Generation of Realistic Social Network Datasets for Testing of Analysis and Simulation Tools

Maksim Tsvetovat Kathleen M. Carley

Sept 26, 2005
CMU-ISRI-05-130

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

To appear in CMU ISRI Technical Reports Series

This work was supported in part by the DOD, and National Science Foundation (MKIDS: IIS0218466), the Office of Naval Research under Dynamic Network Analysis program (N00014-02-1-0973) and the National Science Foundation under the IGERT program for training and research in CASOS, NSF ITR 1040059. Additional support was provided by CASOS - the center for Computational Analysis of Social and Organizational Systems 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 NSF or the U.S. government.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
26 SEP 2005		**00-09-2005 to 00-09-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Generation of Realistic Social Network Datasets for Testing of Analysis and Simulation Tools	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**27**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Abstract

Testing large-scale dynamic network simulation packages such as NetWatch[34] requires a large quantity of test data to be available for each of the experiments. The test data includes initial topologies of agents' social networks and specification of knowledge networks for each of the agents to fit an empirically derived distribution of knowledge.
Testing the software on machine-generated data, as opposed to empirical data only, allows the user to conduct repeatable tests that stress certain aspects of the software and help in debugging and optimization of software performance.

Keywords: Social Networks, social simulation, scale-free networks, cellular networks, random graphs, software testing

1 Introduction

Testing large-scale dynamic network simulation packages such as NetWatch[34] requires a large quantity of test data to be available for each of the experiments. The test data includes initial topologies of agents' social networks and specification of knowledge networks for each of the agents to fit an empirically derived distribution of knowledge. Another task is creation of realistic task structures that could be used to simulate performance of complex interdependent projects by groups of agents.

The main concern in generation of artificial data is its realism. Based on open-source empirical data (such as described in sec. 4), the artificial datasets need to approximate certain qualities or parameters found in the empirical data. However, it is unclear at the outset what parameters need to be emulated to achieve highest fidelity simulation.

Frequently, theories of network topologies in a particular setting are proposed. For example, large amount of social network research relies on assumptions made by Erdös [15] regarding topology and distances in random graphs. As an elaboration of Erdös networks, small-world network topologies[37] retain many properties of random graphs, yet providing a degree of structural realism that maps to macro-level structures of social networks and communities[27] .

However, it is now clear that purely random graphs are not a good approximation of topology of social networks. Other proposed topologies include scale-free networks[3], whose role in modeling social networks we discuss in section 2.

While none of these theories has emerged as a clear winner and new ideas of network topologies in large-scale social networks are frequently published, it is important to make simulation tools independent of the models and theories of initial network topology. Furthermore, a simulation tool that is proven and validated through docking and comparison with empirical results can be used as a means to test validity of multiple theories of network topology - or test its own assumptions against all possible networks.

Testing the software on machine-generated data, as opposed to empirical data only, allows the user to conduct repeatable tests that stress certain aspects of the software and help in debugging and optimization of software performance.

As number and complexity of social network analysis algorithms grows, it becomes more and more important to test these algorithms for accuracy,

scalability, robustness. We define robustness of a measurement algorithm as a function of degradation of quality of measurements with decay of the data modelled as introduction of noise into inputs of the algorithms.

Robustness studies such as [6] and [11] measure impact of decay in random networks on accuracy of computation of standard social network analysis metrics. Such rigorous tests require large amounts of data which can be easily manipulated to introduce errors. Networks used as input to the robustness study need to span different sizes and topologies, and be easily manipulated to introduce a quantifiable amount of noise for robustness testing. This problem is much easier to solve using synthetic (generated) data, where size and topology of the network are controlled by generation functions[9]. SNA algorithms need to be then tested against multiple network topologies. Moreover, parameters of the network generator can be manipulated in a scientific fashion, thus allowing the measurement algorithm to be also tested on possible variation of the topology.

2 Terrorist Organizations and Scale-Free Networks

An argument has been made[30] that terrorist networks may exhibit features of scale-free networks and can thus be treated as such in analysis and derivation of attack scenarios.

Scale-free networks have been observed in many contexts ranging from networks of airline traffic to sexual networks and Web link patterns. The high probability of emergence of scale-free networks, as opposed to evenly distributed random networks, is due to a number of factors, including:

- Rapid growth confers preference to early entrants. The longer a node has been in place the greater the number of links to it. First mover advantage is very important.
- In an environment of too much information people link to nodes that are easier to find - thus nodes that are highly connected. Thus preferential linking is self-reinforcing.
- The greater the capacity of the hub (bandwidth, work ethic, etc.) the faster its growth.

It has also been observed that scale-free networks are extremely tolerant of random failures. In a random network, a small number of random failures can collapse the network. A scale-free network can absorb random failures up to 80% of its nodes before it collapses. The reason for this is the inhomogeneity of the nodes on the network – failures are much more likely to occur on relatively small nodes.

However, scale-free networks are extremely vulnerable to intentional attacks on their hubs. Attacks that simultaneously eliminate as few as 5-15% of a scale-free network's hubs can collapse the network. Simultaneity of an attack on hubs is important. Scale-free networks can heal themselves rapidly if an insufficient number of hubs necessary for a systemic collapse are removed.

