On Effectiveness of Wiretap Programs in Mapping Social Networks [*]

Maksim Tsvetovat[†]   Kathleen M. Carley[‡]

Abstract

On December 16, 2005, a New York Times article[31] revealed that in the immediate aftermath of the September 11th attacks, the National Security Agency began a broad program of domestic signal intelligence collection. As press reports indicated [28], NSA implemented its new collections program based on the snowball sampling methods, which is generally used in surveying hidden populations and networks.

However, snowball method is known to be a biased toward highly connected actors[21] and consequently produces core-periphery networks when these may not necessarily be present. In case of terrorist networks, the last statement is particularly important in light of the "smoking gun" arguments presented by the government.

In a further argument, the government argues that wiretaps on suspect terrorist operatives need to be put in place extremely fast and in large quantities, thus overloading the FISA[37] court system. In the use of snowball sampling, overload of information collection system does present a distinct problem due to exponential growth of the number of suspects to be monitored.

In this paper, we will focus on evaluating the effectiveness of the wiretapping program in terms of mapping fast-changing networks of a covert organization. By running a series of simulation-based experiments, we are able to give a number of information gathering regimes a fair evaluation based on a consistent criteria. Further, we propose a set of information gathering programs that achieve higher effectiveness then snowball sampling, at a lower cost.

1   Introduction

On December 16, 2005, a New York Times article[31] revealed that in the immediate aftermath of the September 11th attacks, the National Security Agency began a broad program of domestic signal intelligence collection. Despite the fact that the NSA is barred by its charter from collecting information on US citizens on constitutional grounds, the Senate Joint Resolution 23 [36] of Sept. 14, 2001 was treated as a declaration of war, thus granting intelligence agencies broad surveillance powers.

As press reports indicated [28], NSA implemented its new collections program based on the following method: monitoring was started with a small number of known suspects within the US. Then, contacts of these subjects (e.g. telephone numbers that the subjects have called) were added to the list of suspects and monitored as well, thus expanding the dragnet with the speed of a combinatorial explosion.

In the social-science methodology, this method is known as "snowball sampling" and is generally used in surveying hidden populations and networks, such as these of drug users. Despite the fact that the snowball method is used frequently, it is not a panacea and is known to be a biased sampling methodology[7]. In particular, the snowball method gives preferential treatment to highly connected actors[21] and consequently produce core-periphery networks when these may not necessarily be present.

In case of terrorist networks, the last statement is particularly important in light of the "smoking gun" arguments presented by the government. It is known[33] that suicide terrorist operations such as these conducted by Al Qaeda are not conducted by core actors, but rather "operative cells" - tightly woven small groups of operatives on the periphery of the large network. Thus, surveillance of a peripheral operative not directly involved with an operative cell is likely to lead the investigation to central actors located overseas and not to a local sleeper cell, which in fact may be preparing a large-scale attack.

In a further argument, the government argues that wiretaps on suspect terrorist operatives need to be put in place extremely fast and in large quantities, thus overloading the FISA[37] court system. In the use of snowball sampling, overload of information collection system does present a distinct problem due to exponential growth of the number of suspects to be monitored. Thus, size of the program quickly becomes a liability and by necessity has to be controlled. This vast increase in need for computational power could be the cause of a significant expenditures to enhance U.S. wiretapping capabilities[24].

In this paper, we are not going to comment on

[†]George Mason University, mtsvetov@gmu.edu
[‡]Carnegie Mellon University, kathleen.carley@cmu.edu

the legality of methods employed by the government, but will instead focus on evaluating the effectiveness of the wiretapping program in terms of mapping fast-changing networks of a covert organization. By running a series of simulation-based experiments, we are able to repeatedly recreate scenarios of network evolution, thus giving a number of information gathering regimes a fair evaluation based on a consistent criteria.

Further, we propose a set of information gathering programs that achieve higher effectiveness then snowball sampling, at a lower cost. If implemented, such programs would increase intelligence-gathering capabilities without sacrificing or overloading the legal framework set by FISA Act and FISA court system.

This paper also shows an annealing-based information capture strategy that is a more effective means of sampling social networks of hidden populations. We show that this strategy is capable of gleaning more accurate information about the network from lesser number of captured messages, and does not exhibit a strong bias towards highly connected nodes. Moreover, we show how this strategy can be parametrically tuned to balance between breadth of survey and amount of captured information per node – a tunable bias towards core or periphery of the network.

The paper is organized as follows: in section 2 we establish the mechanism of structural analysis of covert network and its applicability to analysis of cellular networks such as Al Qaeda (section 3). Then, we introduce the simulation methodology that enables experimentation with wiretapping policies (section 6) and proceed with establishment of the baseline evaluation method and cost metrics (section 7). We proceed to evaluate performance of snowball sampling techniques (section 8) and simple SNA-based sampling policies (section 9). Finally, we introduce and evaluate an optimization-based wiretapping policy that lowers resource requirements and raises the quality of SIGINT collection (section 10).

2 Background

For reasons of national security it is important to understand the properties of terrorist organizations that make such organizations efficient and flexible. We must understand what does the underlying organization look like, how does it evolve, and how can the evolution of its structure be mapped through observation?

