Introduction to Dynamic Network Analysis

Geoffrey P. Morgan
William Frankenstein
Kathleen M. Carley
Carnegie Mellon University
School of Computer Science
Institute for Software Research
5000 Forbes Ave
Pittsburgh, PA 15213
gmorgan@cs.cmu.edu, wfranken@cs.cmu.edu, kathleen.carley@cs.cmu.edu

Keywords:
Network Analysis, Key Entity Identification, Grouping, Network Comparison, Network-Centric Modeling

ABSTRACT: We discuss dynamic network analysis and how it can be useful both for instantiating simulations and for novel simulation outputs. The tutorial focuses on providing attendees grounding in network analysis, allowing attendees to incorporate network data as inputs or as outcomes of their simulation tools. We first introduce key network analysis concepts and statistics including centrality measures and brokerage. To ground these concepts and provide some intuition for thinking about different types of centrality measures, we run a short hands-on analysis of network data in ORA based on intelligence from the Revolutionary War. We then introduce the data structures commonly used in network analysis and help attendees import their own data into ORA. We then introduce attendees to grouping algorithms including clique identification, Concor, and Newman-Girvan. We conclude with additional methodologies to compare networks such as QAP. These methods are particularly useful for comparing simulation inputs and outputs, and inform analysis of simulation outputs. Attendees are encouraged to bring their own data, but sample data will be provided.

1. Introduction
Network analysis, the evaluation of how nodes (including agents, resources, locations, and more) relate to one another, has proven useful in many diverse fields, from ecology, organizational psychology, and cyber-vulnerability analysis. This tutorial is designed to provide attendants with a grounding in network analysis methodology and ideas on how to incorporate networks into simulation inputs and outputs. Networks are useful ways of representing virtual proximity, limits on an actor’s social cognition (Simon, 1991), and interactions between nodes. We present both theoretical information and how to apply these analysis techniques using ORA (Carley, Pfeffer, Reminga, Storrick, & Columbus, 2013).

1.1 Why is this material important?
This material is useful to anyone trying to incorporate large data-sources into their simulation tools. Networks provide a useful and effective way of leveraging large-scale event data that is well-suited to informing agent-based models. Further, network analysis techniques may provide new ways of leveraging and analyzing existing simulation data. Finally, network techniques are becoming more common, and having some familiarity with the area should provide some utility to many attendees.

1.2 Tutorial Purpose
The tutorial intends to introduce attendants to network analysis and offers ideas of how network analysis can be used to both inform simulation and examine simulation outputs. It will present material refined yearly for a Summer Institute in network methodologies held at Carnegie Mellon since 2001.

1.3 Intended Audience
The intended audience of this tutorial includes people first learning about network analysis techniques or those who are interested in evaluating existing data via network analysis. It provides the broader theoretical context on network analysis that may be useful to these attendees while still focusing on applications. People
also interested in learning about network-centric simulation, either to use existing simulations or to incorporate elements into their own tools, will also be well served. Attendees are encouraged to bring their own data (ideally a sub-sample).

2. Contents
In the course of this tutorial, we will discuss how to identify key entities in a network, various methods of grouping entities together within a network, ways of transforming data from multi-mode (agents x locations, for example) into single-mode data, and finally ways of comparing networks.

The tutorial will include lecture and exercise based elements.

2.1 Characterizing a network and Key Entity Identification
We introduce tutorial attendees to some of the key terms and ideas used in network analysis literature, such as nodes, ties, modes and classes, and the concept of dynamic networks. Nodes are analogous to agents; they represent individuals in a network. Ties are analogous to links or edges between nodes in a network. Modes and classes highlight how networks can represent ties between nodes of the same type or class, such as a social network representing ties between agents; networks can also represent ties between different types of classes, such as membership lists of organizations. In doing so, we review standard notation for representing these networks as matrices, as well as statistics used to characterize networks.

We introduce and define network statistics characterized at the level of the network and node. At the network level, these statistics include density, link count, isolate count, component count, reciprocity, characteristic tie length, clustering coefficient, average distance, and diameter. At the node level, these statistics include centrality measures including degree, betweenness, eigenvector, and closeness. By considering these node level statistics and the type of network being analyzed, we can identify network ‘elite’ and key entities (Wasserman, 1994).

In identifying key entities, we consider the type of network being analyzed as well as the interplay of different network statistics – and highlight the role of the meta-network (Carley, 2002; 2005) in helping characterize these networks for organizations. To characterize the behavior of key entities, we consider brokerage behavior, bridges, and structural holes in networks. We utilize different centrality measures to match key entities to these behaviors and network features. Degree centrality, which considers a node’s total number of connections, can be useful for identifying resource-rich individuals. High betweenness centrality indicates an ability to connect disparate groups, but may also limit the actual paths available for action to the node due to politics. Eigenvector centrality, which rewards being connected to highly degree central nodes, indicates agents with strong social capital. Closeness centrality can indicate nodes able to utilize several different resources very quickly.

