Seven Deadly Sins of Social Network Analysis

Common Pitfalls in the Application of SNA for Defense and Intelligence Uses

Daniel B. Horn, Michael L. Haxton
Booz Allen Hamilton
McLean, VA, USA
horn_dan@bah.com, haxton_michael@bah.com

Drew Conway
Department of Politics
New York University
New York, NY, USA
drew.conway@nyu.edu

***Abstract*—The application of Social Network Analysis (SNA) methods in support of defense and intelligence efforts has become increasingly common in the decade after the September 11, 2001 attacks. Although SNA has a great deal to offer these communities, care must be taken to avoid a range of potential pitfalls that may undermine the effectiveness or validity of these methods. Based on a series of interviews, observations, and experiences working with social network analysts, we have identified seven critical pitfalls that are commonly faced. In this paper we provide an overview and explanation of each pitfall and suggestions for avoiding and mitigating it. The list we provide is by no means comprehensive, and some pitfalls are difficult or impossible to overcome. However, we believe that an awareness of how they may impact the results or interpretation of an SNA may help reduce overconfidence or misapplication of these techniques.**

Keywords-Social Network Analysis, Methodology

I. INTRODUCTION

In the months and years after the September 11, 2001 terrorist attacks, Social Network Analysis (SNA) garnered a great deal of attention within the defense, intelligence, and law enforcement communities due in part to the recognition of the decentralized structure of terrorist organizations [1]. SNA methods have had a place in investigative approaches for decades [2], but this new urgency, coupled with advances in computing and visualization capabilities, has led to a renaissance within these communities. With this new prominence came new tools and new users of these techniques. Where once SNA was primarily the domain of academics or methodologists, metrics were now becoming available to a broad range of analysts. One unfortunate consequence of this wider presence is the increased ease with which SNA methods can be incorrectly applied, leading to the potential for flawed interpretation.

In an effort to understand and mitigate the pitfalls that face network analysts, we conducted a series of interviews with social network analysts across the intelligence and defense communities. Drawing on these interviews and our own observations and experiences, we have identified seven prevalent and concerning pitfalls around the application of SNA in the intelligence and defense arenas. In discussing each pitfall, we present a description of the pitfall as well as some potential steps to avoid or mitigate it. This list is not meant to be comprehensive, and should not be seen as a hard and fast set of rules. Instead, we hope this will create a discussion among stakeholders including methodologists, practitioners, tool builders, and consumers of analytic products.

This work was funded through contract #SP0700-03-D-1380 by the Defense Threat Reduction Agency (DTRA) to SURVIAC. The views, opinions, and/or findings contained in this article are solely those of the authors and should not be construed as an official DTRA, SURVIAC, or DOD position, policy, or decision, unless so designated by other documentation.

II. PROCEDURE

In the fall of 2008 we conducted a series of interviews with social network analysts and methodologists at 4 agencies across the US Department of Defense and Intelligence Community. These interviews focused on a range of topics including tools, methodology and challenges. Each interview was loosely organized around a set of interview questions meant to focus on ten topics:

1) Background information on the organization and the conditions surrounding the development and structure of its SNA capability.
2) Target set(s) the organization focuses on, and the established process for managing multiple target sets at once.
3) Data and sources the organization uses to populate databases, spreadsheets, and tools that will be analyzed, as well as, the procedures for managing this data.
4) Analysis procedures, techniques, and metrics the organization applies to the data it has gathered.
5) Analytic tools the organization uses or is developing and their opinions and preferences regarding each tool.
6) Staffing and training the organization's staff receives in order to perform SNA, as well as the general expertise the staff possesses.
7) Production process the organization carries out after analysis has been done and the formats and presentation of final products.
8) Consumers the organization serves, how they use the products, and what feedback they provide.
9) Lessons learned, best practices, and challenges.
10) Interviewee recommendations of other US Government organizations that perform SNA that may provide additional useful information.

In addition to these interviews, we drew on our experiences working with analysts using SNA and non-SNA approaches to a variety of related issues and a review of commonly used network- and link-analysis tools used within academia, and the defense and intelligence communities.

III. Analysis

To analyze the data, we used inductive qualitative techniques [3][4] with a focus on factors associated with the organization and implementation of SNA efforts. After the four interviews, the team analyzed the interview responses looking specifically for both trends and differences in answers of each organization's responses. Data analysis consisted of reading and discussing interview notes. This focused dialogue guided another examination of our data in which we further developed the themes and assessed the extent to which the data accurately represent current practices, based on both our subjects' descriptions and our own knowledge of the practice of SNA.

Defining Social Network Analysis

Our use of the term "Social Network Analysis" refers to the application of computational, mathematical, and statistical analysis to complex relational data [5]. At a fundamental level, SNA is a mixture of theoretically-based metrics and statistical techniques, as well as common heuristics that, over the past 60 years, have been proven effective at understanding social network structure. The theoretical basis of most metrics and techniques derive from sociology, social psychology, and anthropology [5, pp. 10-11]. The mathematical and technical foundation for most of the techniques and algorithms comes from graph theory, statistics and linear algebra. There are a host of techniques that range from simple statistics to complex computational algorithms, including SNA metrics concerned with individual entities within the network (e.g., centrality metrics), SNA metrics concerned with pairwise comparisons of structure (e.g., structural equivalence), SNA techniques that find community or group structure (e.g., cluster analysis) and SNA metrics concerned with properties of an entire network (e.g., transitivity).