Terrorist organizations are often characterized as cellular — composed of quasi-independent cells and featuring a distributed chain of command. This is a non-traditional organizational configuration; hence, much of the knowledge in traditional organizational theory, particularly that focused on hierarchies or markets, is not directly applicable. Some lessons can be learned from previous work on distributed and decentralized organizations. This work demonstrates that such structures are often adaptive, useful in a volatile environment, and capable of rapid response [29][27]. In other words, one should expect terrorist organizations to adapt, and adapt rapidly.

Terrorist organizations are often characterized as dynamic networks where the connections among personnel define the nature of that evolution. This suggests that social network analysis will be useful in characterizing the underlying structure and in locating vulnerabilities in terms of key actors.

A further complication relates to the fact that the only way to obtain information about terrorist networks is by gathering intelligence — via signal interception (SIGINT) or human intelligence (HUMINT) means. By their nature, SIGINT and HUMINT techniques provide incomplete and frequently inaccurate data, and the heuristics for learning shapes of covert networks need to take this uncertainty into account. A cost factor is present as well - each piece of information comes with a price and it would be prudent to maximize its utility.

It is important at the outset to note that this examination of intelligence gathering strategies is highly exploratory. We make no claims that it is comprehensive, nor that the types of "error" in the data that intelligence agencies can collect is completely described. Further, our estimate of the structure of the covert network is based on publicly available data much of which is qualitative and requires interpretation. This work should therefore be read as a study in the power of an empirically grounded simulation approach and a call for future research.

We restrict my analysis to a structural or network analysis and focus on what the covert network looks like, how its structure influences its performance and its ability to pass information, how it evolves, and how its path of evolution can be altered. Admittedly, in this complex arena there are many other factors that are critical but they are beyond the scope of this study. Thus, from a straight social network perspective, this study suggests the types of methodological issues that will emerge when working with dynamic large scale networks under uncertainty.

3 Covert Terrorist Networks - the Al Qaeda

Al Qaeda, arabic for "The Base", is the largest known extra-national terrorist organization. In 2002, it was estimated to have the support of six to seven million radical Muslims worldwide, of which 120,000 are willing to take up arms [23]. Its reach is global with outposts reported in Europe, Middle East, East Asia and both Americas. In the Islamic world, its task is to purify so-

cieties and governments according to a strict interpretation of the Koran and to use religion as a unification force for the creation of an Islamic superpower state.

As Goolsby[20] stated, Al Qaeda extends its reach and recruits new member cells via the adoption of local Islamic insurgency groups. Beginning with provision of operational support and resources to facilitate growth, Al Qaeda representatives work to transform an insurgency group such as Jemaah Islamiyya (Indonesia) from a group seeking political change to a full-fledged terrorist organization executing multi-casualty attacks such as the Bali bombing in 2002[22].

Although the *modus operandi* of Al Qaeda is cellular, familial relationships play a key role. As an Islamic cultural and social network, Al Qaeda members recruit from among their own nationalities, families and friends. What gives Al Qaeda its global reach is its ability to appeal to Muslims irrespective of their nationality, enabling it to function in East Asia, Russia, Western Europe, Sub-Saharan Africa and North America with equal facility.

Unlike conventional military forces which are hierarchical and centralized, terrorist militant units are generally small, geographically dispersed and, at the first glance, disorganized. Nevertheless, they have been able to effectively counter much larger conventional armies. Large terrorist organizations operate in small, dispersed cells that can deploy anytime and anywhere [32]. Dispersed forms of organization allow these networks to operate elusively and secretly.

The need for security dictates that terrorist organizations must be structured in a way that minimizes damage to the organization from arrest or removal of one or more members [17]. This damage may be direct - making key expertise, knowledge or resources inaccessible for the organization, or indirect - exposing other members of the organization during interrogations. There are several factors that allow a terrorist organization to remain covert, including:

- Strong religious or ideological views that allow members to bond within a cell.

- Physical proximity among cell members, often to the extent of sharing living quarters, working and training together.

- Lack of rosters on who is in which cell.

- Cell members being given little knowledge of the organizational structure and the size of the organization.

- Inter-cell communication on as-needed basis only.

- Information about tasks issued on a need-to-know basis so very few people within the organization know about the operational plans in their entirety.

A need-to-know information policy can be counterproductive when an organization needs to complete a task that is larger than any one cell. Further, such policies tend to lead to duplication of effort and reduce the ability of one cell to learn from another. To fix these inefficiencies, terrorist organizations have been known to employ "sleeper links" - where a small number of members of each cell have non-operational ties (such as family ties, ties emerging from common training, etc) to members of other cells [25]. These links are rarely activated and are used mainly for coordinating actions of multiple cells in preparation for a larger operation.

To remain covert, the Al Qaeda has structured itself as a leaderless design characterized by its organic structure, horizontal coordination, and distributed decision making. However, the need to maintain a strong ideological foundation and resolve coordination issues has led to the need for strong leadership. One apparent solution has been to have multiple leaders diffused throughout the network and engaged in coordinating activities without central control or a hierarchy among the cells. Whether the leaders are themselves hierarchically organized, even though the cells are not, is less clear.

