Advancing Future Network Science through Content Understanding

Paul Hyden, Stephen Russell
Information Management and Decision Architectures
Naval Research Laboratory
4555 Overlook Ave SW
Washington, DC 20375
UNITED STATES

{paul.hyden, stephen.russell}@nrl.navy.mil

David Jakubek
Office of the Assistant Secretary of Defense for Research and Engineering
4800 Mark Center Drive
Alexandria, VA 22350
UNITED STATES

david.a.jakubek.civ@mail.mil

Napoleon Paxton, Ira S. Moskowitz
Center for High Assurance Computer Systems
Naval Research Laboratory
4555 Overlook Ave SW
Washington, DC 20375
UNITED STATES

{napoleon.paxton, ira.moskowitz}@nrl.navy.mil

ABSTRACT

We summarize the future network relevant results of a Joint Interactive Content Understanding Forum held on 25 March 2014 at the Naval Research Laboratory, Washington DC, USA. We report to NATO STO on the aspects of future networking to content understanding analyzed at this forum via our paper. The workshop objectives were to introduce new perspectives, expose government research labs work in related areas, illustrate how content understanding applies to cyber security, and introduce operational needs in Information Management for algorithms focused on content understanding. Overall, the outcomes were greater awareness of the problem and opportunities, establishment of common needs, introduction of a working group, and documenting a path forward.

1.0 INTRODUCTION

One of the greatest achievements in the development of the Internet is that communication can happen without the transport layer needing to understand the meaning of the actual content being transported. The often-used analogy of shipping containers being moved across a highway or ocean nicely illustrates the technological beauty of cyber applications. However, despite this achievement, the increasing security demands placed on cyber-traffic have mandated a need for comprehension of the underlying meaning of the packets being moved.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Advancing Future Network Science through Content Understanding	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory ,Information Management and Decision Architectures,4555 Overlook Ave SW,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

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

13. SUPPLEMENTARY NOTES
Cognitive Radio and Future Networks, The Netherlands, 12-13 May 2014.

14. ABSTRACT
We summarize the future network relevant results of a Joint Interactive Content Understanding Forum held on 25 March 2014 at the Naval Research Laboratory, Washington DC, USA. We report to NATO STO on the aspects of future networking to content understanding analyzed at this forum via our paper. The workshop objectives were to introduce new perspectives, expose government research labs work in related areas, illustrate how content understanding applies to cyber security, and introduce operational needs in Information Management for algorithms focused on content understanding. Overall, the outcomes were greater awareness of the problem and opportunities, establishment of common needs, introduction of a working group, and documenting a path forward.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**17**	

Moreover, with advances in cognitive radio technologies, where the characteristics of the utilized transmission medium may be dynamic, the ability to characterize transport may have an increased dependency on content understanding in future networks. By example, entire business segments of Silicon Valley concentrate on monetizing the substance and nature of the content involved in interactions with their modern services and products. Increasingly, it is critical for the basic missions of government to bring together every tool and technique that allow knowledge workers to leverage information content and apply it to higher order decision making capabilities. The application of these algorithms and technology allows knowledge workers to communicate and operate with semantic fidelity across vastly different domains. In network-constrained environments where tactical decision-making is time sensitive, it is imperative to deliver prioritized, aggregated, and summarized information and visualizations that serve the needs of the end user. In many cases, a properly enriched understanding of this content will not only improve security, but also reduce demands on the network. It is insufficient to model topologies, physical properties, and usage patterns alone. Future networks must take content understanding into account.

We summarize the results of the 4th Joint Interactive Content Understanding Forum held on 25 March 2014 at the Naval Research Laboratory, Washington DC, USA. This forum was focused on cyber security as a content understanding challenge. In addition to the US Naval Research Laboratory, there was participation from the Institute for Defense Analyses, UT-Austin Applied Research Lab, the Office of the Secretary of Defense, the Office of Naval Research, the US Army Research Laboratory, MITRE, MIT Lincoln Laboratory, and the Central Intelligence Agency, and others from the academic and research communities.

The purpose of the forum was to gather cross-governmental agencies and law enforcement together with government researchers to raise awareness of the interdisciplinary nature of content understanding and cyber security. Towards this vision, the forum served to increase the awareness that the two domains shared a boundary at a critical juncture: decision-making. Moreover compartmentalized research and technical solutions ultimately leads to inhibited capability and potentially wasted resources. Both content understanding and cyber security are difficult problems and rather than relying on only a few major contributors to solutions, it is important to share processes, approaches and success stories as everyone does their best to stretch a finite amount of human attention and resources. Real world, user-driven problems appropriately matched to research efforts can enable true information dominance and organizational advantage. Subsequently part of the vision for the forum is to inform leadership of the need to commit resources to these significant challenges and connect researchers with end users to allow for exploitation of past research successes and permit more work to continue.