IV. Results

In our analysis of our interview notes, reviews of our own observations, and discussions with others, seven themes emerged, each representing a common, critical pitfall encountered in the application of SNA in investigative and analytic contexts. The themes were:

1) *Confusing Link Analysis with SNA* – both are valid, but confusing them dilutes the value and strength of SNA
2) *Failing to differentiate modes and layers in your networks* – inaccuracies result from mixing different types of nodes and relationship types
3) *Invalid applications of SNA* – either applying unproven techniques for SNA, or applying proven techniques incorrectly; both are commonly done
4) *Designing a methodology around a tool* – methodology should derive from analytic requirements, not from the available technology
5) *Ignoring context in SNA* – qualitative expertise and information is crucial to effective SNA
6) *Inappropriately handling error in network data* – either ignoring the impact of error, or delaying analysis to wait for perfect data; both represent serious issues
7) *Applying SNA to inappropriate problems* – there are no clear guidelines, but some rules of thumb exist

These pitfalls have been faced by analysts and methodologists in various places in the analytic community, and have evoked a wide range of responses and strategies across organizations. Below, we provide an overview of each pitfall and suggest ways to avoid or mitigate their impact.

1. Confusing Link Analysis with SNA

Over the past decade, there has been a surge in the study of networks. A search of scientific journals finds that during the period from 1991-2000, 2,771 articles were published containing the phrase "network analysis," while from 2001 to 2010 that same search delivers 7,838 articles (nearly 3 times as many) [1]. For many researchers—particularly those in the defense and intelligence communities—this sudden interest in networks was predicated on the realization that terrorist organizations were better represented as networks of indeterminate form rather than the governmental agencies they were used to dealing with. As the demand for network analysis rose, many researchers rushed to meet this need. However, much of the research resembled network analysis only in name, and severely lacked the fundamental methodological foundations of actual SNA.

This trend continues today, and constitutes one of the major pitfalls in SNA: the misrepresentation of Link Analysis as SNA. Link Analysis is the reporting of network data. This difference is not subtle, yet many in the defense and intelligence communities fail to make the distinction, leading to poor analysis and misleading results. Link Analysis often takes the form of a link chart, which simply displays the network ties, and is occasionally combined with the reporting of simple centrality metrics. Reorganization of the entities, grouping by analyst judgment or observed attributes in the visual display, is the most common methodological technique used in Link Analysis. Conversely, SNA involves a large and expanding set of graph theoretic and statistical methods, and visualization techniques grounded in social science theory and supported by robust empirical results.

The goal of Link Analysis has historically been to organize and communicate relational data. The use of link charting tools has become increasingly common in defense, intelligence, and law enforcement communities. These tools enable analysts to quickly create a tangible visual representation of complex data. This can tremendously aid the establishment of a shared

[1] Search was performed on the ScienceDirect database, a clearinghouse for hundreds of scientific journals across several disciplines.

situational awareness between analysts or from analysts to decision makers. One of the great strengths of Link Analysis is the ease with which different classes of entities ('modes') or relations can be intermingled in a single visualization (see Pitfall 2). Because the goal is communication to a person, an analyst using Link Analysis can introduce a wide range of data modes or relation types in idiosyncratic ways. As an example, an analyst describing a terrorist network may add a single event to a link chart to show the connection between otherwise disparate groups. In using SNA, the analyst would be well served to consider whether other, less illustratively important events should be captured in the network. Such events might 'muddy the waters' of the link chart, but could enable statistical analysis. Without adding these events, a 2-mode network of people-by-events with only a single event would be misleading.

Relational data has become one of the most crucial forms of data for intelligence analysis today. Relational data is at the heart of SNA – i.e., data about the relations among a defined set of entities – and it is also used in many other analytic approaches and methodologies. Relational data is also used in Link Analysis to manage and display complex relational datasets. Link Analysis can be a useful approach on its own or for supplementing SNA studies, but it is not SNA. While Link Analysis helps an analyst visualize how individuals in a network are connected, it does not provide the statistical and mathematical assessment of what those relationships signify – SNA does. The metrics, algorithms and techniques associated with SNA are based on tested and established theories from academia. This is the key aspect of SNA that distinguishes it from other widely used methods for analyzing relational data: Social Network Analysis has a sound theoretical foundation and track record that can support intelligence analysis confidently.

The blurring of these two approaches is problematic for SNA in many ways. First, when Link Analysis is presented as SNA, it gives the illusion that pictures and lists constitute a consistent, quantitative, analytic methodology, which dilutes the strength and legitimacy of SNA. Second, the consumers of intelligence analysis are either left with a false notion of what network analysis is, or imbue the analysis with greater confidence than is warranted. Both of these aspects of the problem can lead to issues with future use and acceptance of a valid and worthwhile technique, and also to possible errors in decision-making resulting from over-confidence in the analysis.

Avoiding the Pitfall

In order to avoid this pitfall, agencies and analysts alike must educate themselves in the methodological techniques underpinning SNA, rather than just the software and database tools available. Embracing the technology and data is insufficient for avoiding this pitfall. Distinguishing between SNA and Link Analysis requires adequate education and training. Education provides the foundation to understand when analysts are doing Link Analysis and when they are doing SNA. Education is also crucial to the appropriate application of SNA – it provides the fundamental understanding of the techniques to ensure they are being applied appropriately – ensuring appropriate quality levels are maintained. Many of the techniques in SNA are complex, and while technology can help, analysts will need to ensure that the techniques are appropriate to the data and correctly interpreted.