Substantial intelligence effort is needed to piece together the massive amount of information, both *postfactum* investigations and "logs" of activity, to generate a picture of the entire organization. Nevertheless, the picture that is emerging suggests that terrorist organizations are organized at the operational level as *networks* rather than as hierarchies [10].

4 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 [33][12][14]. 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[20]. 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[39].

Cellular networks[10] 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 abil-

ity 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[11].

Rothenberg[33] 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[40], which will ameliorate the loss of an individual. The redundancy in communication channels may also be mirrored in the redundancy 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[34].

This leads to a hypothesis that cells throughout the network contain structurally equivalent[18] 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. We 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.

5 Agent-based Network Modeling

NetWatch agents are intelligent adaptive information processing systems, constrained and enabled by the networks in which they are embedded. These networks evolve as individuals interact, learn and perform tasks. The design of the NetWatch multi-agent model is based on the principles of agent-based models of complex adaptive systems outlined by Langton[26]:

1. The model consists of a population of simple agents.

2. There is no single agent that directs all of the other agents.

3. Each agent details the way in which a simple entity reacts to local situations in its environment, including encounters with other agents.

4. There is no rule in the system hat dictates global behaviours.

5. Any behaviour at levels higher then individual agents is therefore emergent.

However, We make an important distinction from Langton's ABM techniques. In NetWatch and related models, agents are not defined as simplistic automata following a small set of deterministic rules. Instead, an agent can be viewed as a representation of a human actor involved in the the simulated activities. Using artificial intelligence techniques, the agents can plan and reason about task completion and formation of their social networks and make strategic moves to maximize their utility.

In effect, each agent within NetWatch is built in the same manner as an autonomous robot (sans the hardware) designed to survive on its own in a hostile environment. In greater detail, the methodology of multi-agent network modeling is based on the following principles:

- Agents are independent, autonomous entities endowed with some intelligence, though cognitively limited and boundedly rational. Agents can utilize both deterministic or stochastic rules.

- Agents and the networks in which they are embedded co-evolve. While the initial topology of agent network can be used as an independent variable, the community of agents will create a very different topology at the end of a simulation.

- Agents do not have accurate information about the world or other agents and are limited by their perception.

- Agents can learn the state of the world through interaction. Note that while agents do not have access to a global world-view, they can learn about their non-immediate neighbors through communication and collaboration with other agents.

- Agents can be strategic about their communication. They can use rule-based, decision-theoretic, optimization or other techniques to maximize their utility.

- Agents do not use predefined geometrical locations or neighborhoods. Instead, their choice of communication partners depends on the topology of their social network and evolves over time.

6 Modeling Dynamic Networks

Based on the conceptual framework of multi-agent simulations, we have developed NetWatch, a multi-agent network model for reasoning about the destabilization of covert networks such as organized crime or terrorist organizations under conditions of uncertainty.

NetWatch is built to simulate the communication patterns, information and resource flows in a dynamic organizational network based on cognitive, technological and task based principles. In addition, the model is grounded using information about surveillance technologies and intelligence operations (e.g. [2]) and the covert networks (e.g. [6]).

The design of NetWatch simulation of covert networks and anti-terrorist activity is based on the concept of *red teaming*, a war-gaming approach in which a population of participants is divided into two or more adversarial teams. The teams then are empowered to use any techniques at their disposal to achieve their goals. The main objective of red-teaming is learning the mind-set and *modus operandi* of the adversary and development and testing of strategies fielded against said adversary.

For the purposes of simulation, the covert network is designated as the **Red Team** and the network of
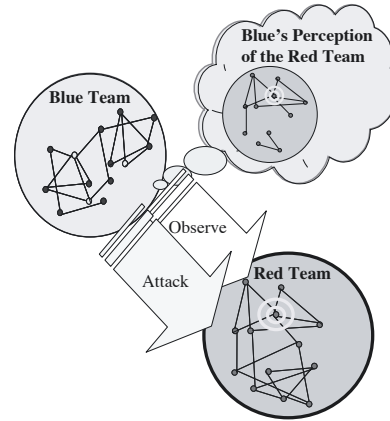

Figure 1: NetWatch Simulation Design

anti-terrorism forces as the **Blue Team** (see figure 1). Both of the teams consist of a number of autonomous, intelligent agents, designed to simulate with highest possible fidelity the activity of individuals and groups present in the subject networks.

A complete description of NetWatch architecture for simulation of dynamic covert networks has been described in [41] and [38]; the complete technical specification of the system is beyond the scope of this paper.

The agents of the Red Team are intelligent, knowledge-driven planning agents that model the process of execution of a logistically complex terrorist attack - complete with gathering required resources, obtaining knowledge and training, and tactical planning.

Red Team network is modelled upon organizational structure of a terrorist organization and constructed to fit a statistical profile of such an organization. The statistical profile mechanism (described in[39]) allows for manipulation of both social network topologies and distribution of information and resources, which leads to robust capabilities for testing of theories of organizational design in covert networks.

The Blue Team represents a set of agencies engaging in anti-terrorist activity and pursues two interconnected goals. The blue team conducts signal intelligence-based information gathering and uses collected information to build a MetaMatrix representation[9] of the Red Team.