2.2 Paul Revere, Broker of the American Revolution – A motivating example
We demonstrate the power and intuition of network analysis with a motivating example of British intelligence from immediately before the American Revolutionary War (Han, 2009). In doing so, tutorial attendees have an immediate, concrete example of how to think about network analysis for their own research, as well practical experience with utilizing ORA.

We start with membership lists for five distinct groups that played a crucial role in pre-revolutionary Boston: St. Andrews Lodge, Loyal Nine, North Caucus, Long Room Club, and the Boston Committee. We also have lists of individuals who would be affiliated with the Tea Party and London Enemies – not official groups yet, but members would be placed on intelligence watchlists. This data will be included in tutorial attendee packages for easy import into ORA (Healy, 2013).

An initial analysis of this data does not reveal significant insight. However, by manipulating these membership lists into social and organizational networks, we gain insight into key entity behaviors. We manipulate the membership data by “folding” networks to infer social (agent by agent) networks and organizational (organization by organization) networks – multiplying the membership lists (an agent by organization network) by its matrix transpose.
We gain further insight and intuition into using network and node statistics by contrasting the networks generated when we consider only the networks generated by actual organization membership lists, deleting the intelligence watchlist information of Tea Party and London Enemies. We also gain insight into how centrality measures behave when computing new measures on the smaller dataset.

In using the Paul Revere data, tutorial attendees will gain familiarity with ORA’s range of features and reports, which will prepare them for using their own data for the remainder of the tutorial.

2.3 Network Data Import to ORA
After providing a motivating example of Paul Revere and his connections in colonial Massachusetts, we use this section to help participants understand the forms network data takes and to import and do basic analyses their own data.

We describe the two foundational network data representations, matrices of data (sometimes called the dense representation), and edge or link-list data (sometimes called the sparse representation). Much modern data not explicitly collected for network analysis can be most conveniently represented in link-list form, where rows from existing table-data can be used to create links.

Further, modern network analysis often finds it useful to be able to add attributes or characterizations to nodes. This can be useful in a variety of ways, including partitioning nodes and visual examination of networks colored by specific attributes.

In this section, we will teach participants how to use ORA to import various forms of network data, ideally including a sample of their own data, how to understand the ORA interface, and how to run network reports to get quick assessments of the measures discussed in the earlier section.

2.4 Grouping Algorithms
In this section, we will provide a primer in various ways groups can be identified in network data. We provide a definition of a group, reasons why grouping is useful, a brief taxonomy of grouping methods, and how we evaluate grouping algorithms. We then turn to specific discussion on Clique identification, the Concor algorithm (Breiger, Boorman, & Arabie, 1975), and the Newman-Girvan algorithm (Newman & Girvan, 2004).

Attendees will then be encouraged to perform grouping analysis either on their provided data, or on sample data provided for the purpose.

2.5 Comparing networks
Comparing networks is particularly useful to the modeler, as change in a specific network of interest could be a useful outcome of a simulated intervention. We discuss why standard statistical approaches are not suitable for comparison of network structures, since neither the independence nor identical distributions assumptions are supported.

We discuss Hamming Distance (Hamming, 1950), which is the number of differences between the two networks, when their matrices are transformed into one large vector. Hamming Distance, while useful for understanding overall differences in a network, does not provide a useful structural comparison.

We then present QAP (Hubert & Schultz, 1976) and MRQAP models (Krackhardt, 1988), which allow analysts to assess how likely one network structure is to appear as the result of another network structure. MRQAP can be thought of, essentially, as a small simulation model evaluating many instances rapidly and then providing a confidence bound on the strength of relationship between two networks.

We provide sample data that allows multiple hypotheses to be tested using QAP/MRQAP procedures, and guide attendees in how to perform this test on this sample, or their own data if it is sufficient for testing purposes.

3. Conclusion
In this tutorial, attendees will learn about network representations, common network measures of centrality and importance in network, commonly applied grouping algorithms, and comparison methods for networks. This tutorial will be useful to people interested in networks, or interested in how to leverage large-scale event-data for use in agent-based simulation.

We believe this hands-on tutorial will be useful and interesting to many modelers.
4. References

Author Biographies
GEOFFREY P MORGAN is a PhD student in Carnegie Mellon’s Computation, Organization, and Society program. Morgan has experience in developing autonomous robotic systems, self-learning systems, cognitive models, and user-assistive systems. He is interested in organizational impacts on individual decision-making and on optimizing organizational performance.

WILLIAM FRANKENSTEIN is a PhD candidate in Engineering and Public Policy at Carnegie Mellon University. His research focuses on issues at the intersection of international security and technology policy, including risk communication and risk management and their application to the field of weapons of mass destruction proliferation. He is especially interested in the application of social media analysis, simulation, and network modeling to this field.

KATHLEEN M CARLEY is a Professor and the director of the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. Her research combines cognitive science, social networks and computer science to address complex social and organizational problems.