2. Failing to differentiate modes and layers in networks

There are two distinct versions of this pitfall, both fundamentally resulting from imprecise data collection. The first involves the entities, or "nodes" in the networks, while the second involves the connections between these entities. Imprecise accounting of the types of nodes in a network is sometimes referred to as "mixing data modes."[2] This involves conflating entities of substantively different types with each other such as people, organizations, geospatial locations, etc. Each of these types of entities should be treated distinctly within SNA. The second involves the relationships between the entities in the network. Imprecise accounting of the types of relationships in a network typically involves conflating distinct types of relationships that fundamentally alter the interpretation and analysis of the social network.[3]

Mixing Entity Types

A data *mode* refers to distinct types of entities in the network. It is extremely valuable to identify multiple types of entities in SNA, but each must be clearly identified and the cross-over between modes must be analyzed using a distinct set of techniques. A mode is considered distinct when it identifies entities with qualitatively different types of connections within and across the modes (e.g. the distinction between movie actors and films), when the fundamental potential for forming ties is distinctly different (e.g. while movie actors and fans are both fundamentally people, the potential for forming ties is very different for each), or when there are no relevant connections among the entities of that type (e.g. ties among movie actors may be considered irrelevant to analysis of directors' choices of actors, so directors and movie actors should be treated as distinct modes in such a study).

Communication accounts, such as social media accounts, telephone numbers, and email addresses, should not be equated with people. Network-based analysis of these modes can be useful and powerful, but should be analyzed as separate data modes in SNA, not as equivalent to the individuals who may utilize these modes, or to each other. There are two fundamental reasons for this: First, a communication mode may be used by multiple people. Analysis of communication patterns can be bolstered with some SNA techniques, but these data modes must be treated distinctly from direct observation of interactions and exchanges amongst people.[4]

Second, a person can use multiple modes of communication, any one of which would provide an incomplete view of the interactions within that network. For example, using email traffic as a representation of the complete network of a set of individuals would bias the data

---

[2] A mode in network data refers to the types of nodes or entities. Multi-modal SNA involves analyzing connections across multiple types of nodes in a network. They should not be conflated or treated equivalently.

[3] Multiple relationship types in a network can be thought of as multiple layers of connections among the nodes in the network.

[4] Except in the case where a telephone or email communication can be definitively linked to specific people. In that case, this can be considered as data within the person-to-person mode of relational data.

substantially. SNA should strive to represent the relational data among the entities through multiple sources of data in order to provide as complete a representation of the network as possible. In practice, communication accounts and their users can be linked as multi-modal networks, and exploration of the networks when communication methods (including non-technologically mediated communication) are collapsed can be fruitful.

The theoretical basis of SNA is grounded in human behavior, where the central mode of interest is individuals or groups. Therefore, while it can be very interesting and powerful to use SNA-based techniques when analyzing the relationships between 'non-human' entities, such as roads, internet servers, and food webs, these other types of Network Analysis should not be considered SNA. These techniques may be valuable, but their application to these other types of entities must be tested and validated – different assumptions about the meaning and interpretation of results must be used when dealing with 'non-human' entities. The track record and body of literature in SNA does not automatically apply to these other applications simply because they are using the same techniques. Caution must be used in these cases (please see section 1 above, "Calling something SNA that is not SNA," as well as section 6 below, "Misrepresenting SNA," for more details). Analysts must pay close attention to the assumptions on which SNA (or any other analytic technique) is founded.

Mixing relation types

Relationship types are the distinction between the qualitative character of the relationships and entities. At a simplistic level, one would not want to confuse negative affect relationships, like 'enemy of' or 'hates', with positive affect relationships, like 'allies with' or 'loves'. Failing to distinguish between the types of relationships does not invalidate the application of SNA necessarily, but it does substantially limit the inferences that can be drawn from the data. If all relationships between entities are conflated (e.g. love, hate, communicates with, relative of, knows of), the presence of a relationship tie between two people simply denotes that they may know of each other. As relationship coding schemes become more precise, analysts are able to draw more fine-grained distinctions in structural features across the relationship types[5]. Analysts must be very careful not to over-extend the analysis beyond the validity of their data, including the granularity in the relationship coding scheme.

While more precise and detailed relationship coding schemes enable SNA to have more depth, it is also true that this requires more involved data collection efforts (that are often more costly). Furthermore, the density of the relationship data within each relationship category of highly precise coding schemes will often be inadequate to support full-blown SNA. In these cases, analysts are often forced to aggregate the more precise relationship data into more aggregate relationships (e.g., 'fought against', 'spoken out against', 'expressed dislike of' would be aggregated into something like 'negative affect') for running SNA metrics, algorithms and techniques. They would then refer to the more precise relationship types to gain a qualitative understanding of specific structures. The increased effort of collecting the more precise relationship data would add some value to the analysis in these cases, but it would likely not be a cost-effective use of government resources.

Avoiding the Pitfall

First, analysts and project managers engaging in an SNA project should begin by developing a clear and concise description of the entities and relationships of interest for data collection. Analysts and managers must focus on the specific types of entities they think will be of interest, and determine how they should be differentiated based on an understanding of (1) the analytic question, and (2) the array of 1-mode and 2-mode SNA techniques that would be appropriate. Analysts must also identify the range of behaviors and relationships that are relevant to the analytic question. A process for distinguishing between those types of entities and relationships of interest should be developed. Second, analysts and project managers should be able to utilize and differentiate between both one-mode relational data (e.g., ties between people) and two-mode (e.g., ties from people to organizations). Most of the commonly available academic short-courses in SNA cover the application of both one-mode and two-mode data analysis. In addition, agencies can procure specialized training to cover these topics in greater detail, if needed. Implementing these practices within an agency will help to ensure that the SNA conducted by the agency will treat the relational data in a consistent manner and apply principles of SNA appropriately.