7 Learning Network Structure though Signal Intelligence: Random Sampling

Random sampling of communications can be considered a baseline against which performance metrics of other network sampling strategies can be compared. While random sampling of wire (or wireless) traffic is rarely used in the real world, in simulation it can be used

to provide a robust notion of signal-to-noise ratio and its effect on acquisition of knowledge about network structure.

As the true goal of intelligent wiretapping heuristics is to *"do more with less"* — i.e. produce a maximally correct result given a certain amount of captured traffic — it is prudent to compare quality of network structure discovered via intelligent heuristics with results of similarly configured random sampling systems.

In this experiment, We establish such a baseline by comparing performance of random sampling techniques across networks of different size and topology, while controlling for signal-to-noise ratio of the sampling apparatus.

7.1 Experimental Design This experiment is based on a matrix design and tests performance of a number of simple wiretapping strategies on simulated organizations of different size and initial topology, testing every possible combination of the following parameters:

Topologies	Number of Agents
Uniform	100
Scale-free	250
Cellular	0

The density of uniform random network[16] is set at 0.2 - i.e. probability $P_{i,j}$ of an edge existing between agents i and j is 20%. The scale-free network are grown using the Barabasi-Albert method of preferential attachment[5], with parameters:

$$k = 0.25$$

$$\gamma = 2$$

The cellular networks are generated using the mechanism described in [39] using the following profile:

Parameter	Value
Mean Cell Size	6
Cell Size St. Deviation	1.7
Internal Cell Density	0.9
Probability of Random Connections	0.01
Density of cell leaders	0.16
Probability of connection between leaders	0.6
Probability of triad closure within cells	0.9
Probability of triad closure outside cells	0.17

Each cell of the experiment was repeated 20 times, generating 20 random networks with the same parametric signature. Results of the experimental repetitions were averaged.
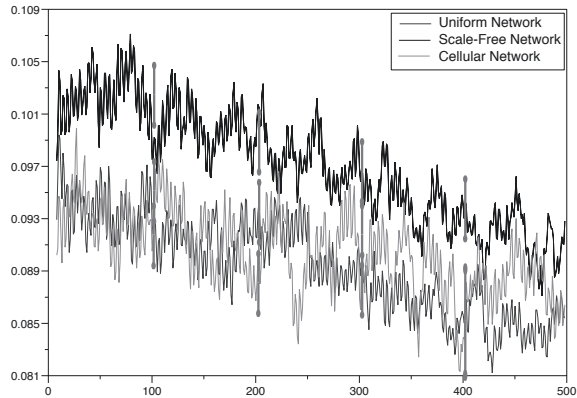

Figure 2: Effect of Initial Network Topology on Wiretap Performance, (100 agents; averaged over 20 runs)

7.2 Performance Measurement We measure performance of wiretapping strategies based on the assumption that the goal of discovering nodes and connection patterns between them is to achieve a maximally complete picture of the overall network structure. The concern in this case is the surveillance technique needs to not only uncover highly visible actors in the network but also discover the breadth of the network structure and minimize the number of undiscovered nodes and edges.

If this assumption is true, the best simple measure of quality of network is *hamming distance*[35], a sum of differences between two graph structures, in this case, the True and Learned Networks. The significance of hamming distance in this particular case is that it illustrates the overall number of errors made by the Blue Team agents. However, to compare performance of an algorithm on networks of different size, the raw hamming distance needs to be normalized:

$$D_{hamming}^{norm} = \frac{D_{hamming}}{e} = \frac{D_{hamming}}{(n-1)^2}$$

where e is the number of possible edges in a graph, an n is the number of nodes. Lower normalized hamming distance signifies higher performance.

7.3 Observations and Discussion

7.3.1 Effects of Signal-to-Noise Ratio on Wiretap Performance This portion of the experiment combines data collected for all initial topologies of the experimental design sampled at different levels of signal-to-noise ratio.

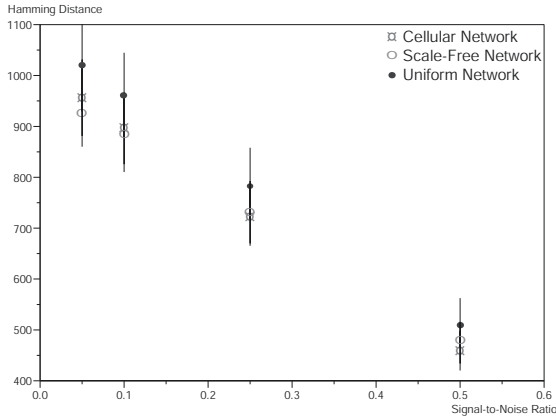Figure 3 shows the effect of signal-to-noise ratio

Figure 3: Effect of Signal-to-Noise Ratio Wiretap Performance; (averaged over 100, 200 and 500-agent networks, 20 runs in each configuration)

of the random sampling wiretap on the quality of acquisition of network data. The conclusion drawn from this figure on its own is fairly obvious — greater signal-to-noise ratio has a significant impact on the quality of learned network. However, this dataset is used in the experiments that follow as a point of reference and a baseline that other results are compared to.