Scale-free networks are also very vulnerable to epidemics. In random networks, epidemics need to surpass a critical threshold (a number of nodes infected) before it propagates system-wide. Below the threshold, the epidemic dies out. Above the threshold, the epidemic spreads exponentially. Recent evidence[28] indicates that the threshold for epidemics on scale-free networks is zero.

However, the reality of terrorist networks does not fit neatly into the scale-free network model. It has been observed[31] that non-state terrorist networks are not only scale-free but also exhibit small world properties. This means that while large hubs still dominate the network, the presence of tight clusters (cells) continue to provide local connectivity when the hubs are removed.

For example, attack on Al Qaeda's Afghanistan training camps did not collapse its network in any meaningful way. Rather, it atomized the network into anonymous clusters of connectivity until the hubs could reassert their priority again. Many of these clusters will still be able to conduct attacks even without the global connectivity provided by the hubs.

Furthermore, critical terrorist social network hubs cannot be identified based on the number of links alone. For example, Krebs observed[25] that strong face-to-face social history is extremely important for trust development in covert networks. Of similar importance is the relevance of skills and training of agents inside a cell to the task at hand. Thus, importance of any individual within the network should be rated on a vector of factors pertaining to its qualities as an individual as well as types and qualities of its links.

Rothenberg[31] notes that postulating a path of a set length from everyone in the global network to everyone else (i.e. scale-free nature of a terrorist network) runs contrary to the instructions for communication infrastructure set forth in the Al Qaeda training manual[1]. Thus, if a terrorist network was observed to be scale-free, it can be argued that its scale-free nature is not a matter of design and can possibly be an artifact of the data collection routines. For example, snowball sampling[19] is biased toward highly connected nodes, so extensive use of this technique may result in observation of scale-free core-periphery structures where none exist[5].

3 Developing the Formalism of a Cellular Network

Given the case studies of Al Qaeda and other terrorist networks, it is clear that terrorist organizations cannot be adequately described as random graphs or as scale-free networks. Therefore, a different model of terrorist networks has emerged, namely cellular networks [31][10][12]. While this model may

not fit a simple mathematical definition such as scale-free or small-world network, its base is in empirical and field data[18]. In section 5.3, I will show that cellular networks in fact are not characterized by a lack of a formal representation but are defined through a more complex process which takes as a goal improvement of fit between the model network and empirical data.

Cellular networks[7] are different from traditional organizational forms as they replace a hierarchical structure and chain of command with sets of quasi-independent cells, distributed command, and rapid ability to build larger cells from sub-cells as the task or situation demands. In these networks, the cells are often small and are only marginally connected to each other. The cells are distributed geographically, and may take on tasks independently of any central authority[8].

Rothenberg[31] observed a number of properties of a cellular network:

- The entire network is a connected component.

  > ...It is likely that on the local level, individual ties are very strong...On the higher level, individual ties are likely to be weaker but the strength of association [people known in common, doctrine] is likely to remain high...

- The network is redundant on every level: Each person can reach other people by multiple routes - which can be used for both transmission of information as well as material. On the local level, there will be a considerable structural equivalence[35], which will ameliorate the loss of an individual. The redundancy in communication channels may also be mirrored in the redundance of groups engaged in a particular task.

- On the local level, the network is small and dynamic, consisting of small cells (4-6 people) that operate with relative independence and little oversight on the operational level.

- The network is not managed in a top-down fashion. Instead, its command structure depends on vague directives and religious decrees, while leaving local leaders the latitude to make operational decisions on their own.

- The organizational structure of a terrorist network was not planned, but emerged from the local constraints that mandated maintenance of secrecy balanced with operational efficiency.

Each cell is, at least in part, functionally self-sufficient and is capable of executing a task independently. Cells are loosely interconnected with each other for purposes of exchanging information and resources. However, the information is usually distributed on a need-to-know basis and new cell members rarely have the same exact skills as current members. This essentially makes each individual cell expendable. The removal of a cell generally does not inflict permanent damage on the overall organization or convey significant information about other cells. Essentially, the cellular network appears to morph and evolve fluidly in response to anti-terrorist activity[32].

This leads to a hypothesis that cells throughout the network contain structurally equivalent[17] and essential roles, such as ideological or charismatic leaders, strategic leaders, resource concentrators and specialized experts.

Given this hypothesis, one can further reason that operations of a particular cell will be affected in a negative way by the removal of an individual filling one of these roles. I further posit that a further development of a cellular network formalism as an empirically driven and yet mathematically sound concept, is necessary for creation of computational models that combine face validity towards real-world data as well as veridicality towards formal models of organizational evolution.

4 Open-Source Data on Terrorist Networks

Social network datasets were extremely difficult to obtain and limited in size and scope, until recently. The prevailing methodology for collecting social network data was by survey, either administered to an entire group of people or collected in a snowball fashion. Collection of social network data was done in a way reminiscent of anthropological data collection - by a human observer embedded into an organization to be studied.

This presented a number of problems. First of all, it was very costly to collect all but the smallest of datasets. While a number of sampling strategies were investigated, it was difficult or infeasible to canvass a larger organization or population. Furthermore, presence of an observer or a survey instrument in an organization inevitably altered the behaviour of individuals in the organization.

Finally, for some networks, especially terrorist networks, it is physically impossible to collect a dataset via direct survey administration. The modus operandi of such networks is covertness and this necessarily limits the data that can be collected on them.

