and organizational factors of groups under diverse socio-technical and cultural conditions. DyNet places the constructural model [8][10] [31] in an information awareness context and enables analysts to explore alternative destabilization strategies 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. In general, such interventions might be targeted misinformation, resource isolation or structure alteration. Then the same DyNetML file that was used as input for ORA is loaded into DyNet to initialize the system and the alternative intervention strategies. Next, a virtual experiment is run, the resultant evolved networks are stored in DyNetML and can then be read into ORA for further evaluation.

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 of many tools are the inability to handle data over time, the inability to meaningfully visualize networks over 1000 nodes, and the inability to visualize large data sets with multiple entity classes.

We do not purport to resolve these issues, just 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. ORA, for example, calls up the SocialInsight visualizer internally and is also able to export data in other formats to other visualizers.

5. Illustrative Application of CASOS Toolkit

We now illustrate the potential of a toolkit in which sets of tools can be linked into toolchains. A case scenario for this exemplary demonstration would be that an analyst has to evaluate various courses of action for an area in that religious or politically motivated acts of hostility are estimated to be likely to happen in the near future. There is no relational data set available. The analyst needs to move from a set of given raw texts with information about the socio-technical systems in the respective area to a what-if analysis in a short period of time.

5.1 Network Data

The analyst is faced with a covert network data set that consists of 368 texts. We refer to this data set as MidEastIV. Of these texts, 158 were collected through LexisNexis Academia via an exact matching Boolean keyword search. Search terms for all texts were the names of 109 people and groups a priori identified by subject matter experts (SMEs) to have been of critical importance in the Middle East region over the last 25 years. Note that data coding was not restricted to these 109 entities. All other individuals and organizations that appeared in the data were also coded. The media searched with LexisNexis included major newspapers, magazines and journals. The articles most relevant according to the LexisNexis sorting function were selected. Sources for the 210

other texts were open source web sites, trial transcripts, scientific articles and excerpts from books. The same search terms were used. Articles that seemed most relevant to independent researchers were selected. The time frame of MidEastIV ranges from material released between 1977 and 2004. The data set contains 17,792 unique terms and 179,702 total terms. The number of unique concepts considers each concept only once per corpus, whereas the number of total concepts also considers repetitions of concepts per text.

In general, the credibility of the coding samples can be increased by using a large corpus that integrates various texts types from a variety of sources. In MidEastIV 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 such as media coverage and observed accounts of the network such as ethnographic summaries and scientific reports. 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 augment the networks extracted from 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 [25] and the Tanzania Embassy bombing data [11].

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

To extract the social structure of the given system 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, acts of hostility and names of individuals and groups under surveillance are part of the domain knowledge peculiar to the discussion of covert networks, whereas words such as “and”, “Star Trek”, and “Johnny Bravo” are not. 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 can involve four techniques [17]: Named-Entity Recognition, which retrieves proper names such as names of people and places, numerals, and abbreviations from texts [26]; stemming, which detects inflections and derivations of concepts in order to convert each concept to its respective morpheme [22]; deletion, which removes non-content bearing concepts such as conjunctions and articles from texts [17]; and thesaurus creation and application, which associates specific 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 [15].

For deletion we built 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 the instances of relevant named entities or ideas, aliases and misspellings with a respective canonical form. For example, Al-Mohsen, Abd Al-Mohsen and Abu Hajjer (all names of one and the same person) were translated into the single-worded concept 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 3,150 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. For that purpose we built and applied a meta-matrix thesaurus. The first column of Table 3 provides quantitative information on the meta-matrix thesaurus. Note that the ontology module in AutoMap is flexible such that entities can be added to the meta-matrix ontology or completely different ontologies can be used. While it is fairly mechanical to create a delete list, thesaurus creation requires domain knowledge.

To illustrate the meta-matrix text analysis technique we code a portion of a sample article [27] from MidEastIV as a meta-matrix network. The following are the names of agents in the considered network 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 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 needs to specify the statement formation rules that determine the proximity of terms 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 [17]). This approach enables the analysts to determine the "sense" of proximity according to their research goals. 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 by running a Sub Matrix Text Analysis in AutoMap [15]. This technique distils networks as requested by the analyst from the meta-matrix in order to analyze them in detail. A single Sub Matrix Text Analysis could consider, for example, a membership network to find out which agent is affiliated with what organization, a knowledge network to study which agents knows what, and an organizational assignment network to detect what organization is associated with what tasks.

We coded the MidEastIV corpus using a window of size 4 and rhetorical adjacency.[4] With this setting we best covered the statements that hand coders were finding. The results from applying this coding schema to the data are shown in Table 3.

[Insert Table 3 here]

Statements between meta-matrix entities were formed from eight entity classes of the meta-matrix. The entity classes 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 unique statements considers each statement only once per text, whereas the number of total statements also takes repetitions of statements into account. Maps extracted with AutoMap are digraphs in order to adequately represent the inherently directed structure of texts. Therefore the lower and the upper triangle of the meta-matrix are not necessarily symmetric. Across the data set, 8,394 unique ties and 19,701 total ties were identified. Figure 4 provides an overview on the distribution of the total edges across the meta-

[4] 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 coding 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 as 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 while maintaining the original distance of the translated terms (rhetorical adjacency).

matrix. Note that connections between and among roles and attributes were not considered for Figure 4, so that the figure represents a total of 13,465 edges.

[Insert Figure 4 here]

5.3 Analyzing Dynamic Social Networks

We extracted one network per text in DyNetML format. The networks were stored in the NetIntel database and further enhanced if nodes in the networks generated in AutoMap matched nodes that were already present in the database and had links to other nodes. We also enriched the data with information provided by SME's on whether an agent was a conservative or a reformist. Then a new DyNetML file was extracted from the database and loaded as input into ORA. Note, one could also go directly from AutoMap to DyNetML to ORA.