The baseline results show a near-linear dependence of accuracy of network mapping on signal-to-noise ration, in case of purely random sampling of communications. This is expected due to the fact that random sampling has an equal chance of discovering all edges of the network, whether they belong to a highly connected agent or to a near-isolate agent. The probability of discovery of an edge at any given time is thus proportional to the ratio of messages that are captured and overall message traffic - which comprises the effective signal-to-noise ratio.

8 Snowball Sampling

A *Snowball Sampling* strategy is based on work of Biernacki and Waldorf[7] and Granovetter[21]. A snowball sampling strategy captures traffic originating from one agent and targets every agent with which it communicated. This essentially is a breadth-first search of the network. In NetWatch the snowball sampling strategy targets agents sequentially, one at a time.

While snowball sampling can quickly map communication in smaller social networks, it exhibits a number of problems. First of all, it can only discover agents that reside inside a single component of the network. This problem is compounded by the fact that cellular net-

Listing 1: Snowball Sampling Algorithm

```
CurrentTarget = a random starting point (A).

REPEAT:
  Add the CurrentTarget to the Taboo List (B)
  Add all agents that CurrentTarget
      communicates with to the Sampling Queue
  REPEAT
    NewTarget = deque from Sampling Queue
  UNTIL NewTarget is NOT on the Taboo List (C)
UNTIL Sampling Queue is empty
```

works consist of semi-isolated groups of agents where frequency of communication inside the group is much greater then frequency of communication outside the group. Mapping out multiple components of the network requires snowball strategy to use multiple, randomly selected entry points.

Further, as agents are targeted sequentially, a problem of oscillation arises. In a sub-network resembling a star topology, a snowball sampler has to return to the center of the star before it can continue to sample communications from other agents. This can be resolved by using a queue to manage a list of unexplored targets and a taboo list to prevent the search algorithm from revisiting the targets that it has already explored.

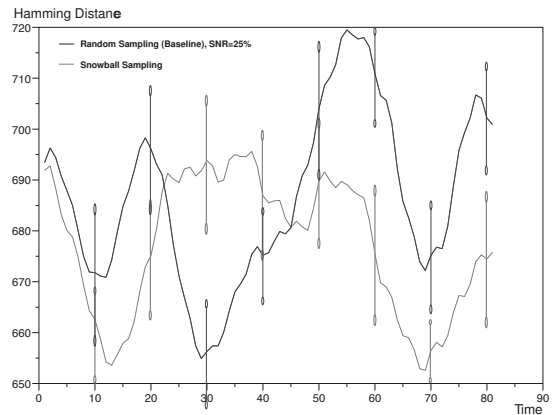Figure ?? and listing 1 illustrates how the snowball sampling algorithm explores a simple graph.

Figure 4: Snowball sampling performance: hamming distance (mean of 20 runs)

Snowball sampling strategy (see figures 4,??) performs at a level comparable to the random sampling baseline at signal-to-noise ratio of 25%. However, the measured signal-to-noise ratio of snowball sampling strategy (i.e. the ratio of number of captured messages

to number of rejected messages) is markedly lower - 16.3%.

Furthermore, snowball sampling[21] has been showed to be biased toward highly connected nodes, so extensive use of this technique may result in observation of scale-free core-periphery structures where none exist[7]. This experiment confirms that, while effective at learning network structures at a higher efficiency than baseline methods, snowball sampling does not discover the breadth of the network by avoiding nodes with low communication rates.

As scale-free models of terrorist networks are easily operationalized, and present a ready tactical model of counter-action, their popularity among intelligence analysts has dramatically increased. However, the reality of terrorist networks does not fit neatly into the scale-free network model. It has been observed[33] 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 autonomous 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[33] 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.

Next section presents a number of strategies that improve upon performance of snowball sampling on dynamic networks via use of socially intelligent traffic analysis and multi-point sampling.

Listing 2: Simple Socially Intelligent Sampling

```
CurrentTarget = random starting point
REPEAT
  Add CurrentTarget to TabooList
  Capture all traffic TO and FROM Current
      Target for a period of time
  Store captured messages in MetaMatrix
      accumulator
  Run Measure of choice on MetaMatrix
  CurrentTarget = agent highest in Measure
FOREVER
```

9 Socially Intelligent Traffic Analysis

As the Blue Team agents receive messages from the wiretap agents, they use their address information to build a representation of the network of the Red Team, or the *Learned Network*.

Thus, while Snowball sampling is myopic (i.e. can only see and survey small portions of the network at each time), a more intelligent Blue Team agent can use its accumulated knowledge of the target network to make intelligent decisions about locations of future wiretaps and configuration of their message filters.

The Blue Team agents implement an analysis toolkit containing a number of common social network analysis algorithms including *degree centrality*, *betweenness centrality* and *closeness centrality* [19]. Also accessible to the agents are methods of MetaMatrix analysis including *cognitive demand* [13], and knowledge and task exclusivity metrics[4],[3].

In its simplest implementation, the socially intelligent wiretap algorithm function is presented in listing 2.