Thus, for study of terrorist organizations, one must obtain information via indirect means. One approach to gathering indirect social network data is via analysis of texts. Originally used as manual coding technique, text analysis has now been automated to extract network structure from corpora of text based on co-appearance of people, organizations and other entities. An example of such text coding is the representation of the Hamas network (figure 1), extracted by AutoMap from a set of documents describing organizational structure and operational constraints of the Hamas terrorist organization.

Between September 14, 2001 and November, 2001 Valdis Krebs[25] assembled a corpus of texts regarding events preceding September 11th attacks. Manual analysis of these texts yielded a dataset which became one of the definitive sources of data on terrorist organizations and structure of a terrorist plot.

Since 2001, much larger datasets on covert networks are available due to both increased interest in the research and improvements in tools for machine collection of network data.

Some of the newer more complete datasets include these collected by IntelCenter[23], R. Renfro[29] and M. Sageman[32]

In the aftermath of the September 11th attacks, it was noted that coher-

Figure 1: Data on Hamas collected by AutoMap

ent information sources on terrorism and terrorist groups were not available to researchers[20]. Information was either available in fragmentary form, not allowing comparison studies across incidents, groups or tactics, or made available in written articles - which are not readily suitable for quantitative analysis of terrorist networks. Data collected by intelligence and law-enforcement agencies, while potentially better organized, is largely not available to the research community due to restrictions in distribution of sensitive information.

To counter the information scarcity, a number of institutions developed unified database services that collected and made available publicly accessible information on terrorist organizations. This information is largely collected from open source media, such as newspaper and magazine articles, and other mass media sources.

Such open-source databases include:

- RAND Terrorism Chronology Database[14] - including international terror incidents between 1968 and 1997
- RAND-MIPT (Memorial Institute for Prevention of Terrorism) Terrorism Incident Database[21], including domestic and international terrorist incidents from 1998 to the present

- MIPT Indictment Database[33] - Terrorist indictments in the United States since 1978.

  Both RAND and MIPT databases rely on publicly available information from reputable information sources, such as newspapers, radio and television.

- IntelCenter Database (ICD)[22] includes information on terrorist incidents, groups and individuals collected from public sources, including not only traditional media outlets and public information (such as indictments), but also information learned from Middle East-based news wire services. Separately, IntelCenter also collects information from Arabic chat-rooms and Internet-based publications - although value of such data is questionable and data may be tainted by propaganda.

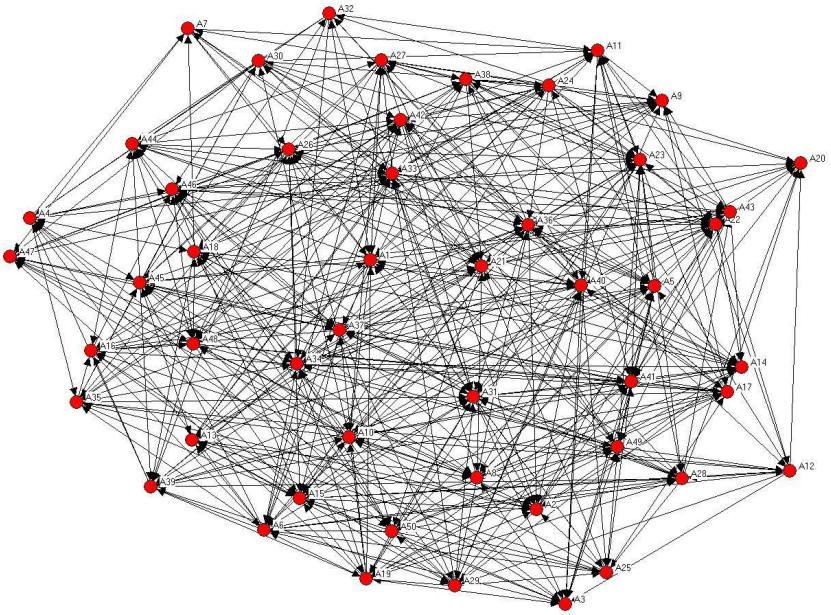


Figure 2: A Uniform Random Network

5 Generating Person-to-Person Networks

5.1 Erdös Random Graphs

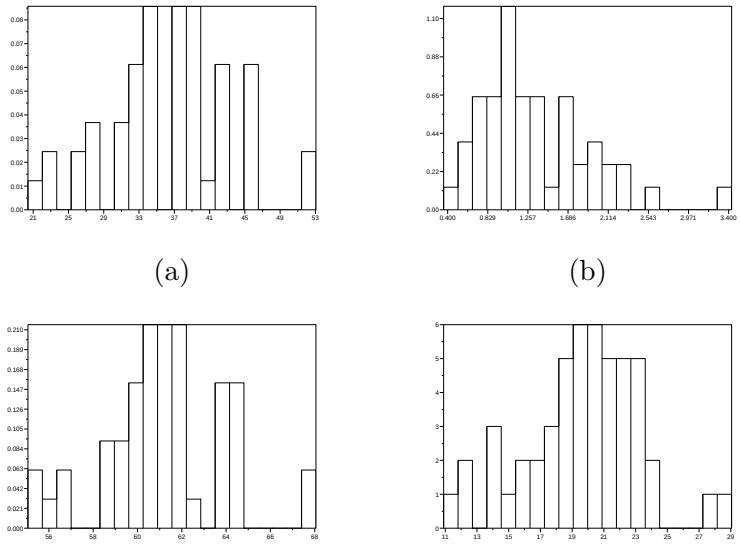
The study of random graphs dates back to the work of Erdös and Rényi whose seminal papers[15][16] laid the foundation for the theory of random graphs.