2.0 FUTURE NETWORKS CHALLENGED BY CONTENT UNDERSTANDING NEEDS

The forum consisted of several panels that highlighted how research in content understanding motivates capabilities necessary in future networks. Historical policy and standards have mandated the operation of networks, most significantly the Internet follow a law of inter communication layer isolation and to that end there are many benefit of layer isolation and independence. However, it is a naïve view of future networks where the concern of content is ignored or unexploited. The may be little debate that future networks will feature a high degree of dynamism, automation, and context-aware intelligence. In autonomous or hybrid manual-autonomous networks, decision-making relies on the cooperation of the elements (hardware/software) and the information/content that participate in the network. However cooperation or synergy between the physical and contextual is not guaranteed. Further, even when cooperation is possible, constraints and objectives may affect the degree of cooperation and agreement. Policy or mission objectives may dictate that two components collaborate, but situational and content artifacts may inhibit the achievement of that collaboration. Such is the domain of cognitive radios that seek to exploit spectrum resources, which form the highway for

content and information movement. Cognitive Radios must be able to demonstrate usage with minimal harm to the "primary user" and from a pragmatic perspective this task is rendered difficult due to challenges in sensing the spectrum in a reliable manner [1]. In this context it is easy to see that content can affect or enrich and transport-related decisions.

Much of the research in cognitive radios and software defined networks is focused on enhancements at the lower layers of the OSI seven-layer model, but inter-layer information sharing can provide critical content-related insight to these enhancements. Moreover there is ample research that illustrates that inter-layer information and specifically content and context awareness can improve human decision-making [2,3]. Thus it follows that autonomous decision-making can be enriched with the same information. Industry views the complexity of introducing content understanding to cyber challenges similarly, as is evident when considering the issues of Net Neutrality. Net Neutrality takes content at its core in making right-of-way and resource allocation decisions within the networks [4]. Further, it may be argued that the net neutrality network environment is adversarial in terms of competition. It is already a fact future networks will include content-driven considerations, because even though Net Neutrality as a moniker has faded from many political discussions, content-based automated communications continues to be debated and advanced [5]. Cast in the light of future networks, content-agnostic and neutral networks may be consumed by the evolution of commercial technology and service provisioning.

Also when considering the implications of cyber security and surveillance the interdependence on content understanding in cyber is readily apparent. Many means of implementing effective cyber intelligence and surveillance mandate the functions implemented at the lower layers of the network, below transport (like the mechanisms of CRs and SDNs), thus controlling/manipulating content, both benign and malicious content, depending on your perspective. The transition to merge content understanding and cyber security is already underway, yet innovations in the individual research domains still remain relatively discrete. If future networks are increasingly autonomous and dynamic (e.g. CRs and SDNs) then content understanding is essential.

3.0 CYBER SECURITY AS CONTENT UNDERSTANDING CHALLENGE

Though cyber security has all of the characteristics of a content understanding challenge, i.e. extracting comprehension from corpora, cyber is seldom viewed from the same perspective as content understanding. The problems are similar in that meaning needs to be automatically extracted from digital content to aid human decision-making. Although cyber security should be a domain rich with integrated capabilities from the content understanding domain, the gaps between those communities remain. Similarly, many of the techniques developed for cyber security could advance the science in the content understanding research. The final panel for the day presented a case and examples of how cyber security research is a content understanding problem and innovations from both domains converge.

Dr. Stephen Russell (Naval Research Laboratory) provided an overview of how cyber problems are really content-understanding problems and vice versus. To illustrate, Russell presented a simple information retrieval exercise using Google to search for the terms cyber warfare, which returned 8.9 million hits and content understanding, which returned 735 million hits. When both phrases were searched together only 136 hits were returned. Russell aptly pointed out that this simple exercise points to the separation of the two domains even though there are many similarities in the problem spaces. The broader point was made suggesting that cyber is just information in motion and content understanding, while grounded in its information retrieval roots, now mandates temporal considerations (movement through time) to have greater efficacy for decision making purposes. To further illustrate, Russell used a metaphor of gaining situational awareness in the context of cars parked at a sports event stadium. Likening this to content understanding, Russell noted it was still a difficult

problem. Extending the metaphor, Russell suggested that cyber was like the same number of cars moving on the highway; an even more difficult problem to gain situational awareness. The real issue in both problems is gaining context and intent and the metaphor shows how solutions in either research area can advance the other.

Towards shared characteristics and research overlap, Russell pointed out that: both content understanding and cyber are large data problems; cyber threats find their opportunity when data meets logic (simply: application/logic abuse using content) where the network is used as a means of transport; content understanding has theoretical grounding in information theory; cyber transactions represent an orthogonal perspective on content through metadata exploitation (the same concepts exist in traditional content understanding, e.g. file access times/locations, user accesses, folder hierarchies, etc.). Russell emphasized that the research innovations in either domain, cyber or content understanding, could and should be applied across both problems spaces. To address these shared characteristics Russell introduced the panel members, Dr. Cheryl Martin (AR Laboratories Texas Austin), Dr. Ira S. Moskowitz (Naval Research Laboratory), Dr. Albert Reuther (MIT Lincoln Laboratory), and Mr. Joseph Mathews (Naval Research Laboratory).