9.1 Observations Figure 5 demonstrates performance of socially intelligent wiretap methods in terms of hamming distance between learned network and true network. Heuristics based on pure SNA metrics of degree and betweenness centrality produce essentially identical performance over time and several times reach the best performance among all techniques. However, on average they do not perform as well.

MetaMatrix-based metrics of cognitive demand and knowledge exclusivity track closely to each other in the first half of the run but diverge as knowledge diffusion increases. The explanation of this divergence lies in the fact that with time, knowledge becomes diffused among the agents. When agents send knowledge requests to other agents, with knowledge diffusion these requests will be sent across a larger group of agents — and thus, cognitive demand at time t becomes a worse predictor of who will communicate at time $t + 1$.
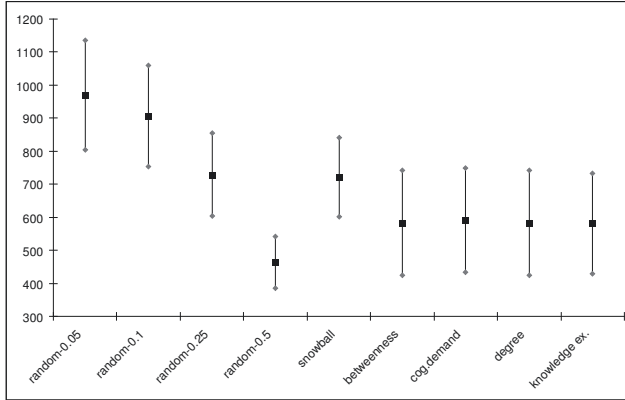
Figure 5: Mean Performance of Socially Intelligent Strategies (3x3 cells, 20 runs/cell)
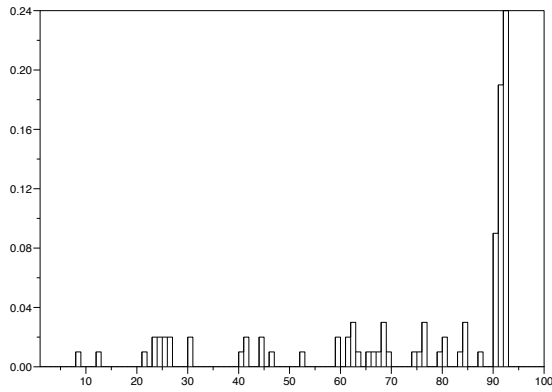

Figure 6: Histogram: frequency of capture of messages per agent; demonstrates adherence to local maximum in simple soc.int. heuristics (single run)

Overall, socially intelligent strategies performed significantly better then baseline established for signal-to-noise ratio of 0.25 (see figure 5) and slightly worse then baseline at signal-to-noise ratio of 0.5 (figure **??**). At the same time, the measured signal-to-noise ratios of socially intelligent strategies were significantly lower then those needed to achieve same performance in the baseline strategies.

Socially intelligent sampling strategies as a whole perform significantly better than baseline strategies at the same signal-to-noise ratio as well as snowball sampling strategy. The highest performance on average comes from strategies that take advantage of knowledge content of communications — cognitive demand and knowledge exclusivity.

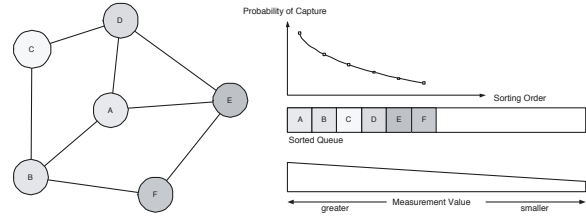However, a significant problem remains in the de-

Figure 7: Socially Intelligent Traffic Sampling with Probabilistic Targeting

sign of the simple heuristics. Essentially, these heuristics can be described as a hill-climbing algorithm where the sampling point (i.e. the wiretap) moves in the direction of highest value of the metric. However, these algorithms are generally unable to discern local maxima from globally optimal solutions. Once such local maximum is discovered, the hill-climber is unlikely to sample any other area of the network.

Figure 6 illustrates the occurrence of local maximum in one of the experiments (cellular network, degree centrality heuristic). The histogram shows frequency with which each of the nodes was targeted by the wiretap. In this particular case, the local maximum is located near Agents 92 and 93 - which together account for close to half of messages captured.

A further complication to the above problem is the fact that initially the Blue Team agents know very little about the Red Team — thus the accuracy of their estimations of centrality metrics is bound to be low[8],[15]. Therefore, the heuristic can fall into a local maximum within one or two time periods from the start.

While problems of local maxima are serious, they are not unsolvable. One of the best solutions for navigating parameter spaces with local maxima is to use an algorithm similar to simulated annealing with a measure of randomization in the beginning to serve as a bootstrapping mechanism. The next section describes such an algorithm, and shows its performance advantages over simple hill-climbing heuristics.