Ideally, analysts and SNA project managers should communicate the coding expectations (the relevant data modes and relationship types to be gathered) in a codebook or by using established guidelines for coding SNA data. If an agency's SNA process is founded on a consistent data source and set of entities, the analytic process should be established with a consistent set of entity and relationship categories. Alternatively, if the SNA applications vary from project to project, along with the accompanying data sources, and entities and relationship of interest, development of a clear and concise description of the entity and relationship categories should be a standard part of the SNA process for each project in that agency. The latter case requires a higher level of resident SNA expertise and education investments at the agency, but the SNA capability will be more flexible and adaptable.

The appropriate use of SNA within and across modes (node types) and layers (relationship types) requires more than a group of well-trained analysts, but also well-educated analysts. There is a critical distinction between education and training in terms of the difference between methodology and tools. Training in the use of software tools for SNA is insufficient unless the analysts also have adequate SNA education. Education provides analysts an understanding of the use of SNA metrics, techniques and algorithms to develop valid inferences for further investigation and analysis. This is crucial to the appropriate use of multi-layered relationship data and mixed modes in SNA.

[5] For instance, an individual that is highly central in negative affect, but also highly central in positive affect suggests that the network is likely polarized – a feature that would lead to a more detailed SNA investigation of the grouping of positive and negative ties.

3. *Invalid applications of SNA*

There are two primary ways to "do SNA wrong": (1) applying unproven metrics and techniques to SNA problems, and (2) applying proven metrics and techniques incorrectly. SNA stands to provide substantial capabilities for intelligence analysis in a number of contexts, but it must be applied soundly, rigorously and consistently.

SNA fundamentally involves the application of metrics, techniques and algorithms that are founded in theories of human behavior and reinforced by empirical evaluation and testing. While innovation is respectable, unproven or untested innovation can be dangerous, leading to invalid results, and spurring a chain of analysis that could mislead a decision-maker. Likewise, the SNA literature has found many fundamental limits to the valid application of these metrics and techniques. It is crucial to carefully apply the metrics and techniques of SNA consistently with established best practices. Analysts must keep in mind that while some techniques may be well-suited for some problems, they may not be for the analyst's intended application.

Applying unproven metrics and techniques

The strength in applying SNA research as a starting point for analyzing relationship data for defense and intelligence applications lies in the track record associated with the literature. The metrics, techniques and algorithms widely used in that literature are based on a sound theoretical foundation from the social sciences. This provides a safer starting point for adapting and developing techniques for intelligence and defense applications.

Developing algorithms and metrics specifically for a single application or set of applications is sometimes warranted; however, agencies must do so with care. There are two main flaws some agencies face when applying unproven metrics and techniques. First, inventing solutions where there is already a viable solution, or even partial solution, can misdirect the effort and cause inefficiencies. If the new method is an improvement, it should provide a direct comparison to the existing approach. This provides a more confident foundation for the innovation.

Second, developing new methods is often done without adequate testing and evaluation due to operational and resource constraints. There are many instances of researchers and analysts developing techniques and algorithms from computer science, applied mathematics and operations research. Often, these approaches are developed for one particular SNA application and never tested for accuracy or validity. Developing new SNA techniques or applying existing techniques to new problems or data sources should include testing and evaluation.

In contrast, some SNA metrics and techniques have been applied on relationship data, identifying power, influence, and latent community groupings in networks over many decades and across many different domains and information sources. The robustness, or sensitivity, of some of these techniques to error in the data has been tested and explored, finding which metrics and techniques are safer to use when analyzing error-prone data. However, caution must be exercised when applying SNA techniques as well. Not all metrics and techniques have been tested, and not all can be consider well-proven for defense and intelligence applications. This, in fact, is the part of the discussion below concerning applying metrics and techniques incorrectly.

Applying metrics and techniques incorrectly

The application of existing metrics, techniques and algorithms incorrectly – applying them in unproven ways, on unproven data sources, or in distinctly invalid situations – should be avoided. SNA offers a vast array of metrics and techniques for analyzing human networks, but most of these metrics, techniques and algorithms are not validated for defense and intelligence applications. Data sources in the defense and intelligence communities provide distinctly subjective views of the network of relationships; they are generally incomplete, biased and provide heterogeneous coverage. This leads to substantial issues in the application of many SNA metrics and algorithms. Analysts and agencies must understand which algorithms have been tested and found robust, and where the limits to validity lie for these techniques. Alternatively, applications using such intelligence as a data source must be tested for sensitivity and robustness. Blindly applying SNA to an untested data source can be just as dangerous as applying an untested algorithm or technique.

SNA metrics, techniques and algorithms can be safely used to find structural features within relationship data, but the interpretation and meaning of those features may not be valid for all applications and data sources. For instance, the use of certain centrality metrics on well-defined (i.e. has relatively low error rates in the data) political elite networks in the United States is well-established, but the application of those same metrics on email networks for a US corporation may provide misleading[6]. More generally, many algorithms for finding cohesive subgroups in networks, for measuring regular equivalence, and other similar features have not been tested on many types of intelligence data and their use should be avoided until such time that these other algorithms have been tested successfully.

Avoiding the pitfall

There are two principal requirements to avoid this pitfall. First, senior technical leaders in an agency must maintain awareness of the evolving literature in SNA. There are SNA communities of practice and conferences that provide valuable instruction and venues for interaction on the state of the art in SNA.

