Data-Driven Diffusion Modeling to examine Deterrence

Michael J. Lanham, Geoffrey P. Morgan, Kathleen M. Carley
The Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15213
mlanham, gmorgan, kathleen.carley@cs.cmu.edu

Abstract— **The combination of social network extraction from texts, network analytics to identify key actors, and then simulation to assess alternative interventions in terms of their impact on the network is a powerful approach for supporting crisis de-escalation activities. In this paper, we describe how researchers used this approach as part of a scenario-driven modeling effort. We demonstrate the strength of going from data-to-model and the advantages of data-driven simulation. We conclude with a discussion of the limitations of this approach for the chosen policy domain and our anticipated future steps.**

Keywords- Text Mining, Network Models, Belief Diffusion, Rapid Prototyping

I. INTRODUCTION

Deterrence calculus between nation-states requires balancing the costs and benefits of restraint with the costs and benefits of action. Many authors have written about deterrence and proposed methods for achieving and maintaining a balanced state of affairs between nations. Of course, achieving deterrence, in particular with respect to long-simmering areas of conflict and between nuclear powers, is fraught with difficulties.

A significant limitation for national decision makers and their advisors is the lack of tool sets for rapid development of meaningful models that allow safe experimentation. Without experimentation, decision makers must rely on their judgment to assess the status quo, the effects of their proposed actions, and the chain of events branching from those actions. Predicting the secondary, tertiary and further consequences of action rapidly becomes impossible

To address the need to experiment and develop an experiential basis for judgments about deterrence, the researchers participated in a multi-year effort to model potential adversaries as well as friendly forces. In that effort, we used a multi-modeling approach to evaluate deterrence but focus this paper on our approach for rapidly developing useful models for examining the diffusion of beliefs in networks of strategic decision makers. Our rapid development approach focuses on converting large amounts of unstructured texts into rich multi-mode and multiplex relational networks for use in dynamic and stochastic simulations.

In the remainder of this paper, we discuss our process for rapidly developing useful simulation models of diffusion through semi-automated analyses of text corpuses; how we applied the approach to a specific crisis scenario, and our lessons learned.

II. THE DATA-TO-MODEL PROCESS

The data-to-model process is intended to be a systematic and computer-assisted repeatable approach with these steps [1, 2]:

1. Collect data
2. Clean the text corpus
3. Ontological Cross Classification
4. Generate Data for Analysis and Simulation

Collecting data is the first step in our process. Data can be structured, semi-structured, or unstructured. Structured data is available from sources such as databases or existing relational data-files. Semi-structured data includes material such as discussion posts, where, for example, the user's ID and the post's time-stamp are known, but the content of the post is unstructured. Unstructured data includes raw text, the content of news articles, and the body of a technical report. Our data-to-model

This work was supported in part by the Office of Naval Research - ONR - N000140811223 and by the Air Force Office of Sponsored Research –MURI, 600322. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Air Force Office of Sponsored Research or the U.S. government.

978-1-4577-1051-3/11/$26.00 ©2011 IEEE

process focuses on the challenges associated with unstructured data, although other forms of data are useful in the final analysis.

Once gathered, the data needs *cleaning*, the second step in our process. Text data, like all written language, is often rife with ambiguity. The purpose of cleaning the data is to remove and/or clarify ambiguous or redundant references. There are several sub-phases of note in this process. De-duplication is the process of removing identical articles in a text corpus – duplicate articles inflate and bias the process results. Named Entity Identification develops a list of proper nouns in the text corpus, which preserves these entities against premature filtering. Text Refinement includes a) converting all verbs to present tense and to base forms, b) removal of stop or noise words (e.g., prepositions, helping verbs), and c) application of n-gram thesauri, which allow multiple words from the corpus to be represented by a single concept word. For example, the distinct verbs {"meet", "call", "email"} might be reduced to {"communicate"}.