There are three standard models for Erdös random graphs[2]. Each has two parameters. One parameter controls the number of nodes in the graph and one controls the density, or number of edges.

For example, the random graph model $G(n, e)$ assigns uniform probability to all graphs with n nodes and e edges while in the random graph model $G(n, p)$ each edge is chosen with probability p.

5.2 Scale-Free Networks

One of the most interesting features of a large class of the complex networks under study now is their scale-free behavior: each node of the network is



(a) (b)

(c) (d)

Figure 3: Distribution of centralities in a Erdös random network: (a)Degree, (b)Closeness, (c)Betweenness, and (d)Eigenvector

connected to some other k nodes. The number of connections obeys a power-law distribution, i.e. $P(k) \sim k^{\gamma}, 2 \leq \gamma \leq 3$ for most networks considered.

Such networks are dubbed "scale-free" because the fluctuations of the distribution around the average value k are infinite (they do not possess any particular scale). The difference between a scale-free network and a random network (where every link between different nodes is present with a probability p, resulting in a Poisson degree distribution) hints towards some mechanisms that generated the observed network features. One of the most celebrated models that explains the emergence of scale-free networks is the Barabasi- Albert (BA) model[4].

According to the BA model, the two essential ingredients for the formation of scale-free networks are growth and preferential attachment. Growth implies that new nodes are added to the network over time at a more or less constant rate. Preferential attachment means that a newly added node connects preferentially to nodes that already have a high degree: a new node tries to attach to authoritative nodes and the degree of a node is an effective representation of its authoritativeness. It has been shown that, if the proba-

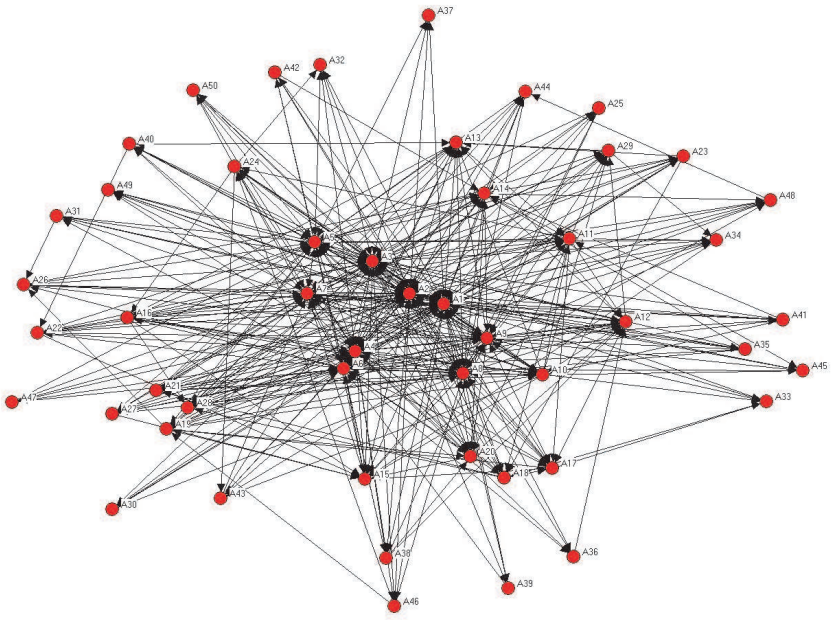


Figure 4: A Scale-Free Network generated by preferential attachment

bility to connect to a site is linearly proportional to its degree, then growth and preferential attachment indeed generate scale-free networks[24].

5.3 Cellular Networks

The above-mentioned algorithms for generating simulated organizational data can be summarized as creating an approximation of real social phenomena (i.e., organizational structure) by means of an analytically solvable function or a statistical mechanism.

Below we present an alternative approach, which relies on the observations of organizational structure of extant covert networks via creation of a network profile.

We define a generative network profile as a collection of observations and measurements that, when taken together, can be used as a generative function for creating networks similar to ones observed in the real world.

The method of generating simulated organizational structures from profiles should be generalizable to many different types of organizations. How-

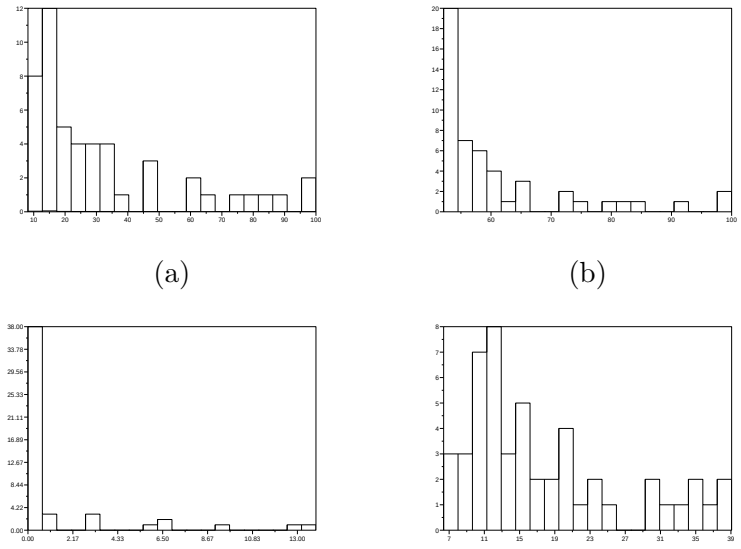


Figure 5: Distribution of centralities in a scale-free network: (a)Degree, (b)Closeness, (c)Betweenness, and (d)Eigenvector

ever, for every type of organization the components of a generative profile would be different.