ORA can be used to visualize the network and generate reports. In order to gain a quick overview on networks, analysts often first visualize them. This can be done with the SocialInsight visualizer from within AutoMap or ORA, or by converting the data into an other format and visualizing it with external tools, such as NetDraw [4], as was done to produce Figure 5 for the social network from MidEastIV. Several features of the visualized network stand out: First, the nodes on the left are isolates – agents who are not directly linked to other agents. 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 1996.

[Insert Figure 5 here]

Next, analysts might ask "who is critical?". The report germane to this question is the ORA Intel report, which provides network analytic measures relevant to the Intel

domain. The Intel report identifies key actors and organizations including limited interpretation of the results. Table 4 shows the part of the Intel report that contains the top five individuals in the given system with respect to measures that determine an individual's prominence or importance in the system. The table 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 exlucded from the organization 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) and is likely to bring as many or as large disjoint groups together as Kharazi would. Were Rafsanjani and Kharazi to work in opposition, the system as a whole could become slightly unstable.

The Intel report also contains information on key organizations in the MidEastIV, some of them shown in Table 5. High degree centrality and large number of members suggests the Islamic Revolutionary Guard Corps and the Guardian Council are the most dominant organizations in the system. The Islamic Coalition Society, however, is the group most likely to connect other groups (highest ranking boundary spanner), but only slightly more so than the Guardian Council.

[Insert Table 5 here]

The Intel report also provides information on the overall nature of the system under consideration – such as its density (0.005 for the MidEast IV). Mathematically, density ranges from 0 to 1, with higher values indicating denser networks. The system under consideration is therefore a very sparse network.

In order to support analysts in putting the measures from the Intel report in a broader context, thus supporting 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) and makes comparisons to values on other networks stored in the NetIntel database.

[Insert Table 6 here]

The data in Table 6 indicates that the network from the MidEastIV data set is much sparser compared to other data sets. This suggests that either there is significant missing data or the system is structured extremely differently from other systems. The value of context information for analysts is to answer questions like "how strong is a .2?".

Recall that Mohammad Khatami stood out as a key agent on almost all dimensions. In Figure 5, he appears in the lower right area of the graph, being highly connected to other people. A question that an analyst might ask is how this individual could be influenced, and whom does he influence. To examine this, we look at the sphere of influence around Khatami (Figure 6). The sphere of influence is the meta-matrix extension of the ego-net. In a standard social network the ego-net is the set of others that an ego is connected to, including the connections among the others. For meta-matrix network analysis this concept has been generalized to the sphere of influence – the sets of nodes from all entity classes that are directly connected to the ego and the connections among them.

[Insert Figure 6 here]

Analysts might also be interested in the immediate impact of a change in the network. This question can be addressed by picking an action and then comparing the network before and after an action has been applied. For example, let's assume that the top five individuals in cognitive demand (the emergent leaders) are being provided with false information. The immediate impact can be seen in changes in the overall network metrics such as the estimated performance or speed of information diffusion.

For the MidEastIV data, when the top five emergent leaders are excluded from the network, 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 leaders. This suggests that the system may enter a fragile state in the case of absence of the original leaders. Furthermore, by comparing the meta-matrix before and after disregarding the original emergent leaders, we find that this change should drop the system's performance by 4% and increase the rate of information diffusion by 67%. This suggests that in a case of an intervention that it might be effective to follow up on a combination of separating the original emergent leaders from the network and information operation.

5.5 Simulating Dynamic Social Networks

To illustrate the use of a MADN computer simulation as part of a toolchain we will use DyNet. The DyNetML file for the MidEastIV data is used to instantiate DyNet. A simple virtual experiment is run with the following strategies: Doing nothing, excluding Khatami, excluding Khamenei, excluding the five most central reformists and conservatives from the data. This virtual experiment was motivated by the results from the Intel report. Each of the conditions was run multiple times in Monte-Carlo fashion.

Within DyNet, the social and knowledge networks co-evolve as individuals interact. This process enables the social network to recover from destabilizations such as attacks. Imagine that agent A and B only interact through agent C. Imagine further that agent C got separated from the network. Over time, through a process of introduction and learning agent A might start to interact with B, or another agent, D, might start to operate as a point of connection between A and B. With either strategy the network dynamically heals itself.

Results from simulations indicate probable ways in which networks will evolve. Given that four of the five intervention strategies are designed to destabilize networks, the results can be thought of as the likely relative strength of each of these operations in destabilizing the network. In addition, the simulation results suggest what might be additional near term impacts of such destabilization operations. Note that we can use ORA to assess the immediate impact of changes such as alterations. The simulation tools let us account for the ability of the network to heal itself, regenerate or add connections in the near term. Hence, the simulation results facilitate explorations of the possible near term changes.

When the virtual experiment in DyNet finishes, the tool outputs a DyNetML file for each of the five conditions. These DyNetML files contain a complete snapshot of the evolved MidEastIV system after 50 time periods. These files can be read into ORA and compared with the original state of the organization. One key thing to explore in this venue is what new relations among nodes are likely to emerge. Thus, near term changes can be assessed by analysts using ORA measures. In addition, DyNet directly outputs changes in information diffusion, organizational performance and agent's beliefs. In

Figure 7, the results of this virtual experiment for information diffusion are shown. As can be seen, any isolation strategy results in delays of information dissemination. The largest impact, however, results from excluding Khatami. This is due in part to the intricate ways in which he is interconnected in his role as an information conduit. His profile is so unique that it is difficult to replace him. However, when he along with reformists is disregarded, the impact on diffusion rates is smaller. 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 can act as information gatekeepers.