Dr. Cheryl Martin (ARL, University of Texas Austin) introduced and presented some of the work conducted by the Center for Content Understanding (CCU), where she is the Director. This center merged with the Cyber Information Assurance and Decision Support (CIADS) group to address the issue raised by Russell in the session introduction. Formed to conduct research in that specific gap, the center investigates and develops academic and theoretical models, as well as systems to advance the science. Martin presented a model developed by the CCU that illustrated how the big data problem of information management and information exploitation overlaps with content understanding capabilities. The model defines the necessary capabilities to achieve information management and exploitation workflows, such as alerting, tracking, routing, classification and identification. Martin highlighted a system developed by the CCU implementing that model that bring together large sets of diverse information to alert, prioritize, predict, summarize, correlate, group, relate, and visualize the “big picture” in a cyber security context. Martin also discussed how “records” are conceptually the same in both the content understanding and cyber domains. She went on to illustrate how the decision-oriented activities relative to the records are also the same: “for each individual record… categorize, sort, translate, interpret, explain, prioritize, etc.”

Martin used an example to illustrate the concepts from the CCU model. Consider threat reports – a report about some computer (IP address, web http URL) has been a known source of malware so we should avoid it. Ideally what network operators/administrators would like to do is automatically put rules into their management systems to automatically block the threat, but the rules must be dynamic and understand shifts in that computer context (perhaps it was cleaned, no longer a threat, a honeypot etc.). Martin emphasized that we have to find a way to understand the context of the malware and other informational content so we can block the appropriate types of data and dynamically adapt policy to be sensitive to shifts in content. Consider such a content-aware system used in the management of automatic alerting, which is a know challenge for cyber network operators. Specifically Martin suggested that content understanding approaches can be used for more than looking at the content of individual network packets. Content understanding can help to identify the utility and context of the data. Protections can be prioritized based on the content and downstream alerting and decision-making can be enriched and enhanced. Martin concluded with a discussion of how content understanding models and algorithms developed for use in cyber can be applied to content understanding in other domains.

A concept to which Martin alluded was that the benefits of content understanding approaches, when applied to any domain and specifically cyber, find their grounding in information theory. Many computer applications and concepts are considered engineering and not a science from an academic perspective. The second panelist, Dr. Ira S. Moskowitz, a Mathematician with the Naval Research Laboratory, specializes in problems in information

theory and has current projects in applied network science.

Many computer applications and concepts are considered engineering, and not a science from an academic perspective. Thus, some important cyber security issues are often approached in an ad hoc manner. Not properly utilizing a scientific approach can lead to erroneous assumptions, as pointed out in the Moskowitz-Kang Small Message Criterion [6] from covert channel analysis. The overlap of content understanding and communications also has grounding in information theory. The similarities between research in covert channel analysis and cognitive radio can be clearly seen in [7], [8], and [9]. Moskowitz went on to describe the history of false flags (where ships flew the flags of their adversaries until the last moment to disguise their true nature) as a practical example of misunderstood content. Using this example as an introduction to blindly applying mathematics to the cyber security problem space, Moskowitz highlighted how the same information theoretic approaches used in steganography can be immediately applied to both cyber and content understanding, yet little work has been done in these areas. Moskowitz went on to discuss how mathematical properties and algorithms from physics can provide the theoretical grounding for techniques to be applied across domains, reinforcing the point that context must still be understood. He illustrated this point with a few examples from his current work in network science that focuses on networks' topological structure and information propagation.

In summary, Moskowitz illustrated the underlying robustness that theoretical grounding can provide to computational techniques that apply to both cyber and content understanding. The major paradigm of cognitive radios is that the changing, or switching, is done behind the curtain, leaving the audience only to see and sense the application. Such a paradigm is ripe for the covert channel analysis that Moskowitz discussed. Do we want the end-user to be aware of the switching? Are we concerned that information is leakable to the public, or to a proprietary user, or to even others at a higher level of sensitivity? The research done in covert and steganographic communication is directly related to this. Imagine a cognitive radio switching between CDMA and 3G, the timing of these switches, if they are visible to the application layer user can pass information. The tools for analysing these type issues are well developed in the cyber community and should be adapted. But, one must also keep in mind that need not be concerned with every possible application. An understanding of the content is imperative for high assurance use of cognitive radios.

The next panelist, Dr. Albert Reuther (MIT Lincoln Laboratory) presents a cyber analytics architecture and system that applies theoretical information and graph theory to deliver content fusions and analytics in the cyber domain. From this presentation it was clear that research conducted at Lincoln Laboratory has taken the approach that content understanding is not only relevant but also necessary for enhanced cyber security. As part of his discussion, Reuther presented a cluster-based architecture that delivered multi-sensor and content information fusion, coupled with large data analytics portrayed using innovative visualizations to aid cyber decision makers. The architecture uses modular "sidecars" that ingest data typical of the content understanding domain (e.g. PDFs, Documents, Powerpoint, etc.), cyber-centric content and sensor data (server logs, traffic capture, device configurations, etc.), and external relational enterprise systems. The system presented by Reuther represents a marked convergence of content understanding innovations converging on the cyber security problem and the MIT Lincoln Laboratory (MIT LL) system implements the technology throughout the entire cyber security workflow, from sensor and content collection through to visualization and presentation. At its core, the MIT LL system sits on a multi-core distributed computing system that applies ubiquitous file system abstractions that are a necessary hallmark of large content understanding systems. This system called LLGrid is a diverse mix of industry-standard storage technologies (Lustre, BitTorrent, PostgreSQL, MySQL, and GRSecurity) and emerging technologies (HadoopDFS, Tokutera, Sector/Sphere, HBase, and other BigTable-like databases). Likewise, the LLGrid storage interface uses industry standard interfaces (NFS, Samba, FUSE, and WebDAV) and high-level interfaces that MIT LL has developed (MatlabMPI, pMatlab, pMatabXVM, gridMatlab, D4M, LLGridZF) [10]. The list of content analytical technologies described in the system alone,

illustrates how the system treats both traditional lower layer network sensor data the in the same context as application layer content. MIT LL has tested the hardware architecture in analysis of botnet networks, which requires integration of both network sensor data (e.g. OSI layer 1-4) and application layer data (e.g. malware documents).