Second, the agency must systematically test the application of SNA for the specific analytic questions and data sources they intend to use. In essence, the analysis process should be validated against the available data sources and intended use, requiring adequate budget and time allocations. Perfect validation is neither necessary nor warranted. Analysis is not a flawless endeavor and analysts must make judgment calls on a

[6] Consider the use of Eigenvector Centrality to uncover highly influential and powerful individuals from email traffic in a corporation. Eigenvector Centrality is heavily influenced by dense regions of relationships – e.g., cliques. Dense regions of connectivity in email traffic are often the result of email exchanges from group distributions, and are more likely an artifact of the email exchange a company sets up that identifying the powerful elites in that corporation.

regular basis. The needed validation speaks only to the requirement for increasing confidence in the analytic methods and data sources. It may be wise to seek out other agencies with similar analytic problems to collaborate on the testing. These other agencies may have already done some or all of the required testing, or could collaborate to perform or fund the requisite testing.

Testing and evaluation of methodology is a recognized field and has established standards and guidelines. Validation and verification standards are commonly applied to analytic methodologies. The US Defense Department has well-established guidelines and standards, governed and managed by the Defense Modeling and Simulation Office (DMSO). The services have established their own extensions and explanations of DMSO guidance as well. These standards identify the critical need to establish the validity of the modeling and analytic approach for a given application and data source, answering the question: Can I be reasonably confident that this methodology will provide the correct answer or insight? Doing so involves evaluating the internal validity – the consistency of the logic underlying the methodology – and the external validity – the consistency of the methodology with reality.

4. Designing a methodology around a tool

SNA tools should follow from the development of a SNA methodology tailored to the specific context of the agency in question. SNA is a flexible methodology that can be adapted to a wide variety of tools and technologies. Sophisticated tools can be brought to bear depending on the requirements and technological infrastructure of the agency. For example, simple scripts can be developed to supplement Microsoft Excel. A methodology should be designed based on a variety of criteria including the analytic requirements of the agency, the expected resource and cognitive capacities of the analysts that will perform the SNA, and the desired work stream that will be implemented. Given this perspective, designing a methodology around a tool, metaphorically speaking, puts the cart before the horse.

More generally, SNA is a set of statistical techniques derived from theoretical models of social interactions and structural formation. As such, when designing a tool to address a specific analytical task, it is paramount to be cognizant of the how the underlying theory of a given network analysis metric or technique applies to the analytic question a tool is intended to address. Ignoring this may result in a misapplication of a technique, or worse, development a tool with endemic flaws that produce misleading or nonsensical results.

Tools are developed with particular requirements in mind, not necessarily the requirements of the agency. This is not meant to imply that custom tool development is the only solution – in fact, a wide variety of flexible SNA tools can be chosen from. Analytic requirements should guide the development of the methodology; the identification, development and integration of tools should follow from the methodology. Software tools, from the simple to the sophisticated, must meet the requirements of the methodology that has been developed. Sometimes custom tool development is an option, but that is rarely cost-effective. Most often a combination of Government-Off-The-Shelf and Commercial-Off-The-Shelf software can be brought to bear to implement even the most sophisticated SNA approaches.

Avoiding the Pitfall

Avoiding this pitfall requires the commitment of resources – time, people and thought leadership – to establish the analytic requirements that SNA is intended to meet, and the development of a methodology in advance of identifying and procuring the software tools. Often, support for SNA at an agency comes as a result of a particular software tool being marketed or even sold to the agency. The result is often a backwards-engineering of methodologies to “fit” that tool. It is preferable that this situation be avoided whenever possible.

Developing the methodology should be accomplished with experts in SNA, keeping in mind capabilities of the analysts, the data sources that will be available and the analytic requirements that the methodology is expected to address. Some consulting firms possess much of the required expertise, but government agencies need to keep in mind that consulting should augment resident expertise within the agency, not serve as a substitute for it. Further, a variety of perspectives are warranted and getting outside expertise from other agencies, consulting firms and academia could substantially strengthen the resulting SNA capability.

Methodology development can be accomplished using a variety of approaches, including bottom-up design, top-down design and hybrid approaches. Bottom-up design is often the one implemented, and it involves developing SNA techniques and processes for a variety of individual analytic requirements. Over time, repetition of the bottom-up approach leads to a variety of SNA techniques and processes that may not seem logical or complimentary. As these individual SNA processes are developed, they can use contradictory or conflicting assumptions about the data, the analysts and the requirements, because they were not coordinated. Alternatively, the top-down design approach alleviates this problem by having a central individual or body identify the intended applications, identify the assumed data sources and intended expertise of analysts.

Effective design from the top-down will usually provide for logical, consistent and complimentary SNA techniques and processes. However, this approach can take considerably longer to implement and involves risks. The primary risk is that ineffective management of the top-down design process will lose touch with the analytic objectives, potentially resulting in an ineffective and inappropriately focused set of techniques and processes for SNA. Sometimes, a hybrid approach that combines these two can be highly effective, combining elements of top-down design and bottom-up design within a spiral development scheme. An option for a hybrid approach is to coordinate bottom-up SNA development through a governing body – analysts driving the development, but the coordinating body approving choices to enforce consistency in approaches, tools and assumptions about analysts and data sources.

5. Ignoring context in SNA

Quantitative analyses – analyses based on numerical data, such as statistics, social network analysis, simulation or some other methodology – can yield very powerful insights. Such approaches add rigor to analytic inquiries. However, one must be very careful when applying these methods and interpreting their results. SNA, like statistics, can be used to find patterns within large datasets. Using SNA only for this reason is trivial, given the advanced state of computational platforms and SNA algorithms available today. The question is not whether the patterns exist, but whether the patterns are meaningful in the context of the data being analyzed.