[Insert Figure 7]

As noted, DyNet can also be used to assess change in beliefs 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 affected the overall level of conservativism (Figure 8). The results indicate that regardless of the intervention, this population is becoming more conservative. Without Khamenei or the top conservatives the move to conservativism is slowed down but not halted, and as more reformists are isolated the trend toward conservativism is exacerbated.

[Insert Figure 8]

6. Discussion

Independently the tools presented in the illustrative example are valuable for addressing core analytical questions. Collectively they provide analysts with the ability to move back and forth between data at various stages of the analysis process, to compare real and

virtual data, and to receive powerful analytical support in thinking about complex, dynamic socio-technical systems.

Each of the individual components has strengths and weaknesses. For extracting networks from texts we used a network text analysis approach implemented in AutoMap. The current operationalization leaves room for the following improvements: First, associated with each specific entity are a number of possible attributes, such as a person's gender or age, and roles, such as individuals' professions or formal positions in organizations. 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 and roles for 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 connection from mixing explosive materials to knowledge of bomb making. Third, pronouns such as he or she are not automatically replaced by the named entity to which this pronoun refers. As a result, ties are lost. In this case, coding could be improved by performing anaphora resolution prior to text data pre-processing.

Finally, many specific entities and connections among them are difficult to assess when a single data source is used. As data is becoming obtainable in a larger variety and

quantity, there will be a decrease in the scarcity of the data. For example, the use of multiple texts will enable a better coding of leadership roles, identifying further roles, and eliminating roles that are irrelevant in a given context (e.g., media conventions). 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 for inference of relations and entities.

AutoMap uses a statistical approach to textual processing coupled with a secondary mapping of the semantic network using an ontology. This is valuable for DNA. As the analysis of the Middle East data illustrates, computer-assisted text coding facilitates systematic analysis and rapid coding of 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 on an ontological and linguistic level.

One final note on text processing is that 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, more systematic, and requires fewer analysts than manual coding. This basically frees up the analysts time to focus on analysis and interpretation.

As a statistical network analysis tool we presented ORA. ORA's core advantage is that it enables analysts to examine multiple networks, multi-mode and multi-link networks. We argue that from a management and intervention perspective measures

utilizing multiple cells of the meta-matrix has more predictive power than measures that consider connections among one class of entities. For example, we used in this paper the measure of 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 ORA are a lack of wizards that facilitate the ease of usage for analysts; more integrated visualization, and better 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. Most studies using relational data have ignored attributes. However, as the field moves 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 such as attributes. Finally, ORA will need to be expanded to facilitate more over time analysis so that it can be used for examining historical trends and contrasting trends in real data with those predicted by simulations.

Moving from ORA to a simulation system is another point where substantial time and person efforts are required. 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. Building the input file for the virtual experiment is time consuming. Moving the output from the simulation tools back into ORA is straight forward as the output can 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.

The simulation tool used herein, DyNet, enables the evolution of networks. 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. An alternative approach is to provide multiple simulations for multiple types of problems and simply link in the tool most appropriate for solving a problem into the toolchain. We argue that in the long run the latter strategy is likely to be more advantageous than creating a single monolithic engine.

The following general issues apply to current simulation engines: First, MADN simulations tend to be relatively slow due to the computational power required for simulating network dynamics. Substantial research is needed on 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 also generates more data than can be reasonably analyzed even with existing tools. Research is needed

in data farming environments and new statistical techniques. Ultimately, such tools should 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 are the use of a common ontology to provide a classification of the underlying data, a common XML interchange language, and an extensible underlying database using a common structure. Whether future work uses the presented ontology, interchange language, or database structure will depend on their appropriateness for the tasks at hand. All three items are open source and will evolve as they are employed by diverse researchers and analysts. The open source approach supports the rapid development, integration and ease of operations in a tool chain.

Toolchains such as the one described in this paper 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 hostile or illegal activities such as suicide bombings or drug trading. Even though sometimes the argument is made that computer-supported text analysis is a poor substitute for the detailed reading and reasoning that analysts provide (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 users in the loop data testing.

Moving beyond methodological issues, we note that given the goal of understanding large, dynamic and complex socio-technical systems, the approach used herein affords analysts 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, events, etc., key insights into diverse behaviors can be gained. If we look only at the social network then the focus of attention is on hierarchies, communication and other social relations. The addition of resources makes it possible to consider issues of economics, adding knowledge enables us to investigate issues of training, learning, education and creativity. By moving beyond social networks, this inherently relational approach now has the promise of enabling effects based operation in areas as diverse as diplomacy, information, military and the economy to be assessed in a relational context. Such analyses will provide powerful insight into multiple aspects of human conditions.

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

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

[12] 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.

[13] 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).

[14] 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).

[15] 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.

[16] 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).

[17] 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).

[18] 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).

[19] T.L. Frantz, K.M. Carley, Relating Network Topology to the Robustness of Centrality Measures. Technical Report, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, CMU-ISRI-05-117.pdf http://reports-archive.adm.cs.cmu.edu/anon/isri2005/abstracts/05-117.html (2005).

[20] T.L. Frantz, K.M. Carley and J. Diesner, *An Automated Methodology for Conducting a Social Network Study of a University Faculty*. Technical Report, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, CMU-ISRI-05-106 (2005).

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

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

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

[24] 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.

[25] 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.

[26] 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).

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

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

[29] 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.

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

[31] 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).

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

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

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

[35] 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.

[36] 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).

[37] 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).