Concepts and words from a particular text corpus may have unique meaning because of the context and content of that text corpus. The third and penultimate phase, *Ontological Cross Classification*, addresses this aspect of text data. As a trivial example, the word "battery" is likely to have very different meanings if the text corpus draws from children's toy manuals as opposed to military history texts. An analyst assigns each concept that remains in the cleaned text corpus to one or more of several classes. These classes are Agents, Roles, Organizations, Events, Locations, Tasks, Beliefs, Resources, and Knowledge [3, 4]. Automation aids this process but tailoring to each text corpus represents the bulk of the human effort in this process.

Iteration through each of the previous three steps occurs as many times as is appropriate to the question(s) at hand. The data-to-model process creates intermediate artifacts, allowing the process to be run both painlessly with new data or to be continually tweaked to improve the resulting data.

The final step, *Generate Data for Analysis and Simulation*, produces the multimode and multiplex networks that an analyst can then be used both for static analysis and as inputs to diffusion simulations. Although the exact type of machine-generated networks will depend on the text corpus, prominent generated networks include social networks "who knows who", knowledge networks "who knows what", and assignment networks "who does what". The networks generated from three node classes are included in Table 1. An important difference between these networks, and traditional network science's focus on agent-by-agent interactions, is the inclusion of the non-agent node classes in the networks and in the analysis [6]. The full set of networks that can be generated based on these entity classes is included in Appendix 1 [3,4,6].

III. THE INDIA-PAKISTAN CRISIS SCENARIO

We used this data-to-model approach as part of an evaluation of the effectiveness and usefulness of a selected set of modeling methodologies. The scenario is entirely fictional, but plausible given past history between these two nations. The scenario for this fictional situation used a mixture of fictional scenario-injects and real-world events and interactions along with real names for people and places. The location of this scenario is along the disputed territorial border regions of Jammu and Kashmir between India, Pakistan and China. This crisis scenario begins with a fictitious raid into the parliament building of Srinagar, India by gunmen on 2 June 2002. The scenario continued to 5 August 2002 with a number of actions by Pakistan, India, the United States and a small number of other countries of interest. We performed the analysis from two perspectives: USPACOM and USCENTCOM.

A. Applied data-to-model process

We used LexisNexis® data, as well as scrapings of governmental web sites from Pakistan, India, and the United States. The LexisNexis® data were 3,000 text files representing newspaper articles meeting the search criteria[1]. We rapidly realized the data had

TABLE I. NETWORKS FROM THREE NODE CLASSES[5]

	***A**gents*	***K**nowledge*	***T**asks*
***A**gents*	Social	Knowledge	Assignment
***K**nowledge*		Information	Needs
***T**asks*			Precedence

[1] The selection criteria within LexisNexis® were the inclusive dates of the three scenario vignettes (20 Jun – 5 Jul, 5 Jul - 22 Jul, 23 Jul – 5 Aug) and the words "India" and "Pakistan."

insufficient overlap for agents and organizations clearly relevant to an international border crisis situation—it was missing Pakistan's Inter-Service Intelligence (ISI) agency, the Director General of the ISI, India's Defense Secretary, US Combatant Commands (COCOMs), the US Department of State (DoS), and agents from those organizations. We then scraped each nation's national security apparatus (their equivalents to the US National Security Council (NSC), Department of Defense (DoD), and DoS) official web sites. To collect data on how the DoD interacts with those countries, we web-scraped the official web sites of USCENTCOM and USPACOM. The US uses these two geographic COCOMs to execute the military and, to a limited extent, other forms of national power in their regions [7, 8]. After these web scrapes, there were approximately 27,000 text files to support creation of network models.

This research built the various thesauri and delete lists from scratch as there was no contingency planning staff from which we could borrow. The development of the generalization thesaurus, meta-thesaurus, and delete lists took approximately 160 man-hours. The project specific thesauri added 962 entries to previously built collections of generalizations and meta-ontology thesauri. Table 2 describes the end-state of the text corpus after data collection and approximately 10 rounds of data reduction.