A significant element in the development of the MIT LL system was to drive toward real time analytics, again a concern typical of content understanding problems. This problem is not trivial in content understanding due to the sheer size of the data as discussed by Martin earlier in the session and others [11, 12]. Analytics in future networks is already a "big data" problem. Reuther highlighted how in this environment real-time, distributed, high performance computing is necessary and illustrated how the MIT LL system allows the map/reduce parallel programming model to be used quickly and efficiently in any language on any compute cluster and analyze virtually any content when using their D4M (Dynamic Distributed Dimensional Data Model) [13]. At the core of Reuther's example was the point that future networks and thus cyber security will only see an increase in spectrum and network use for medical devices and applications, smart grid applications, and other varied content sources. This point is reinforced in the literature by others who go on to demonstrate that cognitive radios, software defined networks, and other cyber-virtualizations will have to not only support but also ***understand*** these upper layer applications [14,15].

In closing his presentation, Reuther highlighted and demonstrated how the underlying analytics in the MIT LL system were delivered to analysts and decision makers using innovative visualization. This point highlights the importance of converging the cyber domain and content understanding: understanding is irrelevant without decision making and understanding has to transcend machine algorithms that will be a commodity component of future networks and impact human-decision making [16-18]. While the mechanics of future networks will largely be autonomous, humans will remain in the loop, particularly in adversarial environments and despite any system-level cyber autonomy there will remain a need for decision-making, which cannot be content agnostic.

Mr. Joseph Mathews (Naval Research Laboratory), the final panelist, presented research that summarized the conceptual focus of the session. In discussing visualizing cyber content, Mathews showed how content understanding approaches could be coupled with cyber networking algorithms. As a researcher who works closely with network operators Mathews re-affirmed that many of the cyber tools target specific function within the network and often lack a global scope. This problem leads to the research question of how to analyze a network the size of the Internet. It is impossible for one presentation to display everything. Mathews pointed out the amount of investment in solving this problem that has been and is currently being made, yet there is still no total solution. In this problem the content is varied and tremendous and despite the plethora of available cyber-related tools, we lack situational awareness at a global level. Moreover decision-making in inhibited by exiting tools inability to deliver understanding at this scoping. Mathews documented how many of the current visualization approaches are either too complex, inadequately handle the volume of data, or lack focus on decision making. As a result network operators and administrators are forced to use simpler tools and limit the scope of the problem. To illustrate a solution that appropriately merges cyber algorithms and sensing with content understand approaches, Mathews presents findings from his project conveying cyber situational awareness on a map using "map metaphors," a map based on synthetic dimensions rather than real geography. As a first step in this research, Mathews discussed the initial research where he took a node and link graph made of authors (nodes) and papers they collaborated on (edges). To this he applied the size of the link between two nodes is the inverse logarithm of how many papers they collaborated on in the context of a specific conference (a content understanding technique). Authors that are far apart collaborated on a few papers and authors that are close collaborated on many papers. The resulting patterns showed a group with few outliers. By applying a synthetic map to the graph it was possible to not only visualize the results, but also make the problem tractable for decision-making. Mathews went on to illustrate how the same approach could be applied to the Internet,

displaying results that clearly illustrated the point. Moreover, Mathews showed how geospatial information layers (another content understanding methodology) could easily be applied to the synthetic map. In this example, Mathews was able to show a technique that included threat alarming, and situational awareness on a global scale across a variety of network traffic and topological concerns.

The session was concluded with a short question and answer session that focused on the need for more approaches such as the ones shown that converge content understanding with algorithmic and sensor driven cyber enhancements. Moreover, solutions developed for cyber can and should be readily applied to the content understanding domain. The real limitation between the two domains lay in segregation of the community and open discussions where researchers from both domains can interact will enhance and enrich the advancement in both.

4.0 GOVERNMENT RESEARCH LAB PERSPECTIVE

The government research lab panel provided a Department of Defense perspective. An increased research effort is now applied to detect and defeat threat networks using structure and unstructured data from sensors and open sources of information. So in the Department of Defense case, the context to optimize the performance of Future Network will require better insight into the threat network. Content understanding is necessary across all decision making in the Department of Defense to include both the warfighting domain and the business domain. As examples across the Department of Defense warfighting domain, content understanding is needed: (1) to improve intelligence, surveillance and reconnaissance support to small unit operations at the tactical level of command, (2) to create new indications and warning to counter regional threats at the operational level of command, and (3) to foster strategic cooperation and partnership at the strategic level of command. The Department of Defense research programs, Data to Decisions and Human Social Cultural Behavior, are Defense-wide programs established to address warfighting mission needs across all three levels of command: tactical, operational and strategic. The panel offers a broad perspective being comprised of a government lab, Army Research Lab, a Federally Funded Research and Development Corporation, Massachusetts Institute of Technology Lincoln Lab, and an industry research organization, Lockheed Martin Advanced Technology Lab.