[38] 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

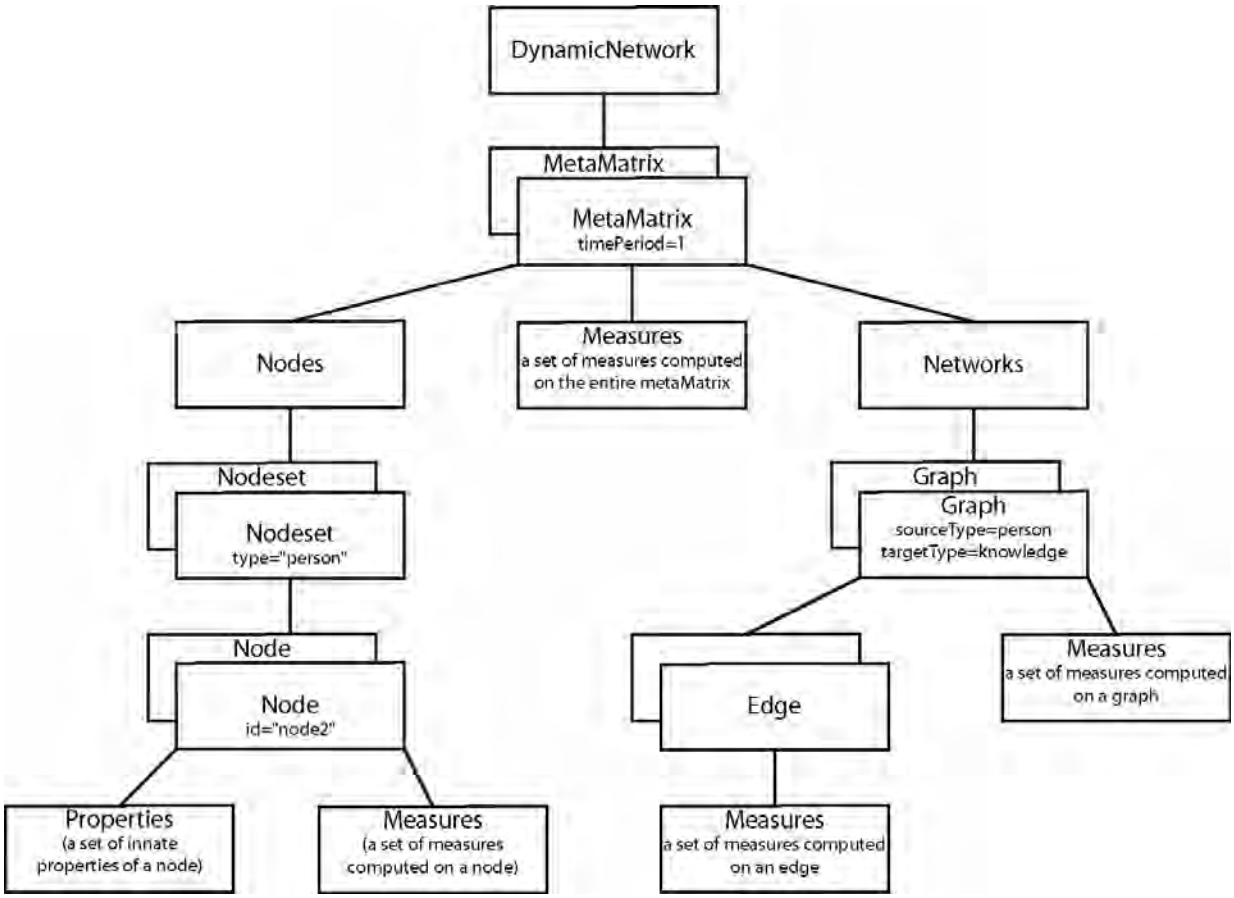


Figure 2: Workflow of Integrated CASOS Toolset

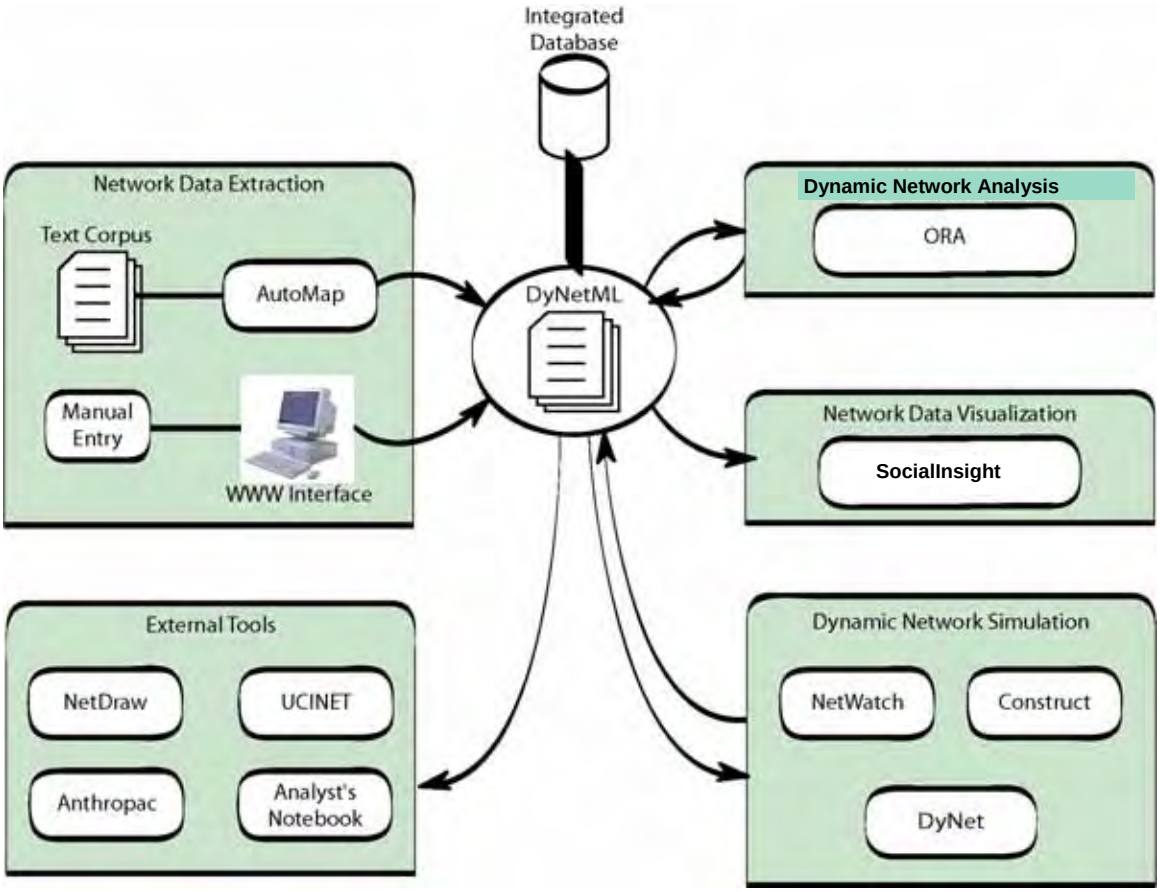