In this section we present a generative profile of a cellular covert network based on the publicly available dataset on September 11th hijackers[25].

Based on publicly available data collected by Krebs[25], the following profile of the structure of covert networks has been derived [12]:

- The network consists of small cells (mean cell size of 6 members) with very low interconnection between cells.
- Internally, the cells exhibit dense communication patterns.
- There is a very low probability of two individuals communicating by chance (0.007).
- The probability of triad closure (link from x to y being more likely if both x and y are linked to third party z) is 0.181.

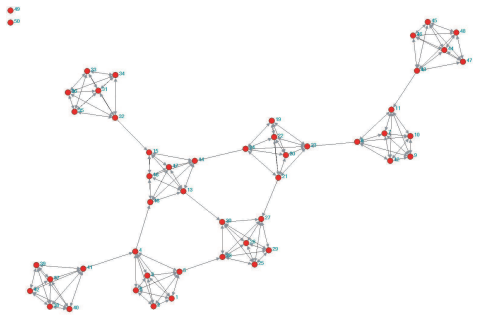


Figure 6: Red Team: A Cellular Covert Network

- Senior members of each of the cells are often also parts of other cells and interact with other senior members on the network.
- Cell leaders are more knowledgeable than other members.
- Cell members share an ideological doctrine but also specialized knowledge (i.e. bombmakers, drivers, operatives).
- Cells use information technologies and electronic communication.

The aforementioned parameters form a statistical profile from which we can generate simulated organizational networks. The plot on figure 6 shows a covert network generated using parameters specified above.

The algorithm for generating a network based on the above profile is represented in listing 1

6 Generalization and Optimization of Network Profiles

At this point, the choice of profile components lies in the hands of the researcher and creation of a profile is a manual task. However, creation of such profiles can be represented as an optimization problem.

Creation of general-purpose generative profiles can be done with using the following assumptions:

- Let the network consist of a finite number of layered groupings. For example, a corporate network may be viewed as a collection of (a)people,

Listing 1: "Generating Cellular Networks"

```
//Generate Cells
CREATE cells with
  cell_size()=normally distributed random variable
     (mean=average cell size, std.dev = 0.17*mean);

//Assign agents to cells
FOR all agents DO
  current_cell=random cell
  IF current_cell is not full THEN
    assign an agent to current_cell
  ELSE pick a new cell; repeat this operation.
  END IF
END FOR

//Fill in connections inside cells
FOR all cells DO
  PICK a random agent inside the cell to serve as a leader

  //Internally, generate a uniform network
  FOR all agents inside the cell DO
    generate links within cell with the given density
  END FOR

  //Bring the probability of triad closure in line with the
     measurements
  IF probability of triad closure significantly less then
     measured value
    Add a small random number of edges; repeat the measurements
  ELSE
    Drop a small random number of edges; repeat the measurements
  END IF
END FOR

FOR all cell leaders picked in previous step
  Generate links among cell leaders to produce required inter-
     cell density
END FOR
```

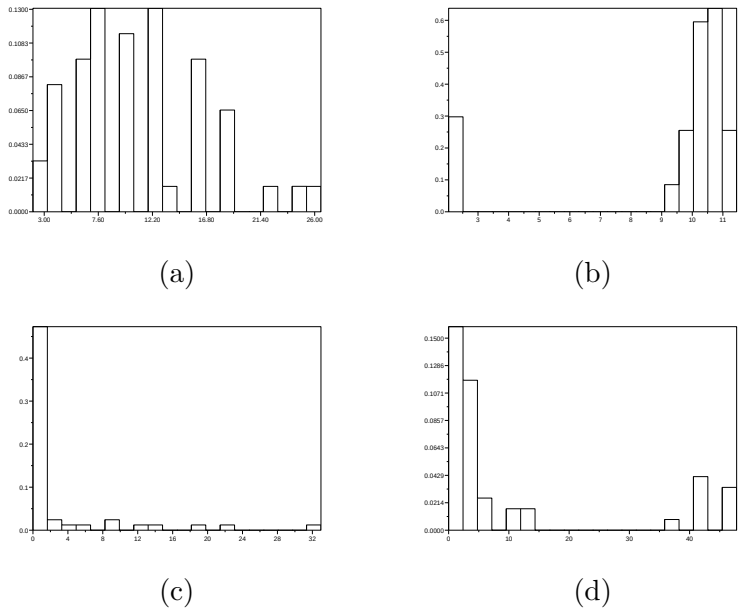


Figure 7: Distribution of centralities in a cellular network: (a)Degree, (b)Closeness, (c)Betweenness, and (d)Eigenvector

(b) workgroups, (c)departments, (d)divisions, and (e)an entire corporation - resulting in a 5 levels of groupings.