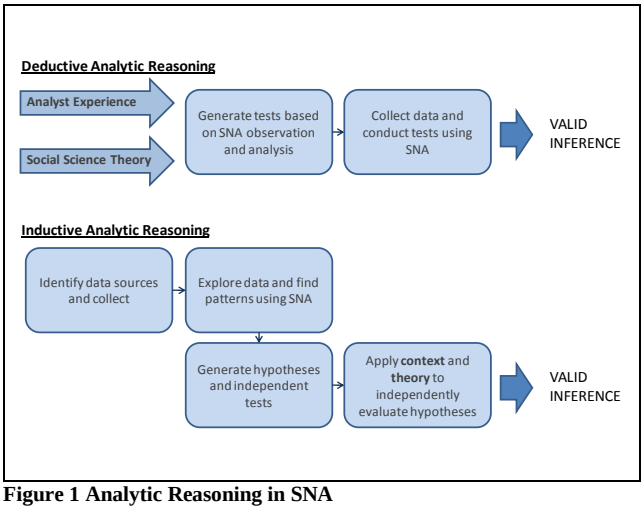

Figure 1 Analytic Reasoning in SNA

Even random networks can have nodes with seemingly interesting structural features, despite the fact that all of the structure is meaningless – it is randomly generated. This fact highlights the risk of doing SNA independently of the underlying human context. To draw inferences about the patterns that exist, analysts must establish a meaningful context in which to interpret and understand those patterns. Analyzing the patterns from the data alone will lead to spurious inferences and false conclusions. This is not meant to imply that inductive reasoning is an invalid approach in these cases. Inductive reasoning can serve as a sound basis for analysis, as long as the appropriate theory serves as a guide to the data used and initial interpretations that result, and as long as secondary, independent information sources are used to confirm the initial interpretations. For example, using SNA techniques on telecommunications data (call patterns) could be perfectly valid as long as the meaning or interpretation of the nodes, ties or groups of interest was independently verified by other sources of information. The SNA would serve as a heuristic to identify items of interest in the network data that would then be investigated to determine the actual meaning of the discovered pattern in the data.

Figure 1 provides a comparison of the logic flow for inductive and deductive analytic reasoning. Both are perfectly valid as long as sufficient precautions are used to avoid the relevant logical fallacies (inferring the general from the specific, and inferring the specific from the general, respectively). Applying qualitative analysis of the context, independent sources of observation and knowledge, and additional social science theory to confirm findings can provide valuable precautions against spurious correlations when performing inductive analytic reasoning with SNA. Sound theoretical reasoning from the beginning and rigorous empirical validation in similar cases is generally required as a precaution in deductive analytic reasoning.

In either case, context is vital to making the SNA output understandable and actionable. The details of the situation being analyzed lend vital clues as to the reason networks have organized the way they have, or why individuals have formed certain structural advantages in a given network. These details are what make the results of the analysis actionable and understandable. Providing the structural features alone (i.e., the output of SNA algorithms) in the analytic report provides little understanding of the network in question. Consider the difference between simply providing a list of centrality scores for actors and providing a descriptive narrative of the network, its history and the attributes of its key members, coupled with network diagrams and centrality scores. Reporting only the structural features provides little for operators or decision-makers to act on, but the broader context can provide sufficient depth and clarity that action could result.

Avoiding the pitfall

Failure to include deep analysis and contextual detail in an SNA will result in analytic products that will fail to connect to the operator, decision-maker or other intelligence analysts. In nearly all cases, the consumer of a network analysis product will not be familiar with the methods or techniques, and therefore will need to understand the analysis in terms of its context to their particular analytic requirement. To avoid this situation, there two primary processes that should be incorporated into any implementation of SNA: 1) Collaboration with Qualitative Experts, and 2) Independent Data and Information Sources. These recommendations should be considered independently. They can be incorporated into an implementation of SNA individually or in combination.

First, it is recommended that SNA be done in collaboration with experts – either external experts or intelligence analysts with more detailed contextual knowledge of the network being studied. This collaboration will enable the SNA analyst to learn and understand the "story" of the network and to communicate that story effectively to others. This approach will provide the added benefit of providing additional knowledge to the qualitative expert on the network of people being studied as well. This benefit should not be considered lightly because in nearly all cases the benefit of the SNA will be greatest for the expert who already understands the people to some extent; then like any informative formal analysis, the SNA can test hypothesis, reinforce or challenge intuition the expert holds, provide deeper insights into observations they have made, and provide new understanding of how the group being analyzed operates.

Second, like any quantitative data analysis, SNA is more effective when multiple data and information sources are used. Multiple data sources are best in feeding the quantitative analysis, but additional qualitative information is also highly valuable to an effective analysis. For instance, reading previous

reporting from intelligence and open sources can provide valuable qualitative insights into the quantitative features in the network data. Other types of independent information could include interviews with people familiar with the members of the network in some instances, biographies of the people involved in the network, and interviews with outside sources of information on affiliated organizations, events or other types of entities contained within the network data.

6. Inappropriately handling error in network data

Network data are never perfect. Effective and accurate SNA must take this into account, and determine the appropriate balance. There are two sides to this issue: First, imperfect data is a fact of life and many of the techniques in SNA should be used with extreme caution when the data are highly imperfect. Second, imperfect data is a fact of life and therefore any aspiration to acquiring complete and perfect data for a network is not realistic. The consequence of these two sides of the same issue is that proper applications of SNA require careful data analysis along with contextual validation and interpretation.

Ignoring errors in the data

Much of the academic literature on SNA ignores the issues of uncertainty and error in network data. Most of the techniques behind SNA implicitly assume perfect information in the data[7]. They are based on graph theory, which reflects the features of networks as they appear in the data, not as they might appear in reality. However, in SNA the reality is often different than what the data shows us, and we must account for this, especially in the realm of covert networks actively attempting to hide their signatures from observation. Analysts are ill-advised to ignore the error in their network data, even if they cannot fully account for that error. Understanding the impacts of these errors is complex, and in most cases cannot be done with understanding where in the network the errors are occurring.