Figure 3: Visualization of sample meta-matrix network

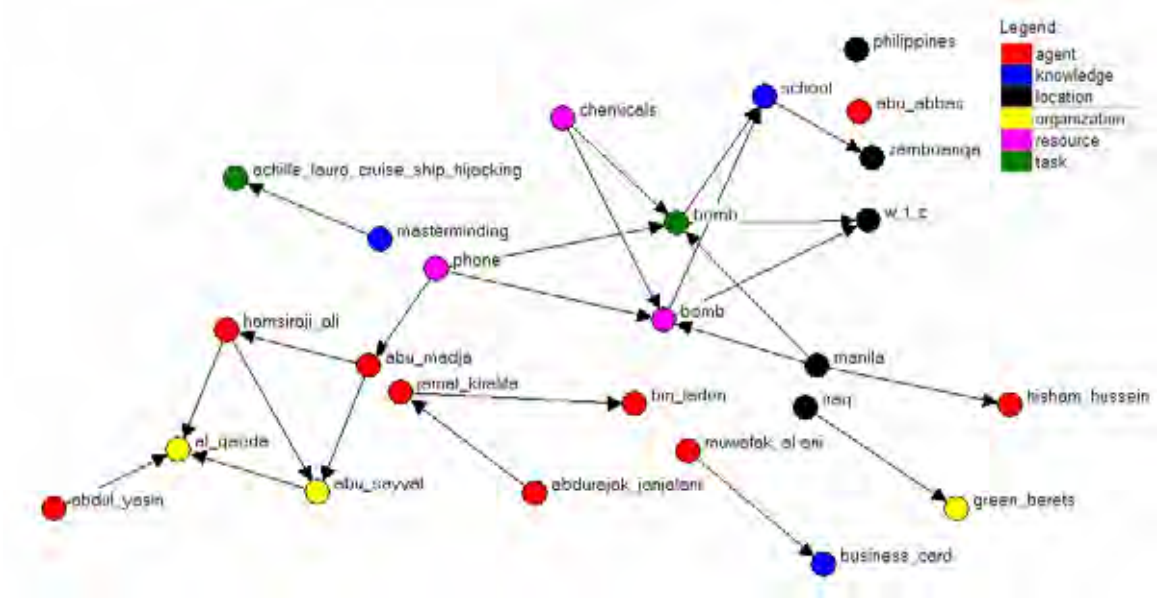


Figure 4: Distribution of edges across extracted meta-matrix