Dr. Liz Bowman from the Army Research Lab offered new perspectives on content understanding which introduces situational awareness of threat networks earlier in the military decision making process. In some cases military decision makers lack valuable insight until it is too late in the process. Using all of the available information to identify patterns and establish an understanding of content can facilitate better decisions earlier in the process. Historically, Department of Defense research investments improve sensing to detect, assess, and deliver data to support decisions, but recently more emphasis has been placed on using open sources of information. Additional insight can be achieved by detecting trends using social models, utilizing open source data available on social media sites to discover conflicts and identify trends. In order to become more agile in future networks, this data must be accessed in a streamlined fashion from social media so that this can be achieved dynamically. The goal is to make predictions and discover influential nodes in the networks. Another critical aspect is effective visualizations based on the analytical results to achieve effective decisions faster.

Research efforts offered by Dr. Bowman that support improved content understanding include the Data to Decision Text Analytic program, collaborative visual analytics, and social cultural awareness. The Data to Decision Text Analytic program blends structured and unstructured data to be analyzed utilizing observational data, a priori knowledge models, and inductive knowledge models to achieve contextual understanding. Improved content understanding is achieved through human and computer processing techniques that leverage human cognitive abilities. Looking at community structure and clusters in social network, semantic analysis can

determine relationships. The next research effort to support content understanding is visual analytics, particularly collaborative visual analytics, which examines how distributed analysts and decision makers work together effectively using analytical tools and visualizations to achieve knowledge transfer and distributed situational awareness. The overall goal is improved presentation. That is, visual representation of analysis results based on automating human analysis processes while integrating agents in the system that perform an ontological search of text, developing a way to share data amongst agents, and creating dashboards to display the visual analysis. The last effort supporting content understanding focuses on prioritizing the information from the social cultural space to see what can be developed technically. This is a multi-theoretical approach, with human-centered analysis and computations for visual analytics, network analysis, the development of cultural structuralism narratives, and naturalistic decision-making. Cultural artifacts are imbedded in a societies organization and structure and people use structure to transmit their cultural behaviors so you can study how inequality in resources the use of power and the outbreak of conflict can be seen as a part of culture analysis. Examining the structure from the data to define function occurs in many areas of computational and network science.

Current prototypes and demonstration efforts used successfully:

- A tool using weak signal analysis on open source data. In support of the United States Army Africa HQ, reports suggested that the Boko Haram was losing their influence in Nigeria. People familiar with the situation did not believe the report based upon anecdotal evidence. The tool developed though a small business innovative research project used open source data, sentiment analysis, and social network analysis to show that Nigerian conflicts spiked in areas coinciding with the re-emergence of Boko Haram in a different location than previously observed.
- Models of Sentiment and Sentiment Transmission. Provides the ability to assess the changing attitudes of a population by exploiting traditionally underutilized yet readily available online sources such as news, blogs, and social media. A dashboard provides several ways to visualize the analytical result.
- Multi-Source Network Pulse Analyzer and Correlator provides course of action planning by enhancing the understanding of the complex dynamics of tightly connected elements spanning demographic, political, military, economic, social, information and infrastructure dimensions, a portal for quickly assessing and understanding the dynamics within a social network, to include theme dynamics, and trending log entries on topics overlaid onto real world incidents.
- Social Network Analysis Realization and Exploitation automatically builds high-fidelity social networks from text data sets too large to be scrutinized in detail through manual effort and provides visualizations of the complex social network interactions, providing the Warfighter with enhanced situational awareness.

Dr. Sanjeev Mohindra from the Intelligence, Surveillance, and Reconnaissance (ISR) Systems Division at MIT Lincoln Laboratory presented his efforts within the Data to Decision program to establish a testbed infrastructure to integrate advanced analytics to support intelligence analysis. One challenge is to combine multiple sources of data that may be considered unrelated. When also combined with human intuition, this capability is used to find people, places, objects and associated activities. Along with traditional sources of sensor intelligence, open source information is added to support content understanding. A key point is no single source of data is good enough to tell the complete story. When building analytics, data sources can be very large so scalability is critical. Scalability includes storage as well as computing resources. The analytical capability is achieved when these resources are made available and exposed to analytical services to achieve a prescribed workflow. Another important lesson learned from the testbed development is that no single organization can capture all the data that is needed. This requires a standard system of interfaces to enable better data sharing and innovation. The interactive nature of the discovery aspect of data analysis relies on human interaction. Data analytics should help

the analysts understand the data and discover patterns. The Data to Decision testbed has the unique aspect of combining high performance computing, cloud computing and a service oriented architecture in one system. Analytics are provided as services that can be composed into workflows to address a specific challenge or analytic need

Another important aspect of the Data to Decision testbed is the assessment of data analytic methods to achieve an operational outcome. Since no single tool addresses all of the data analytic needs, an approach to compose analytics from various sources into a required workflow is used. This provides operational flexibility. In order to understand analysts' challenges and also measure performance, a Red Blue Team experimentation process is used. This provides the analysts perspective and involvement. It allows researchers to conduct independent and objective observations leading to recommendations for automation and innovation. To improve analysis and decision support the person involved needs to understand what the data is about. The analytics must help the analyst understand the data, summarizing text data and videos, find anomalies and clusters in the data - what stands out, extracting entities and events and give the analysts a chance to explore the data. In summary, for content understanding, as it relates to data analysis, will always be open ended, meaning that once insight is gained, more questions will be likely to follow.