The critical realization is that imperfect data has its most direct impact on metrics, not the analysis and conclusions. Given the existence of imperfect data, the burden of interpretation falls on the analyst to derive meaningful conclusions from that flawed data. The analyst must use correct metrics given the likely presence of multiple forms of error in the data, and buttress the interpretation of the SNA output with sound reasoning and secondary source of information and insight. This is where contextual information can be most valuable, as it can help disprove a logical contradiction or false identifications derived from an incomplete view of a network.

Ignoring the value of analysis on imperfect data

At the same time, analysts must also understand that perfect data are not required for the valid application of SNA. SNA on imperfect data simply requires that the analyst place the appropriate caveats on the conclusions regarding the roles, positions or functions of the entities in the network in question, or those findings be verified by independent sources of information (see Pitfall 5 above).

[7] There is a growing literature associated with issues surrounding data quality and data sampling, however, many of the data quality issues associated with analyzing covert networks are not fully understood.

Many techniques in SNA are robust against substantial levels of erroneous data. The key parameter is network density – in regions of the network where density is high (like core regions), there is a high degree of redundancy and higher levels of error can be tolerated for SNA. Where density is lower, the tolerance to error in SNA techniques is much less – more caution must be exercised. For example, if we assume that some terrorist networks are centrally organized, and that those central regions of the network have higher density of connections, analysis of the leadership using SNA will be more effective than on the peripheral regions where the operators (e.g., IED implanters, suicide bombers) sit. Analyzing those peripheral regions can still be accomplished effectively, but the SNA techniques must be more conservative in nature, relying more on qualitative data and context.

Some research has focused on comparing the robustness and sensitivity of specific SNA techniques and metrics to errors in network data. For instance, In-Degree Centrality has been shown to be highly stable in sampled networks [6]. Additionally, the nature of the errors also impacts the robustness of other SNA techniques. This is a critical reality that must be realized in conducting SNA. Patterns in the error in network data can have direct impacts on the applicability of network analysis techniques. Most research into sensitivity and robustness has assumed errors randomly occurring in the network data, while it can be assumed that the biases in collection of data on covert networks are systematic.

The appropriate application of SNA must recognize the impact of error in the data, and verify SNA findings with secondary information sources and expert inputs.

Avoiding the pitfall

The solution to avoiding this pitfall lies both in analytic policy and in effective education and training. Agency policies on the appropriate use of data sources and quality of those data sources must be carefully designed to support analysts, while safeguarding the customers of those analysts. Further, effective education and training is needed to provide analysts the knowledge, skills and capabilities to operate with imperfect data more effectively and with confidence.

Establishing effective policies for data standards is crucial to supporting useful, effective and reliable analytic products using SNA. Analysts must operate under policies that do not require excessive certainty in every individual piece of data in a network dataset. However, the data quality standards must be established to provide the users of the analysis appropriate confidence in the findings. In setting these standards, the agency must recognize that the analysis does not rest on every individual piece of data for valid SNA and that, in general, missing data is more impactful than false data. Policies should also establish the quality control into the process, hopefully involving senior SNA experts in the review of analyses prior to the analysis becoming a product. Involving experts in the review of analysis prior to production beginning can save time, establish additional high value SNA techniques to utilize in the analysis, and provide these inputs while there is still time to impact the product without delaying delivery to the customer.

To enable these effective policies, analysts must be given the knowledge, skills and capabilities to operate on imperfect

data. The requirement is to bolster the analysts' education and training so that they know what analysis is supported by the amount and quality of data that they have for any given project. They must develop sufficient confidence in their analytic results to be effective at operating with the data that is available, while not waiting for perfect data to be found. What constitutes "sufficient confidence" will vary by the intended use of SNA. The more critical the impact of incorrect assessments resulting from SNA relative to the impact of a correct assessment, the more conservative the standard for "sufficient confidence" should be. For instance, if the impact of the SNA application revolves around resource allocations, "sufficient confidence" standards can be much lower than if the application impacts decisions of life and death.

The key to enabling this feature is education and training. Analysts must be educated in "network thinking" and understand how people operate in networks. We also recommend that an educational element include effective mentoring of junior SNA analysts by more seasoned SNA analysts. This enables consistent quality standards and applications across an organization. Training provides analysts the skills to use the technology available at that agency to understand what analytic conclusions are supported by the data. Effective education and training will enable analysts to utilize the relational data that is available, understanding when they can reach conclusions about the network and its actors safely without requiring unrealistic and operationally infeasible timelines for the analysis.

7. Applying SNA to Inappropriate Problems

SNA is not always the right tool for intelligence and operational analyses. Recognizing when SNA is and is not appropriate is critical for the effective application of this methodology. There are no clear guidelines as to what SNA is not appropriate for, but there are some rules of thumb that should help guide decisions on the use of SNA.

Consider a simple example where SNA is applied to studying road networks in the United States in order to identify the roads that are most important to commerce and need to be protected and maintained. The roads can be considered links between junctions and the junctions can be the nodes. SNA holds that centralities based on shortest paths are a valid approach to understanding the importance of people – this has been reinforced in countless studies. However, if we were to apply this assumption to road networks, we would be missing something very important – shortest paths are not the only paths of relevance in road networks. There are a variety of secondary factors that influence route choice in road networks, and this fact can seriously threaten the validity of shortest path based SNA centrality metrics for road networks. It does not mean that SNA techniques will not find interesting features and patterns in road networks, however, the interpretation and meaning must be carefully reconsidered based on a different set of assumptions than have proven effective in human networks.