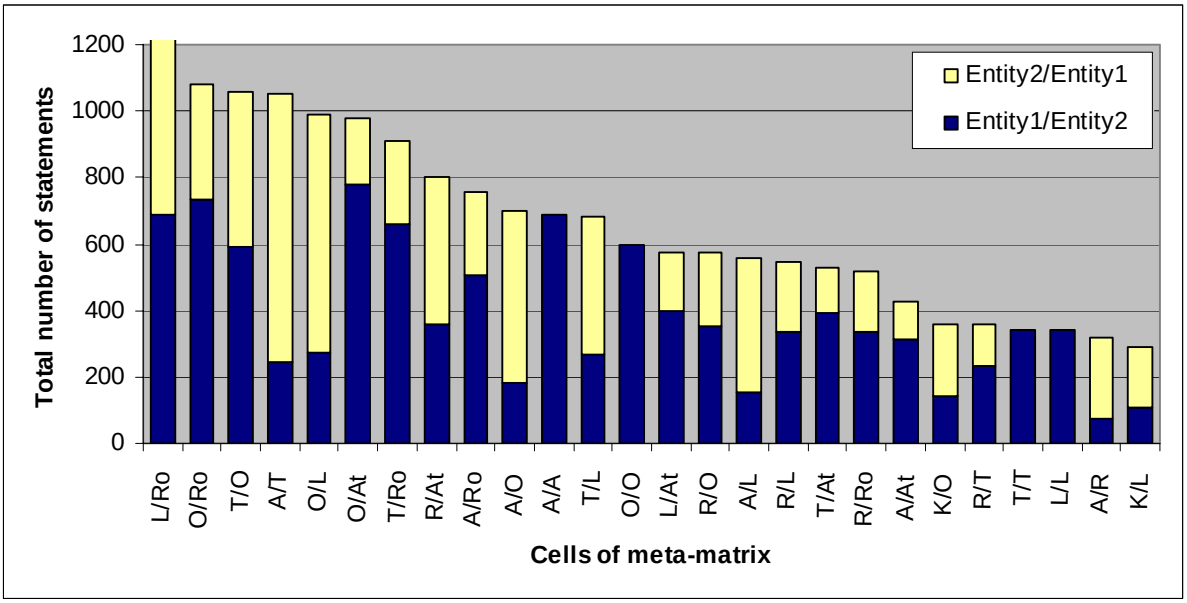


Figure 5: Social Network in MidEastIV

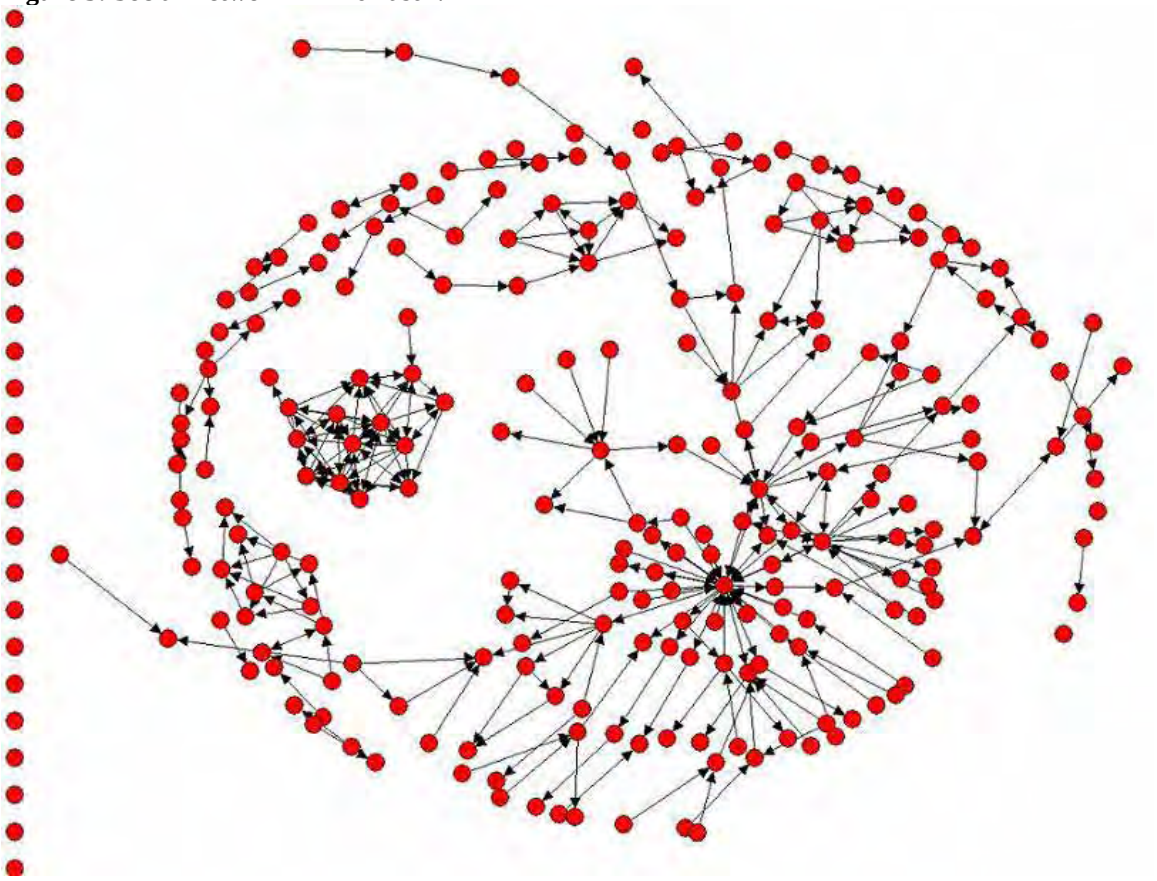


Figure 6: Sphere of Influence for Mohammad Khatami