10 Socially Intelligent Traffic Sampling with Probabilistic Targeting

We propose a more robust solution to a traffic sampling: an algorithm with probabilistic targeting. This algorithm allows the much greater coverage of the network and is less prone to finding local maxima. The algorithm maintains a set of nodes that comprise its region of interest (ROI). After random initialization and a period of traffic capture, the captured MetaMatrix is analyzed and a new ROI is constructed based on the results of this analysis. However, nodes to comprise the new ROI
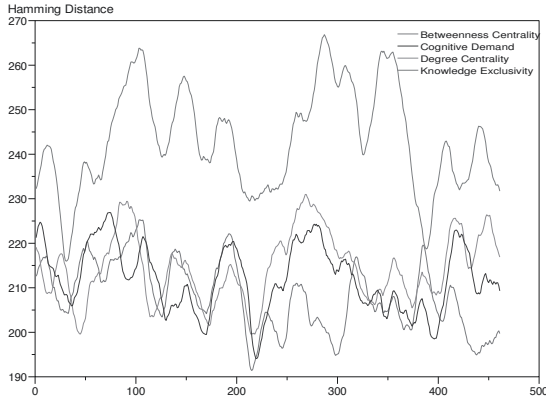
Figure 8: Performance of Probabilistic Socially Intelligent Strategies, cellular networks; 100 agents, single run

```
ROI = a small random set of nodes
Let Exp = Exponentially Distributed Random
        Variable (lambda)

REPEAT
  Capture traffic TO and FROM nodes in
        CurrentTargets
  Store captured messages in MetaMatrix
        accumulator

  Run Measure of choice on MetaMatrix
  Insert all nodes into an Array SORTED by
        value of Measures

  ROI = empty list
  FOR i = 0 to number of nodes
     Probability(i) = Exp(i)
     r = 0 < random number < 1
     IF (Probability(i) < r)
       ADD Array(i) to ROI
     END IF
  END FOR
FOREVER
```

are not picked deterministically. Rather they are picked by an exponentially distributed random variable. The distribution is composed in a way such that the most prominent nodes (i.e. these highest in the measure of interest) have the highest probability of being included in the ROI - but there is non-zero probability of the least well-connected nodes included as well.

The algorithm is illustrated on figure 7 and listing 3.

The exponentially distributed random variable is initialized as follows:

$$(10.1) \qquad P(x) = \lambda e^{-\lambda x}$$

Where parameter λ dictates the speed of fall-off of the probability distribution function. Thus, λ dictates how "adventurous" the algorithm would be in including little-known agents in the ROI.

Furthermore, manipulation of the λ parameter during the running of the algorithm results in behaviour similar to that of simulated annealing - the ROI becomes more constrained as more information on the network is obtained.

Results of evaluation of this algorithm are presented in the next section.

11 Performance of Intelligent Network Sampling Heuristics

In the previous experiment, We have shown that well-understood sampling mechanisms of snowball sampling and hill-climbing socially intelligent sampling outperform the baseline strategies but still perform suboptimally. In this experiment, we test a heuristic targeted at achieving the maximum breadth of coverage of nodes on the network - perhaps at the expense of depth of knowledge or number of undiscovered edges.

Performance is evaluated on the basis of hamming distance between the learned network and true network and in terms of effective signal-to-noise ratio. I further introduce a wiretap effectiveness metric that embodies the *"get more from less"* philosophy by combining the two metrics to study incremental efficiency of each of the algorithms.

Figure 8 shows that annealing-based sampling strategy clearly outperforms the simpler SNA-based algorithms described in section 10. At signal-to-noise ratios similar to the simple strategies (figure 9) the annealing-based strategies achieve a mean hamming distance of about 50% of simple strategies.

11.1 Sampling Biases Simulated Annealing algorithm as described above is a probabilistic method, where probability of capture varies depending on value of a cost function. This method achieves its low hamming distance metric by sampling more efficiently. Every agent that has been discovered has a chance of being sampled in a particular period as opposed to only agents with high levels of SNA metrics. The heuristic does not get stuck in local maxima and the randomization of search allows the heuristic to bootstrap itself efficiently. The level of randomization goes down slowly thus focusing the search and preventing waste of resources on
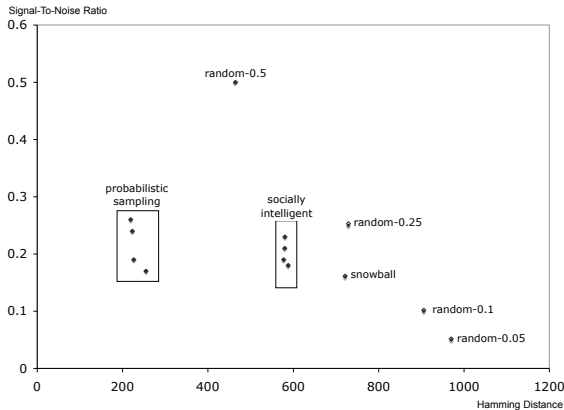
Figure 9: Overall performance of all sampling strategies; lower values on both axes represent better performance

agents that don't communicate a lot.

As we have mentioned in sections 8 and 10, both snowball and simple socially-intelligent strategies exhibit significant biases in relation to the network position of nodes. Snowball sampling tends to collect significantly more information on nodes that are highly connected; simple socially intelligent methods are biased towards prominent nodes detected by the SNA metric chosen by the user.