We propose a set of rules of thumb for the appropriate application of SNA. These rules of thumb are not universally agreed upon, but they are a starting point for consideration of when SNA is and is not appropriate. The following are problems considered to be appropriate for SNA:

- Problems involving human behavior, involving a large number of people (10's to 1000's of actors). Larger numbers may require different graph theoretic techniques than have been found in SNA[8]. Smaller problems do not necessarily require the application of SNA, though it can be used.
- Problems dealing with functional groups, for example:
    - Uncovering illicit business dealings among a limited number of companies
    - Determining how to influence political decision-makers in a government
- Problems dealing with the diffusion or flow of resources, information, influence among people
- Problems dealing with collaboration, coordination or reactions across interpersonal connections of people

SNA is likely not appropriate for:

- Problems in which the perspective of entities and their relations does not map well (e.g., aggregating opinions of unconnected individuals)
- Problems dealing with long-term forecasts of behavior, patterns or outcomes
- Problems analyzing military operations, logistics and force planning
- Problems where behavior of very large numbers of people is involved – on the order of hundreds of thousands of people or more (e.g., forecasting extremist Islamic support)
- Problems where precise models of physical or human behavior are available (e.g., where game theory could be used)

Applying graph theory and SNA techniques to non-human data should not be considered doing SNA. These techniques may be validly applied to non-human networks, but the application is not SNA.

Avoiding the Pitfall

The best solution to this pitfall is to apply an appropriate level of expertise and common sense. Agencies and analysts must carefully consider a problem or set of problems, evaluating what available methodological approach is appropriate, and whether SNA, specifically, can confidently lend any understanding to the problem.

V. CONCLUSION

In this paper, we have described seven common pitfalls in the attempted application of SNA in the defense and intelligence communities. These seven pitfalls are:

[8] Although many SNA algorithms can be run on very large networks, there meaning may over-interpreted. For example, a single phone call between otherwise disparate parts of the global telephone call graph could radically change the betweenness centrality score for a node hundreds of hops from both the sender and receiver of the call.

1) Confusing Link Analysis with SNA
2) Failing to differentiate modes and layers in networks
3) Invalid applications of SNA
4) Designing a methodology around a tool
5) Ignoring context in SNA
6) Inappropriately handling error in network data
7) Applying SNA to inappropriate problems

These pitfalls are found in various places within the community, in government agencies as well as in industry and academia. Agencies already using, or considering using SNA to support their missions are urged to consider the recommendations of this report carefully.

- *Establish effective education and training requirements for SNA analysts.* Use external education programs, or design tailored programs for the agency. Training programs should focus on the tools available for SNA within the agency, and should include material that is tailored to handling the data sources relevant to that agency. Analysts and managers must educate themselves in
    - SNA applications and techniques at a basic level, not just the database and analytic tools.
    - The array of 1-mode and 2-mode analytic techniques, understanding the differences
- *Establish cadre of senior technical SNA leaders* – either internal to the organization or contracted from outside. Senior technical leaders must maintain awareness of evolving SNA literature – keep current. These senior technical leaders should also be involved quality control, as well as the mentoring, education and training of junior analysts.
- *Effective analytic policies must be put in place*. These policies should strive to balance the need to enforce minimum standards for SNA data in order to protect customers of the analysis with the need to allow analysts to take risks in the data, using the SNA process to weed out the bad data from the conclusions. These policies must also establish effective quality control procedures involving senior technical leaders to review the approaches and analyses prior to the final production stage.
- *Develop clear, concise coding guidelines*. Preferably, these guidelines should be captured in official documentation, or in a project-specific codebook.
- *SNA applications should be tested to the extent feasible and cost-effective*. This does not mean agencies must avoid applying SNA until academic levels of rigor have been achieved, but rather, they should collaborate with other agencies that may have tested SNA applications or could fund testing together.
- *Agencies should establish expected analytic requirements for SNA up front*, rather than in response to the observed capabilities of a software tool. This endeavor should involve both internal and external experts in both the analytic problems the agency faces and in SNA applications. The outside experts can come from other agencies, academia, or industry.
- *SNA analyses must be buttressed with qualitative expertise and information*. SNA analyses should involve:
    - Qualitative experts in the network in question, or at a minimum, the SNA analyst should be intimately familiar with the underlying data
    - Secondary data and information sources to corroborate the SNA findings

These specific pitfalls are not meant to be a comprehensive list, nor are they all hard-and-fast rules. Instead, we view this as the beginning of a conversation. For SNA to be most effective, we must be careful to apply it correctly, knowledgeably, and in the proper contexts.

REFERENCES

[1] S. Ressler, "Social network analysis as an approach to combat terrorism: Past, present, and future research," Homeland Security Affairs, vol. 2, article 8, July 2006, http://www.hsaj.org/?article=2.2.8.

[2] M. K. Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects," Social Networks, vol. 13, pp. 251-274, September 1991.

[3] A. M. Huberman and M. B. Miles, "Data management and analysis methods," in Handbook of Qualitative Research, Y. S. Lincoln and N. K Denzin, Eds., Thousand Oaks, CA: Sage, 1994, pp. 428-445.

[4] M. B. Miles and A. M. Huberman, Qualitative Data Analysis: An Expanded Sourcebook, Thousand Oaks: Sage, 1994.

[5] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.

[6] E. Costenbader and T. Valente, "The stability of centrality measures when networks are sampled," Social Networks, vol. 25, pp. 283-307, 2003.