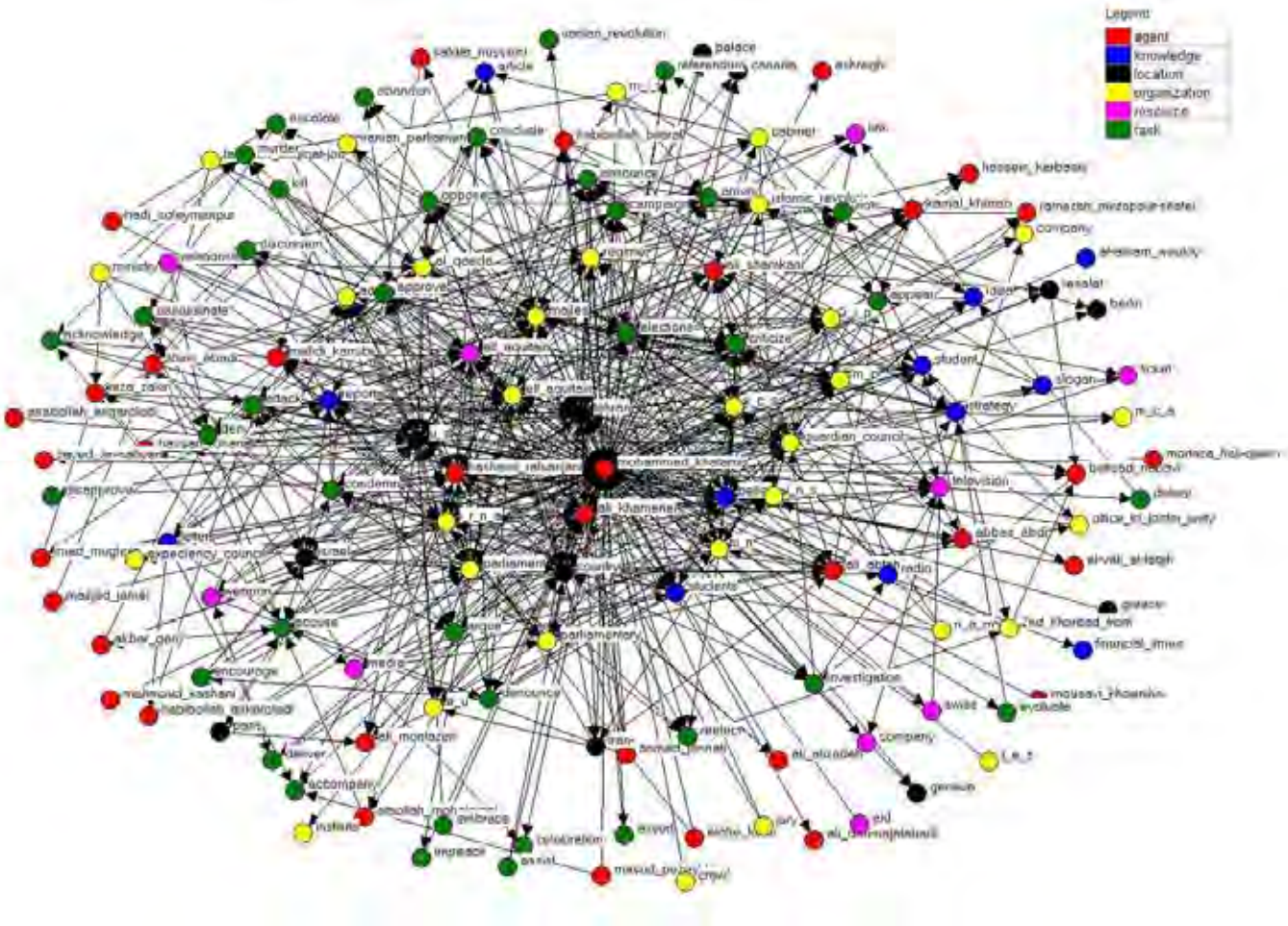


Figure 7: Diffusion Results from Virtual Experiment

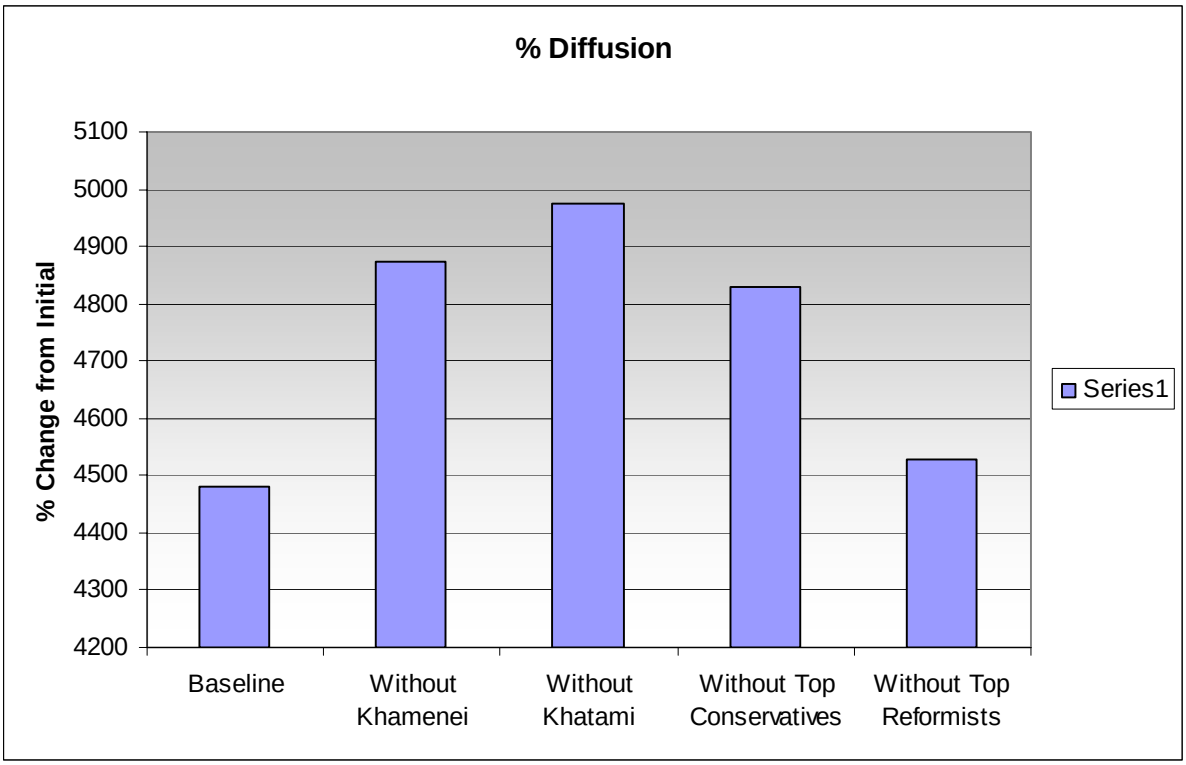
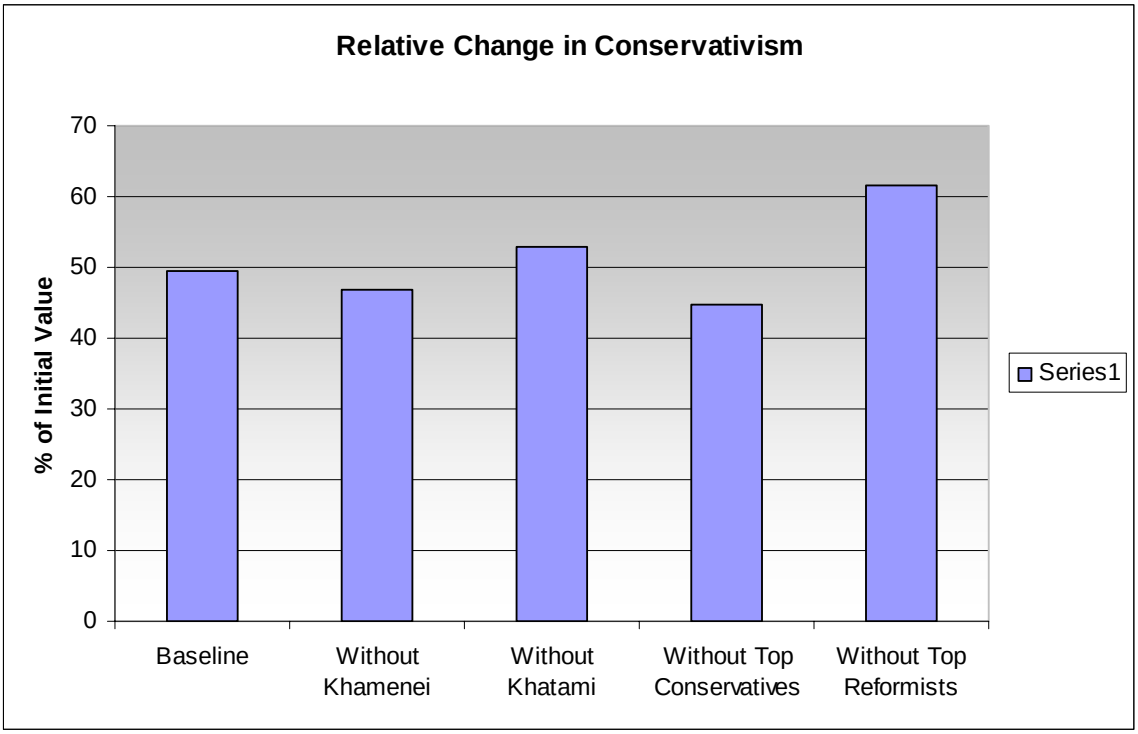