- Assume that groupings at each of the levels (e.g. departments) connect to each other with a network structure that can be expressed with a generative function (unform, scale-free, etc).

A generalized algorithm for generation of complex organization network can be described as a traversal of the hierarchy of layered groupings from most specific to most general while applying a generative function for each of the layers to generate edges at the given layer.

Thus, generation of a complex network can be parameterized with a profile consisting of (a)number of layers , (b)size of groupings at each layer, and (c)a simple generative function for each layer.

Given that number of simple generative functions is finite such parametrization can be then viewed as an optimization problem, defined as traversal of a state-space of generative profiles and evaluating the fit of each generative profile to a population of known networks.

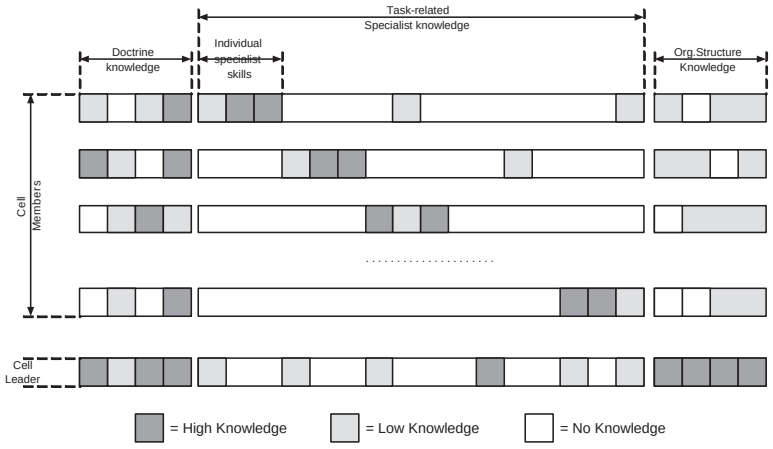


Figure 8: Knowledge Distribution of NetWatch Agents

7 Generating Knowledge Networks

Knowledge is represented in the MetaMatrix as a set of nodes, with each representing facts or groups of facts. Knowledge that an agent possesses is referred to as an edge between an *Agent* node and a *Knowledge*; knowledge that is required to accomplish an primitive task is represented as an edge between *Task* and *Knowledge* nodes; etc.

Based on data available on structure of terrorist training[12], NetWatch generates agent-knowledge networks using a profile of the knowledge network of a cellular organization.

The knowledge that the agents possess is divided into a three main categories. These categories encompass (a) general doctrine and ideology of the organization, (b) shared training and skills in MO of the organization (e.g. communication procedures, clandestine operations), (c) specialist task-related skills (e.g. bomb-making, sniper skills, getaway car driving), and (c) knowledge of overall organizational structure.

The algorithm for generating the knowledge network presumes the existence of well-formed cells, as generated by the algorithm in section 5.3. The following principles are followed:

- Cell leaders are more knowledgeable than other members. As cell leaders are recruited from the ranks of experienced operatives, their doctrinal knowledge is high and they possess many of the shared skills of the

other agents. They also possess a small amount of knowledge in each of the specialist areas. This knowledge is not sufficient to replace specialist agents but is sufficient to proficiently delegate subtasks during execution of a complex operation.

- Cell members share an ideological doctrine and a *modus operandi*, further referred to as *"shared knowledge"*. Adherence to a militant ideology is a driving factor in recruiting of operatives in terrorist organizations and is further amplified during training of studies in an a militant religious academy.

  Shared M.O. skills are derived from shared training camp experiences that terrorist organization recruits undergo. Shared skills include communication procedures, clandestine operation skills, preservation of secrecy during planning and preparation of operations.

- Cell members possess specialized knowledge that outlines their specific function within a cell; these facts are further referred to as *"specialist knowledge"*.

- A specialized portion of the knowledge network deals with overall knowledge of the organizational structure and policies. This knowledge is privileged information distributed only to cell leaders and is further referred to as *privileged knowledge*. However, rank-and-file cell members may obtain small amounts of the privileged information through interaction with other agents outside the primary cell.

The algorithm that generates knowledge networks as outlined above is fairly simple. The knowledge network is divided into portions based on purpose of each fact(e.g. shared knowledge, specialist knowledge, privileged knowledge)(see figure 8).

Then, for each agent a_i and fact f_k the algorithm generates a probability $P_{i,k}$ of existence of a an edge $a_i - f_k$ based on the group that the agent belongs to (i.e. cell leader vs. rank-and-file) and what group the fact belongs to (i.e. shared, specialist or privileged).

The edges are then instantiated with a roll of the dice.

7.0.1 Algorithm Parameters

The knowledge network generator depends on the following parameters:

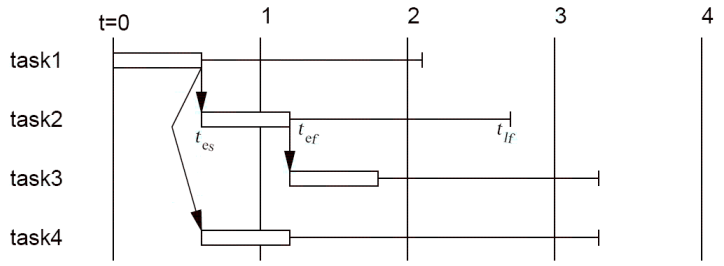


Figure 9: Construction of a Task Network as a Precedence Graph

- Proportion of shared knowledge
- Proportion of specialist knowledge
- Proportion of privileged knowledge

8 Generating Task Structures

The task network consists of a set of *primitive* and *compound tasks* with their precedence relations expressed as $Task - Task$ edges in the MetaMatrix. The complexity of the task network in terms of feasibility of execution can be controlled by varying the *average connectivity* (sum of predecessors and successors) of a task[13]. This parameter can be essentially thought as controlling the parallelism within the task network.