The probabilistic sampling method allows the user to effectively tune the algorithm's biases to achieve the best outcome for a particular problem. This is done by manipulating the fitness function that determines placement of nodes in the sorted queue. Strength of the bias is controlled by λ parameter of the exponential distribution (see equation 10.1), where higher values of λ correspond to stronger biases.

In its standard configuration, the algorithm exhibits a weak depth bias – a weak preference for sampling highly rated nodes. To achieve a breadth bias (i.e. to exhibit a prefence for discovery of new nodes), the user needs to reverse the sort order, thus placing newly discovered nodes at a priority for communication capture.

The tunable bias presents an opportunity for users of the algorithm to sample hidden populations such as covert networks with an account for changing priorities. In the beginning of the search, a breadth bias will achieve greater network coverage in shortest amount of time, and an increasing depth bias is suitable for later stages of data collection, where finding out as much as possible about prominent nodes takes precedence over

mapping the periphery of the network.

12 Conclusions

Our experiments show that it is possible to obtain high-quality data on covert networks without using random traffic sampling (e.g. Echelon) or snowball sampling, both of which capture too much unnecessary data, and do not make good use of the data that has been already captured.

As an alternative, use of optimization-based and socially intelligent sampling techniques allows for a tightly targeted and resource-thrifty SIGINT gathering program. Not only these techniques are significantly less costly, they also produce better overall intelligence data with a closer match to the covert network being studied.

The annealing-based information capture is not limited to sampling communications within covert networks, but rather is a flexible methodology for surveying networks of hidden populations. As the optimization-based sampling techniques do not require as much information capture as snowball sampling, they will present less of a resource strain on data collectors and are a more cost-efficient way to sample communication for analysis of social networks.

References

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

[2] D. Alberts, J. Garstka, and F. Stein. *Network Centric Warfare: Developing and Leveraging Information Superiority*. CCRP Publication Series, 1999.

[3] Michael Ashworth. Identifying key contributors to performance in organizations: The case for knowledge-based measures. CASOS Working Paper, 2003.

[4] Michael Ashworth and Kathleen M. Carley. Identifying critical human capital in organizations. CASOS Working Paper, 2002.

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

[6] N. Berry. The international islamic terrorist network. *CDI Terrorism Project*, September 2001.

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

[8] 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/CentralityRobusti 2004.

[9] Kathleen M. Carley. Smart agents and organizations of the future. In Leah Lievrouw and Sonia Livingstone, editors, *The Handbook of New Media*, chapter 12, pages 206–220. Sage, Thousand Oaks, CA, 2002.

[10] 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.

[11] 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.

[12] 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.

[13] Kathleen M. Carley and Yuquing Ren. Tradeoffs between performance and adaptability for c3i architectures. *Proceedings of the 2000 International Symposium on Command and Control Research and Technology*, 2001.

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

[15] E. Costenbader and T.W Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25:283–307, 2003.

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

[17] Bonnie H. Erickson. Secret societies and social structure. *Social Forces*, 60(1):188–210, 1981.

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

[19] L. C. Freeman. Centrality in social networks: Conceptual clarication. *Social Networks*, 1:215–239, 1979.

[20] 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.

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

[22] International Crisis Group. Indonesia backgrounder: How the jemaah islamiyaa terrorist network operates. *Asia Paper*, (43), December 2002.

[23] Rohan Gunaratna. *Inside Al Quaeda: Global Network of Terror*. New York University Press, New York, NY, 2002.

[24] Hannibal. The new technology at the root of the nsa wiretap scandal. *Ars Technica*, December 20 2005.

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

[26] C.G. Langton. *Artificial Life*, pages 1–47. SFI Studies in the Sciences of COmplexity. Addison-Wesley, Redwood City, CA, 1989.

[27] P. Lawrence and J. Lorsch. Differentiation and integration in complex organizations. *Administrative Science Quarterly*, (12):1–47, 1967.

[28] ERIC LICHTBLAU and JAMES RISEN. Spy agency mined vast data trove, officials report. *New York Times*, December 24 2005.

[29] Zhiang Lin and Kathleen M. Carley. *Designing Stress Resistant Organizations: Computational Theorizing and Crisis Applications*. Kluwer, Boston, MA, 2003.

[30] Marvin L. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, N. J., 1967.

[31] James Risen and Eric Lichtblau. Bush lets u.s. spy on callers without courts. *New York Times*, December 16 2005.

[32] D. Ronfeldt and J. Arquilla. Networks, netwars and the fight for the future. *First Monday*, 6(10), 2001.

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

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

[35] A. Sanfeliu and K.S. Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:353–362, 1983.

[36] U.S. Senate. Joint resolution 23: "authorization for use of military force.", 2001.

[37] Lee Tien. Foreign intelligence surveillance act: Frequently asked questions, 2001.

[38] 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.

[39] M. Tsvetovat and K.M. Carley. Generation of realistic social network datasets for testing of analysis and simulation tools. Technical Report Technical Report CMU-ISRI-05-130, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, 2005.

[40] 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.

[41] Maksim Tsvetovat and Kathleen M. Carley. Simulation of human systems requires multiple levels of complexity. *IEEE Proceedings on System, Man and Cybernetics*, to appear, 2006.