Dr. Daniel McFarlane is Principal Research Engineer and LM Fellow Emeritus in the Informatics Lab at Lockheed Martin Advanced Technology Laboratories. He emphasized the need to achieve content understanding through context modelling. Context Modelling requires that meaning is delivered in terms of the receiver and requires the ability to adapt to unforeseen change and needs. The statistical model is used for experiments together with the individualized meta-cognitive model. Two key observations follow from field study. First, meaning is relative to the changing situation and the people involved in it. Second, adversaries are creative and always introducing things we do not expect. Additionally he found that people are simply better at cognitive reasoning, as they have a better way of organizing their thinking to solve problems. Humans see something critical in the situation and not just basic information. Machines are not at the point where they can replace humans in cognitive reasoning. However, expert information can be put into the system dynamically: patterns, stored perspective models, stored analytic models. Three types of information delivery define the space with some overlap.

- Autonomy: Intelligence-based systems that can respond to dynamic environment
- Automation: System performs as it was programmed with little operator involvement
- Human augmentation: mitigate cognitive limits and improve human performance

Indeed, technology can predict when people fail. Here, Dr. McFarlane introduces a key warfighter insight. Autonomy augmentation as a fourth type of information delivery engages warfighters' uniquely-human insight to mitigate system limitations. Since autonomy has strengths and weaknesses, adding in a human performer to see a situation after data has been reduced enough for a decision offers a novel approach.

Jim Starz from the Lockheed Martin Advanced Technologies Laboratory and presented his research work that transitioned from the Human Social Culture Behavior (HSCB) Program. The world today is much more dynamic and it is easy to find like-minded people to share ideas. Today ideas can be shared instantly so there is an increasing need to understand culture and society to determine how we should interact with different groups. Four major thrusts defined the HSBC program.

- Understanding - can you understand what society's thinking and what the behavior is?
- Detect shifts - can you see when society is shifting?

- Detect forecasting - when is there a change in behavior?
- Mitigation - can you shape policies when you detect things and maybe get populations to act in a different way?

The World-wide-Integrated Crisis Early Warning System (W-ICEWS) at Lockheed Martin is the program that has instantiated these efforts.

Instability forecasting, the identification of computational social science models for forecasting destabilizing events of interests, has achieved 80 to 85 percent accuracy for labelling correct events, actors, etc. However, it is not very accurate when moving to different domains. It is now a program of record and available as ISPAN on SIPRNet. The test and evaluation version is on unclassified networks. In actual use, it was discovered that most users didn't simply want to interact with the forecast but were interested in the products that went into building the forecasts. As such, ISPAN breaks out into three major components.

- iTRACE - Discovering Events and Trends (first component of W-ICEWS). Looks at historical events, trends, and news data, using event extraction to find who is doing what to whom and when and where. Events are classified as positive or negative events. Groups are characterized for their level cooperation. People use the data to discover their own analytics. Initial accuracy was about 60%, but utilizing clients in the loop early prompted the engineers to go back and improve the accuracy of the detection. Current accuracy is about 80%. The engineers report it as a good example of using clients to help the engineering process. People use the data more than the visualizations
- iCAST - Forecasting destabilizing events of interest across 167 countries worldwide.
- iSENT - Sentiment and Social Media. Measures the positive and negative sentiment across multiple types of social media.

In summary, the government research panel provides examples of Defense-wide research in content understanding to defeat threat networks. This research broadens the type of information required to develop context to achieve mission effects against threat networks. The use of "open" sources of data increases the volume of data delivered by future networks and the inclusion of streaming data sources stresses the velocity of data supported by future networks. The dynamic nature of the military operational environment to include rapidly changing world events and evolving partnerships also place demands on the adaptability and agility of future networks.

5.0 CONCLUSUION

Content extends beyond the topological, physical, and mechanical nature of networks, encompassing the substance and use of cyber as well as cognitive communications networks. Thus, efficient content understanding requires a collection of people, technology, and algorithms to allow organizations to address ever-larger problem domains that form the reality of network applications. Content understanding related challenges are not problems suitable for closed-loop solutions, but rely on intuitive and experienced practitioners who face domain-specific problems every day and researchers who can quickly adjust to unforeseeable changes in context, data formats, or real-time needs that are in constant flux and require ever more sophisticated levels of engagement. The capacity for human understanding of large corpora of documents, images, and other unstructured data has become a key feature and potential competitive advantage of large systems of people and networked information technology. This common challenge appears in numerous problem spaces across the government and is not limited to network sensing.