If the people-to-people network was generated as a cellular network, assignments of people to subtasks ($Person - Task$ edges) are uniformly distributed within each cell. This results in various degrees of subtask difficulty (amount of resource seeking and delegation required to accomplish the task). When people-to-people networks are created as random or scale-free graphs, the task assignments are distributed uniformly throughout the entire network which results in some tasks being not feasible.

9 Scalability

To estimate efficiency of the network generation algorithms, we have conducted timing runs of each of the algorithms for generation of people-to-people networks: Erdös random graphs, scale-free networks with preferential

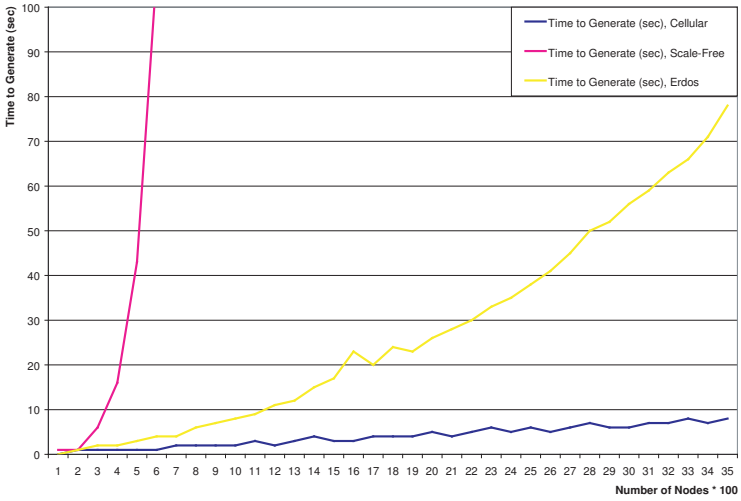


Figure 10: Time requirements to generate networks

attachment, and cellular networks. We varied the size of the network to be generated from 100 to 3500 nodes.

Figure 10 shows the time in seconds to generate a network of a given size with each of the algorithms. The least efficient of the algorithms is the preferential attachment algorithm, which grows exponentially. Use of this algorithm becomes impractical for networks over 2000 agents, where generation of the graph took approximately 10000 seconds, or a little under 3 hours. While the computational complexity of this algorithm is very high, it can be executed by a parallel machine in near linear time [26].

Erdös random graphs have been shown[15] to have a quadratic complexity ($\Theta(n^2)$). However, one iteration of edge generation is a very fast operation, so the algorithm remains practical in generating networks of up to 20000 nodes (generation time is 120 seconds).

The cellular network generation algorithm performs in near-linear time due to the fact that cells are small and self-contained. The computational complexity of the cellular network generator is $\Theta(\sigma_{cell}\frac{n}{k}k^2 + \sigma_{intercell}n) = \Theta(\sigma_{cell}nk + \sigma_{intercell}n)$ where n is the number of nodes, k is the mean size of a cell, and σ_{cell} and $\sigma_{intercell}$ are, respectively, densities inside the cell and between cells. Thus, when k is much smaller then n, the complexity of the cellular network generator is close to $\Theta(n)$. In practical terms, this means that even very large networks can be generated in relatively short times, with

a 20,000 node network taking less then 20 seconds to generate.

10 Conclusion

All of the network generation algorithms described above are used as a means of testing NetWatch, a large-scale multi-agent simulation of covert networks.

While realism of data generated by any of these algorithms can be disputed and nothing is more realistic then empirical data, the use of diverse techniques for generating initial data allows the simulation researcher to test the multi-agent system on networks of widely varying sizes and topologies. Due to small quantities of available empirical data, this is currently not possible to do without resorting to artificially generated data.

This report is not comprehensive in regards to generation of all possible network topologies. In this work, we did not consider small-world networks, as generation of small-world topologies is addressed well in [27] and [37]. Further, we did not consider issues of generating hierarchical networks.

In the field of modeling social and organizational networks, it is important to address organizations as comprehensive network structures, incorporating structures of task interdependency, information and resource requirements as well as person-to-person structures. This comprehensive approach would allow modeling organizations based on their form, e.g. departmental, functional, or matrix organizations.

While the generalized generative approach described in section 6 allows for wide flexibility in the topology of generated networks, it is not designed for modeling organically emerging network forms, such as these of markets. For example, market-driven network may exhibit emergent segmentation processes[36], which, due to the complexity of the market process, can be only generated via simulation of the market environment.

As a software engineering tool, the network generation package provides a consistent interface to all of its generation functions - therefore enabling the user (e.g. NetWatch) to test performance of the simulation tools on a wide variety of source networks. This also forces the simulation to remain independent of the initial network topology and thus allow for multi-theory testing of simulation tools.