Disambiguation of terms, as well as accurate categorization of concepts into the meta-matrix ontological categories proved tedious but not complex. Because the researchers are not SMEs in Pakistan, India, or Kashmir-Jammu, categorization, deletion, and generalization required an investment of resources to reconcile the multiple references to the same concept (e.g. person, place) in multiple texts with slightly different verbiage. We established three categories for persons of interest to this effort: NSC-level agents; diplomat agents; and national-level political agents (e.g. leaders of political parties). Identification of specific persons relevant to a border-crisis scenario was an iterative process of identifying a concept or set of concepts (e.g. a n-gram) then using web-based searches to determine the source actor for those concepts. For example, given the scenario time-frame, {"The President of the United States", "President Bush", and "George W. Bush"} all represent the same person and could be reduced to "President George W. Bush". This iterative process allowed us to reduce the agent set to a total of 47 named individuals across three countries.

TABLE II. NODE COUNTS, PER COCOM, VIGNETTE A & B

Node Type	**Vignette A**		**Vignette B**	
	PACOM Count	***CENTCOM Count***	***PACOM Count***	***CENTCOM Count***
Agents	42	47	42	47
Belief	21	21	32	32
Event	40	40	22	22
Knowledge	145	145	148	148
Location	3250	3250	3303	3303
Organization	321	321	325	325
Resource	116	116	117	117
Role	247	247	244	244
Task	418	418	424	424

To further reduce the set of actors, we eliminated agents not directly connected to strategic decision-makers. We used measures such as degree centrality, betweenness centrality, and eigenvector centrality – these measures provide similar but different aspects of a node's criticality in a network – to understand the importance and relevance of individual nodes. Relevance was, at times, immediately apparent, and at other times required a return to source documents as well as web-based searching.

B. Network Analysis

We divided the data set into the three time periods in the scenario. These three periods, which we called vignettes, represented the initial crisis incident plus eight days (Vignette A); the mid-crisis period when the two US COCOMs were using independent analysis and actions (Vignette B); and the last period was when the two COCOMs would, in the scenario, collaborate and merge their respective models and COAs to present to US national leadership (Vignette C). We merged the modern-day (circa Aug/Sep 2010) web scrapings with the time-period data drawn from LexisNexis®. For each vignette, we used network analysis software to calculate important node and network measures as well as to visualize the interconnections of these strategic decision-makers

Over the three vignettes, we were able to discern shifts in relative rankings of the nodes of interest. Fig. 1 is a key agent report from the USCENTCOM perspective, thus focused on US and Pakistani

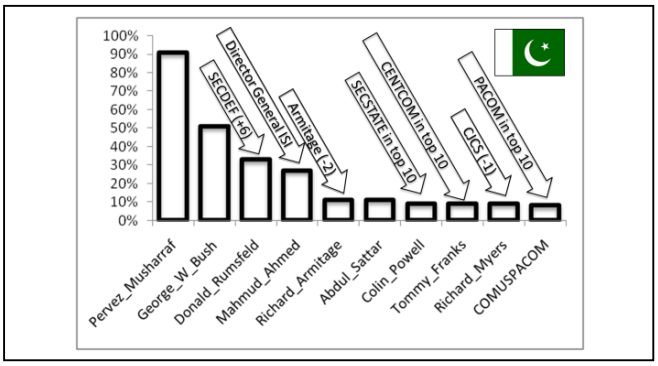

Figure 1. The change in actor relevance indicates that the scenario is shifting from a diplomatic to a military situation.

agents, at the conclusion of Vignette B. This chart, generated with ORA 2.2.2a, depicts the top ten agents that consistently appeared in 22 different network measurements. As can be seen in Fig. 1, the Presidents of the two countries are clearly in the top ten, as well as the head of the ISI, the Secretary and Deputy Secretary of State, and key US military commanders and advisors. In this chart, but not in the chart from the previous time period, Vignette A, are the commanders of USPACOM and USCENTCOM as well as the SecState. Their presence in this chart, in this time period, is consistent with an interpretation that the scenario is rapidly moving from a diplomacy-centric situation to one involving the US military. At the same time, the diplomacy instrument of national power is increasing its level of effort by incorporating the SecState himself, and not simply his subordinates. The drop in relative ranking of the Chairman of the Joint Chiefs is consistent with an increasing presence of both COCOM commanders, in direct discussions and interactions with the President. Their direct involvement with the President is consistent with the DoD moving from planning for action with the CJCS as the principal military advisor to executing action through the NCS to the COCOMs.