6.0 REFERENCES

[1] Mishra, S.M.; Sahai, A.; Brodersen, R.W., "Cooperative Sensing among Cognitive Radios," *Communications, 2006. ICC '06. IEEE International Conference on* , vol.4, no., pp.1658,1663, June 2006

[2] Russell, S., Forgionne, G., and Yoon, V. (2009), "Presence and Availability Awareness for Decision Support Systems in Pervasive Computing Environments" *International Journal of Decision Support System Technology*, 1(1)

[3] Blasch, E., Russell, S., Seetharaman, G. (2011), Joint Data Management for MOVINT: Data-To-Decision Making", *14th International Conference on Information Fusion (Fusion 2011)*, Chicago, Il. USA

[4] Crowcroft, J. 2007. Net neutrality: the technical side of the debate: a white paper. *SIGCOMM Computer Communication Review*. 37, 1 (January 2007), 49-56.

[5] Hiltzik, M. (2014, January 14). Net neutrality is dead. Bow to Comcast and Verizon, your overlords. *The Los Angeles Times*. Retrieved from http://www.latimes.com/business/hiltzik/la-fi-mh-net-neutrality-20140114,0,522106.story

[6] Moskowitz, I.S.; Kang, M.H., "Covert Channels --- Here to Stay?," Proc. Computer Assurance, COMPASS 1994, Gaithersburg, MD, July 1994.

[7] Moskowitz, I.S.; Greenwald, S.J.; Kang, M.H.,"An Analysis of the Timed Z-Channel," IEEE Transactions on Information Theory, V. 44, No. 7, pp. 3162-3168, November 1998.

[8] Moskowitz, I.S.; Newman, R. "Multiple Access Covert Channels," IASTED Conference on Network and Information Security (CNIS 2005), Nov. 14-16, 2005, Phoenix, AZ, pp. 182-188

[9] Goldsmith, A.; Jafar, S.A.; Maric, I.; Srinivasa, S. "Breaking Spectrum Gridlock With Cognitive Radios: An Information Theoretic Perspective," Goldsmith, A.; Jafar, S.A.; Maric, I.; Srinivasa, S., Proceedings of the IEEE, V. 97, No. 5, pp. 894-914, May 2009.

[10] Kepner, J., Byun, C., Arcand, W., Bergeron, W., Hubbell, M., McCabe, A., Michaleas, P., & Reuther, A., (2011) Persistent Surveillance Supercomputing using the LLGrid Filesystem, Proceedings of The High Performance Computing Modernization Program Users Group Conference 2011

[11] Singh, S.; Singh, N., "Big Data analytics," Communication, Information & Computing Technology (ICCICT), 2012 International Conference on , vol., no., pp.1,4, 19-20 Oct. 2012

[12] Jeffrey Cohen, Brian Dolan, Mark Dunlap, Joseph M. Hellerstein, and Caleb Welton. 2009. MAD skills: new analysis practices for big data. Proc. VLDB Endow. 2, 2 (August 2009), 1481-1492.

[13] Chansup Byun; Arcand, W.; Bestor, D.; Bergeron, B.; Hubbell, M.; Kepner, J.; McCabe, A.; Michaleas, P.; Mullen, J.; O'Gwynn, D.; Prout, A.; Reuther, A.; Rosa, A.; Yee, C., "Driving big data with big compute," High Performance Extreme Computing (HPEC), 2012 IEEE Conference on , vol., no., pp.1,6, 10-12 Sept. 2012

[14] Jianfeng Wang; Ghosh, M.; Challapali, K., "Emerging cognitive radio applications: A survey,"

Communications Magazine, IEEE , vol.49, no.3, pp.74,81, March 2011

[15] N.M. Mosharaf Kabir Chowdhury, Raouf Boutaba, A survey of network virtualization, Computer Networks, Volume 54, Issue 5, 8 April 2010, Pages 862-876, ISSN 1389-1286

[16] He, A.; Kyung Kyoon Bae; Newman, T.R.; Gaeddert, J.; Kyouwoong Kim; Menon, R.; Morales-Tirado, L.; Neel, J.J.; Youping Zhao; Reed, J.H.; Tranter, W.H., "A Survey of Artificial Intelligence for Cognitive Radios," *IEEE Transactions on Vehicular Technology*, vol.59, no.4, pp.1578,1592, May 2010

[17] Castells, M. (2011), The Rise of the Network Society: The Information Age: Economy, Society, and Culture, Volume 1, John Wiley & Sons, NY, NY, USA.

[18] Vermesan, O. and Friess, P. (2011), Internet of Things - Global Technological and Societal Trends From Smart Environments and Spaces to Green ICT, River Publishers, Aalborg, Denmark.

ANNEX A

For reference, we include discussion from an additional panel and two speakers from the Joint Interactive Content Understanding Forum that were judged to be out of scope relative to the focus on future networks in this paper.

Keynote Speaker

The Keynote Speaker was Paul A. Brinkley, the President and Chief Executive Officer of North America Western Asia Holdings (NAWAH), LLC., and author of the newly released book, “War Front to Store Front.” Prior to the launch of NAWAH, Brinkley served as Deputy Under Secretary of Defense and Director of the Task Force for Business and Stability Operations, reporting directly to U.S. Secretary of Defense Robert Gates. As part of these duties, he spent five years overseeing economic improvement in Iraq and Afghanistan, along with several African countries. Mr. Brinkley’s professional experience has given him the leadership perspective on the value of enriched content understanding.