Figure 8: Conservativism Results from Virtual Experiment



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	0.06	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	0.06	Ali Khamenei		
	3	0.04	Hashemi Rafsanjani		
	4	0.02	Kamal Kharazi		
	5	0.02	Ali Montazeri		
Degree Centrality	1	0.16	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	0.10	Ali Khamenei		
	3	0.07	Hashemi Rafsanjani		
	4	0.04	Hashemi Shahroudi		
	5	0.04	Ali Montazeri		
Boundary Spanner	1	1.00	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	0.89	Ali Khamenei		
	3	0.87	Mohammad Reza Aref		
	4	0.83	Kamal Kharazi		
	5	0.57	Hashemi Rafsanjani		
Eigenvector Centrality	1	1.00	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	0.77	Ali Khamenei		
	3	0.58	Hashemi Rafsanjani		
	4	0.41	Ali Montazeri		
	5	0.40	Ahmad Jannati		
Task Exclusivity	1	0.03	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	0.01	Kamal Kharazi		
	3	0.01	Mohammad Khatami		
	4	0.01	Reza Asefi		
	5	0.01	Hashemi Rafsanjani		

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

Measure	Rank	Value	Name of Agent
Degree Centrality	1	0.21	Islamic Revolutionary Guard Corps
	2	0.21	Guardian Council
	3	0.17	Majles-e-Shura-ye-Eslami, Islamic Consultative Assembly
	4	0.13	Islamic Coalition Society
	5	0.13	Islamic Republic of Iran Broadcasting
Boundary Spanner	1	1.00	Islamic Coalition Society
	2	0.97	Guardian Council
	3	0.85	Mojahedin-e Khalq
	4	0.75	Islamic Revolutionary Guard Corps
	5	0.67	Majles-e-Shura-ye-Eslami, Islamic Consultative Assembly
Membership	1	0.07	Majles-e-Shura-ye-Eslami, Islamic Consultative Assembly
	2	0.06	Islamic Republic of Iran Broadcasting
	3	0.04	Islamic Revolutionary Guard Corps
	4	0.04	Guardian Council
	5	0.03	Atomic Energy Organization of Iran

Table 6: ORA context report

Measure	Type	MidEast IV	Other Networks	Interpretation
Centrality-Betweenness	Mean	0.00	0.05	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	0.00	0.38	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	0.01	0.17	On average this group is less cohesive than other groups.
Centrality-In Degree	Mean	0.00	0.28	On average each person in this group is connected to fewer others than is typical for other groups.
Centrality-Information	Mean	0.00	0.06	On average each person in this group has less access to information than is typical for other groups.
Centrality-Inverse Closeness	Mean	0.01	0.47	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	0.00	0.28	On average each person in this group sends information (messages/goods/advice) to fewer others than is typical for people in other groups.
Clustering Coefficient-	Mean	0.05	0.38	This group is less cohesive than other groups.
Component Count-Strong	Mean	511.00	8.50	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.
Constraint-Burt	Mean	0.18	0.32	On average people in this group are less constrained in their action than is typical in other groups.
Diameter	Mean	545.00	22.00	On average information will take more time to flow through this group than other groups.