C. *Dynamic Analysis through Diffusion Simulations*

We performed not only a static analysis of the networks as generated, but also used those networks as an input to a model of information diffusion. We used a validated and publicly available model of information diffusion called Construct [9, 10], we used version 4.2. Construct is an agent-based and turn-based simulation. In this simulation, agents pick communication partners based on two preferences: similarity, also called homophily [11]; and knowledge seeking, a preference to interact with agents who possess rare knowledge [12, 13]. These agents exchange information, which informs their beliefs. The primary output measure of the interest was the number of strategic decision-makers who possessed a "pro-war" belief.

Using this simulation, we explored a set of questions. These were: 1) If the United States does not intervene, how many strategic decision-makers will possess the pro-war belief; and 2) given the scenario as defined, if all deterrence actions are taken, how many decision-makers will possess pro-war beliefs? Secondary questions included: 1) At what time-point in the crisis would the interventions have maximum impact; 2) how large a set of interventions is necessary to produce significant impact in the number of strategic decision-makers with a pro-war belief; and 3) Is there an interaction between the number of required interventions and the timing of those interventions?

To examine these questions, we implemented the specified scenario as a set of "provocations" and "responses" with a magnitude, a start-time, and an end-time. Provocations provided pro-war knowledge to the strategic decision makers. Responses provided evidence against the pro-war belief to these decision makers.

In our virtual experiments, our model predicted that, if the US and other outside states do not work to tamp down tensions, within thirty days of the scenario's start-date more than 60% of the strategic

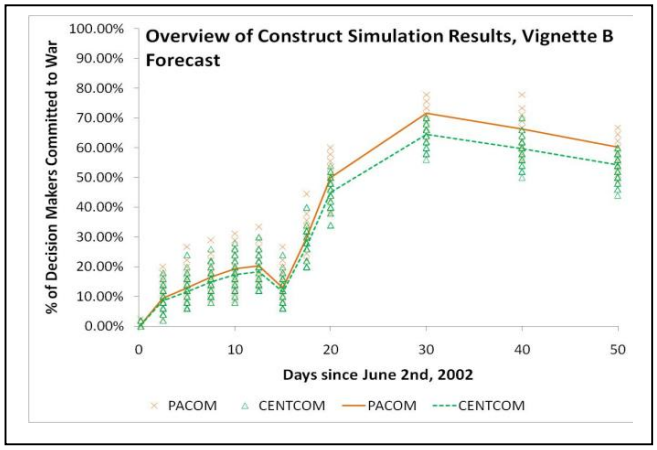

Figure 2. If left unchecked, our simulation predicts that the majority of strategic decision makers in both India and Pakistan will possess the pro-war belief within 30 days..

decision makers of both Pakistan and India believe that war is the right choice (as shown in Fig. 2). Our model indicated that the conventional US response outlined in the scenario document to this situation was insufficient – it produced useful changes in the minds of decision-makers but critical days after the majority of these actors have the "pro-war" belief. As Fig 3 shows, early interventions produced the most significant impact – as agents then chose to pass along anti-war knowledge on their own.

D. *Implications for policy*

The results of this analysis effort and the simulations of this scenario indicate several items of interest: other nations in the midst of hostile tensions can perceive US military action (even in the form of reconnaissance flights) as provocative; levers of deterrence must be found and used quickly and, at times, repeatedly to have an effect; continued provocations as perceived on each side will rapidly overwhelm US DIME options; early and fast action may not win the day, but it does buy time before models predict continued rising tensions, thereby allowing for additional actions to help de-escalate the situation.