References

[1] Al-Quaeda. Al quaeda training manual: Declaration of jihad against unholy tyrants. URL: `http://www.usdoj.gov/ag/trainingmanual.htm`, 2001.

[2] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley and Sons, New York, 1992.

[3] A. Barabasi. *Linked: The New Science of Networks*. Perseus Publishing, Cambridge, 2002.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.

[5] P. Biernacki and D. Waldorf. Snowball samplong. problems and techniques of chain referral sampling. *Sociological Methods Research*, 10(2):141–163, 1981.

[6] S. Borgatti, K.M. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. http://www.casos.cs.cmu.edu/publications/papers/CentralityRobustness5b.pdf, 2004.

[7] Kathleen M. Carley. *Dynamic Network Analysis*, pages 133–145. Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. Comittee on Human Factors and National Research Council, ronald breiger and kathleen m. carley and philippa pattison, (eds.) edition, 2003.

[8] Kathleen M. Carley. Dynamic network analysis. In K. Carley R. Breiger and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 361–370. Committee on Human Factors, National Research Council, 2003.

[9] Kathleen M. Carley. Linking capabilities to needs. In K. Carley R. Breiger and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 361–370. Committee on Human Factors, National Research Council, 2003.

[10] Kathleen M. Carley, Matthew Dombroski, Maksim Tsvetovat, Jeffrey Reminga, and Natasha Kamneva. Destabilizing dynamic covert networks. *Proceedings of the 8th International Command and Control Research and Technology Symposium*, 2003.

[11] Kathleen M. Carley, Jeffrey Reminga, and Steve Borgatti. Destabilizing dynamic networks under conditions of uncertainty. In *IEEE KIMAS*, Boston, MA, 2003.

[12] K.M. Carley, J.S. Lee, and D. Krackhardt. Destabilizing networks. *Connections*, 24(3):79–92, 2002.

[13] John Collins, Maksim Tsvetovat, Bamshad Mobasher, and Maria Gini. Magnet: A multi-agent contracting system for plan execution. In *Proceedings of SIGMAN 98*, 1998.

[14] RAND Corporation. Purpose and description of information found in the incident databases, 2003. http://www.tkb.org/RandSummary.jsp.

[15] Erdös and Rényi. On the evolution of random graphs. *Publication of Mathematics Institute of Hungian Academy of Sciences*, 5:1761, 1960.

[16] Erdös and Rényi. On the strength of connectedness of random graphs. *Acta Math. Acad. Sci. Hungar*, 12:261–267, 1961.

[17] F.Lorrain and H.C. White. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1, 1971.

[18] Rebecca Goolsby. Combating terrorist networks: An evolutionary approach. In *Proceedings of the 8th International Command and Control Research and Technology Symposium*. Conference held at National Defence War College Washington DC, Evidence Based Research Vienna VA, 2003.

[19] M. Granovetter. Network sampling: Some first steps. *American Journal of Sociology*, 81:1267–1303, 1976.

[20] Le Gruenwald, Gary McNutt, and Adrien Mercier. Using an ontology to improve search in a terrorism database system. *Proceedings of the 14th International Workshop on Database and Expert System Applications (DEXA'03)*, 2003.

[21] Brian Houghton. Understanding the terrorism database. National Memorial Institute for Prevention of Terrorism Quarterly Bulletin, 2002.

[22] IntelCenter. Intelcenter database (icd), 2005. http://www.intelcenter.com/icd/index.html.

[23] IntelCenter.com. Mapping al-qaeda v1.0. www.intelcenter.com, 2003.

[24] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physics Review Letters*, 85:4629–4632, 2000.

[25] Valdis E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2001.

[26] Benjamin Machta and Jonathan Machta. Parallel dynamics and computational complexity of network growth models. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 71(2):026704, 2005.

[27] M. E. J. Newman, C. Moore, and D. J. Watts. Mean-field solution of the small-world network model. *Physics Review Letters*, 84(14):32013204, 2000.

[28] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physics Review Letters*, 86(14), april 2001.

[29] Robert S. Renfro. *Modelling and Analysis of Social Networks*. PhD thesis, Department of Air Force, Air Force Institute of Technology, 2003.

[30] John Robb. Scale-free terrorist networks. Jef Allbrights Web Files; URL: `www.jefallbright.net/node/view/2632`, 2004.

[31] Richard Rothenberg. From whole cloth: Making up the terrorist netwrok. *Connections*, 24(3):36–42, 2002.

[32] M. Sageman. *Understanding Terror Networks*. University of Pennsylvania Press, 2004.

[33] Brent L. Smith and Kelly R. Damphousse. The american terrorism study: Indictment database, 2002.

[34] M. Tsvetovat and K.M. Carley. Modeling complex socio-technical systems using multi-agent simulation methods. *Kunstliche Intelligenz (Artificial Intelligence Journal)*, Special Issue on Applications of Intelligent Agents(2), 2004.

[35] Maksim Tsvetovat and Kathleen M. Carley. Structural knowledge and success of anti-terrorist activity: The downside of structural equivalence. *Journal of Social Structure (www.joss.org)*, forthcoming, 2005.

[36] Maksim Tsvetovat, Kathleen M. Carley, and Katia Sycara. Specialists, generalists and emergent market segmentation: a multi-agent model. *Journal of Computational and Mathematical Organizational Theory*, (8):221–234, 2002.

[37] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 1998.