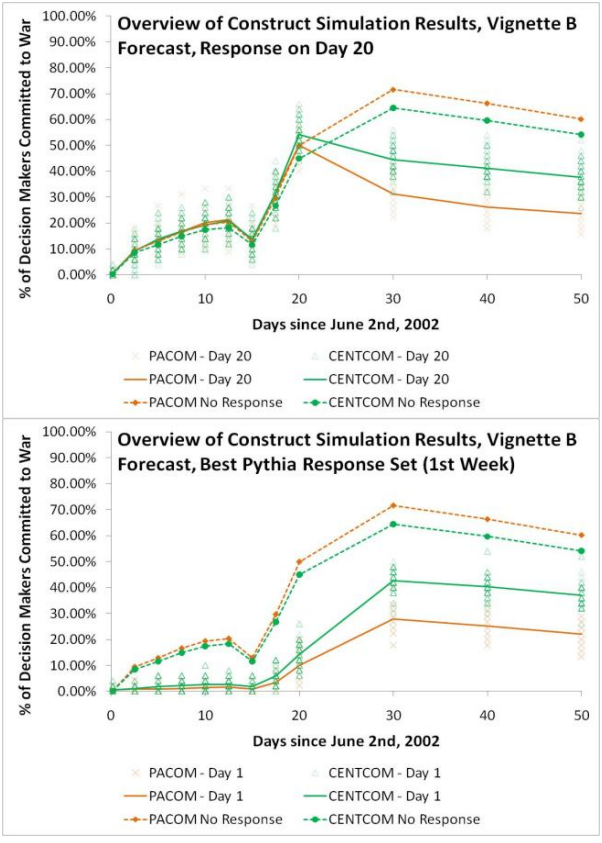

Figure 3. Early interventions produce a much more useful impact on strategic decision maker beliefs than late interventions.

Though none of these conclusions are necessarily earth-shattering, they were consistent across multiple sets of models informed by different assumptions and paradigms. This similarity in outcomes informs the decision makers that there is precious little decision space with which to maneuver.

IV. DISCUSSION

In this work, we have demonstrated a rapid and semi-automated process for converting large text corpuses into useful models of belief diffusion. Both static and dynamic analysis offered useful guidance to policy makers considering how best to apply national power to pursue deterrence objectives.

This data-to-model approach requires using a large number of tools and services. One contribution of this research was to define appropriate workflows. In the future, the use of such workflows can be facilitated by using a workflow management system for multi-modeling, such as SORASCS [14].

This research effort was a first iteration of the approach, applying the procedure to model deterrence. Although a useful demonstrator, there were some bumps in the road and problems to overcome.

The data collection method, data drawn from LexisNexis® based on date ranges and nation names, is not an accurate portrayal of the distinct information available to each COCOM staff. We believe that the utility and explanatory power of such models will improve as the data improves in quality and topicality. Further, each COCOM's information assets are likely to have distinct and important differences – and these differences may well lead to diverse final results. The models in the realized process are thus more likely to differ and yet also more likely to be useful.

Data cleaning efforts were focused on the agent and knowledge networks with relatively little effort spent removing 'noise' nodes from the other data sets. Given infinite time and effort, all node types would be cleaned. However, post-experiment cleaning of the data reduced the location node set

from over 3000 concepts to just over 1500 with no significant impacts on the analysis—validating the decision to conserve effort by cleaning only the node classes that factored directly into analysis. Thus, from our experience, errors seem unlikely to cross over into evaluations of other node-classes.

The diffusion simulations made some additional simplifying assumptions. Except in very limited cases, we did not attempt to profile the strategic actor set to determine their starting inclinations towards the pro-war belief. The simulation is able to use such information, but without access to domain experts that level of precision for actor beliefs seemed problematic.

The diffusion simulations relied on the agent-by-agent networks produced through the data-to-model methodology, but did not take advantage of the knowledge and resource networks generated by the approach. This was partially due to a relatively scarcity of knowledge and resource concepts being tied to these actors. With richer data sets, using these additional networks may be a better method of informing actor knowledge at start-time.

Policy makers are likely to not only want to know the number of strategic decision makers who possess the pro-war belief, but also *which* decision-makers are likely to have the pro-war belief at any particular time in the scenario. We believe that our simulations results are robust to trends but unlikely to be robust to individual prediction.

Other limitations included the deliberate exclusion of India's Cold Start doctrine [15-18] as well as India's 'no first use against non-nuclear states' policy [19]. Researchers also omitted Pakistan's published responses to the doctrine of Cold Start. One of the goals of Cold Start is to hide information from outside states that may attempt to interfere in India's strategic objectives. This goal, information hiding, outlines another limitation of simulation as used – we did not incorporate meta-cognitive reasoning into the simulation — agents being aware that others are attempting to influence them. This could be modeled with the existing simulation method as both a) adding resistance to information provided by outside actors, and b) introducing error in the perceptions of outside actors as to what information the strategic decision makers possess.

Follow-on efforts will need to incorporate a more sustained collection of textual and other unstructured data. Sustained collection and processing of such data will support the use-case's assertion that planning teams can rapidly and efficiently feed data into the developed processes. Sustained collection will also demonstrate analysis techniques to continually and effectively refine planning models. The collection of textual and other unstructured data will need to closely correlate to the developed scenario to ensure appropriate overlap of data without requiring researchers to create data injections to support automated analysis. The need to keep the scenario and collected data synchronized may suggest experiments using historical events in lieu of fictional actions.

ACKNOWLEDGMENT

The authors wish to acknowledge and thank the partner researchers at George Mason University's System Architectures Lab. Without the combined efforts of Alex Levis, PhD, Sayed Abbas K. Zaidi, PhD, Lee Wagenhals, PhD, Robert Elder, PhD, Tod Levitt, PhD, Ahmed Jbara Y Abu, and Syed Hasan Ali Rizvi, this project would not have met the successes it did. This paper represents only a small part of the work this project entailed.

References

[1] K. Carley, *et al.*, "Experimentation Testbeds: Using SORASCS to Run and Process HSCB Virtual Experiments," in *Human Social Culture and Behavioral Modeling (HSCB) Focus 2011: Integrating Social Science Theory and Analytic Methods for Operational Use*, Chantilly, Virginia, USA, 2011.

[2] K. Carley, *et al.*, "Rapid Ethnographic Assessment: Data-To-Model," in *Human Social Culture and Behavioral Modeling (HSCB) Focus 2011: Integrating Social Science Theory and Analytic Methods for Operational Use*, Chantilly, Virginia, USA, 2011.

[3] K. M. Carley, *et al.* (2010, 7 March 2011). *AutoMap User's Guide 2010*. Available: http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-10-121.pdf

[4] J. Diesner and K. M. Carley, "Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a Novel Method for Network Text Analysis," in *Causal Mapping for Research in Information Technology*, V. K. Narayanan and D. J. Armstrong, Eds., 2005, pp. 81-10.

[5] K. M. Carley, "Dynamic Network Analysis," in *Summary of the NRC workshop on Social Network Modeling and Analysis*, R. Breiger and K. M. Carley, Eds.: National Research Council, 2003, pp. 133-145.

[6] K. M. Carley, "Smart Agents and Organizations of the Future," in *The Handbook of New Media*, L. Lievrouw and S. Livingston, Eds., Thousand Oaks, CA, USA, 2002, pp. 206-220.

[7] Joint Staff J7, "Department of Defense Dictionary of Military and Associated Terms," vol. Joint Publication 1-02 (JP 1-02), Department of Defense*, et al.*, Eds. Washington, D.C.: Joint Staff, 2010.

[8] Training and Doctrine Command, "Field Manual 3-0 (FM 3-0) Operations ", D. o. t. Army, Ed. Washington, D.C.: Headquarters, Department of the Army, 2008.

[9] K. M. Carley, "Group Stability: A Socio-Cognitive Approach. ," in *Advances in Group Processes, Advances in Group Processes: Theory and Research*, Greenwich, CT: JAI Press, 1990, pp. 1-44.

[10] K. M. Carley*, et al.*, "The Etiology of Social Change," *Topics in Cognitive Science,* vol. 1, pp. 621-650, 26 June 2009.

[11] M. McPherson*, et al.*, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology,* vol. 27, pp. 415-444, 2001.

[12] C. Schreiber*, et al.* (2004, 7 January 2011). *Construct - A Multi-agent Network Model for the Co-evolution of Agents and Socio-cultural Enviroments*. Available: http://www.casos.cs.cmu.edu/publications/papers/schreiber_2004_constructmultiagent.pdf

[13] D. Krackhardt, "Social Networks," in *Encyclopedia of group processes and intergroup relations*. vol. 2, S. Otten*, et al.*, Eds., Los Angeles: Sage, 2010, pp. 817-821.

[14] D. Garlan, *et al.* "Using Service-Oriented Architectures for Socio-Cultural Analysis," in *21st International Conference on Software Engineering and Knowledge Engineering* (SEKE2009), Boston, MA, 2009.

[15] B. Raman. (2001, 6 August). *Pakistan's Inter-Services Intelligence (ISI) (1/8/2001 ed.)*. Available: http://www.acsa2000.net/isi/index.html

[16] A. Ahmed. (2010, 4 January 2011). The 'Cold Start and Stop' strategy. *Insitute for Defence Studies and Analyses - Comment* [Electronic OpEd]. Available: http://www.idsa.in/idsacomments/TheColdStartandStopstrategy_aahmed_280910

[17] G. B. R. Kanwal. (2010, 4 January 2011). India's Cold Start Doctrine and Strategic Stability. *Insitute for Defence Studies and Analyses - Comment* [Electronic OpEd]. Available: http://www.idsa.in/node/5442/372

[18] H. V. Pant. (2010, 4 January 2011). India's quick-strike doctrine causes flutter. *The Japan Times* [Electronic Newspaper]. Available: http://search.japantimes.co.jp/cgi-bin/eo20100202a1.html

[19] S. S. Menon, "Speech by [Indian National Security Advisor] NSA Shri Shivshankar Menon at NDC on "The Role of Force in Strategic Affairs"," N. S. Council, Ed. New Delhi, India: Government of India, 2010.

Networks		Node Types								
		Agent	***Knowledge***	***Resource***	***Task***	***Event***	***Organization***	***Location***	***Role***	***Belief***
Node Types	***Agent***	Social *"Who knows who"*	Knowledge *"Who knows what"*	Capabilities *"Who has what"*	Assignment *"Who does what"*	Attendance *"Who attends what"*	Membership *"Who belongs to what org"*	Agent Location *"Who is where"*	Role *"Who has what roles"*	Belief *"Who believes what"*
	Knowledge		Information *"What informs what"*	Training *"What resources are needed for training"*	Knowledge Requirements *"What knowledge is task critical"*	Education *"What event teaches what"*	Organizational Knowledge *"What org knows what"*	Knowledge Location *"Where is what learned"*	Role Requirements *"What must be known to perform what"*	Knowledge Influence *"What knowledge informs what?"*
	Resource			Substitution *"What can replace what"*	Resource Requirements *"What tasks require what"*	Event Requirements *"What events require what"*	Organizational Capability *"What org can do what"*	Resource Location *"Where is what"*	Role Requirements *"Who needs what resource to do what "*	Resource Beliefs *"What beliefs are required to use what"*
	Task				Task Precedence *"What must happen before what"*	Event Agenda *"What tasks occur at what?"*	Organizational Assignment *"What org does what"*	Task Location *"Where is what done"*	Role Assignment *"What roles do what"*	Belief Requirements *"What beliefs require what tasks"*
	Event					Event Precedence *"What events happen before what"*	Organizational Responsibility *"What org is putting on what"*	Event Location *"Where is what event"*	Role-Event Requirements *"What roles are often present at what"*	Belief Attendance *"What beliefs influence participation at what"*
	Organization						Inter-Organization *"What org works with what"*	Organization Location *"Where is the organization"*	Organization Role *"What org has what roles"*	Organizational Culture *"What beliefs are common"*
	Location							Proximity *"What is near what"*	Location Roles *"What roles are common where"*	Significant Locations *"Where is associated with what beliefs"*
	Roles								Inter-Role *"Who knows what"*	Significant Roles *"What roles are associated with what beliefs"*
	Beliefs									Belief Influence *"What beliefs influence what"*

Appendix 1 Meta-Matrix table of nodes and 45 networks –precise semantics will depend on text corpus. Derived from [4, 6]