He discussed economic development in war torn countries to describe the overall problem of content understanding across multiple agencies and competing semantic frameworks. Our system of information sharing is failing because we have an unrealistic operating environment. Our organizations are getting larger and have less of a centralized information sharing structure. We have a transitional problem where we over rely on service providers. For one transaction, many companies are involved in delivering the final content and if one of them fails the service is delayed and expectations are not met. Incentive structures may be able to motivate companies to share information better. He discussed the Marshall Plan and how it "helped" Europe recover after World War II by giving them business incentives. Notably, smaller, specialized groups were able to tackle individual problems better because they are the experts in those areas. He pointed to the U.S. Agency for International Development as an example of a large organization being tasked to do things that it does not specialize in, and he believes it has become less effective because of this. He discussed how the approach of a larger military and less business incentives have not improved the economy in Iraq and Afghanistan, and people have to rebuild things themselves with proper incentives and incentives rebuild organizations. Incentives are the key to prompt businesses and government organizations to share information to gain understanding of the content. As we’ve seen in numerous industries, the flow of contextualized, content rich information allows local actors to optimize their behaviour toward the broader goals of the enterprise.

New Perspectives on Content Understanding Panel

This panel was convened to stretch the conversation and introduce notions of content understanding from domains that hadn’t been introduced at prior meetings. We see a common theme emerging of the central need to understand the meaning of the component flows on network, be they are bits, pings, dollars, likes, retweets, or intellectual property.

Jeff Johnson is Executive Director within Ernst and Young’s Technology Risk and Security Center of Excellence. He serves as the Coordinating Partner for EY's Cyber Economic & Business Risk Solution efforts. The Technology Risk and Security COE is a key practice within the Global EY Performance Technology Advisory Service Line - a service line with over 9,000 seasoned practitioner-consultants in the Americas and more than 24,000 world-wide. Among many other accomplishments, Johnson contributed to Brinkley’s efforts in the Task Force for Business and Stability Operations (TFBSO) from 2007 to 2009.

He has been looking at the challenge of relating the business and financial meaning of cyberattacks in the context of global competition and financial markets. He comments that life and the challenges we face shapes the way we look at data. When we work in the financial world we view content through a financial lens. When we work in the technical world we view content through a technical lens. Security people look at technical data but very little is done to determine why a breach takes place. Normally they discover the intrusion and patch up the system and the investigation ends there. Clearly, the need to be proactive in protecting the enterprise starts with understanding the larger objectives of these attacks.

There is a problem with Nation States. Why are they hanging out in networks where they haven't been invited? Introducing the concept of Conversion Analysis helps to answer this question. Conversion Analysis is the process of examining asset losses and linking them to where those assets reappear, helping to provide context and meaning to otherwise disconnected phenomena. It appears there are only a few of reasons for Nation State Involvement:

- Harm the company
- Obtain the R&D to compete with the company
- Combination of the two goals

Johnson developed a cyber economic analysis program. They have teams that split to discover why hacks take place, such as the Night Dragon Campaign. One team looks at financial data. One team looks at technical data. Results showed a strong correlation between profit data and Nation State Attacks. For example, the retailer Target Corporation had a decrease in sales of 30% after attacks took place. It is possible that a nation state could be behind the attack to drive the value of the company down so it can purchase it at a lower price. Johnson observes that someone could have made a lot more money in shorting Target's sales in the market than they did with fraud on the credit card numbers. The glaring need is to merge the expertise of people in different lenses so we can effectively solve problems.

Dr. Philippe Loustaunau has over twenty five years of progressive leadership, business development, strategic planning, managerial, scientific, analytic, and teaching experience, with significant expertise and experience in Human Social Cultural Behavior (HSCB) modeling and deep technical expertise that bridges the social science community with the technical community.

Loustaunau offers that we need to take unstructured data and make it structured for better sharing and understanding. Emphasis on the data is required: Traditional text (newspaper, websites, blogs (social media), Social media across multiple sites for information about the users and their activities and images and videos, training classifiers to detect violence in image and videos. Current Intelligence Advanced Research Projects Activity (IARPA) research on data types includes time series activities which can forecast things such as smiles on Tumblr to detect violence in certain parts of the world. Weather data is very important to understanding health information all over the world. Opentable time series analysis tracks the vitality of neighborhoods. Parking lot analysis, which involves counting the number of cars in parking lots, can help determine health issues over time.

Emphasizing the understanding requires us to attaching meaning to social media. Features, text frequency of the same words over time, topic classification over time, and dynamic algorithms can tell us what we should be looking at over time and show how the topology changes over time. Coding and extracting events transcends feature extraction. Extracting a piece of scientific knowledge allows for inference of new knowledge from the extracted knowledge. Checking the modality of the event is more complicated algorithmically. Other information can be added but it makes the algorithms more difficult to code. The integrity of different types of

data, such as traditional text, weather data, patent data, etc. must be valued. This data can be combined to discover patterns and interesting facts. There is a large wealth of data publicly available on the Internet that can be leveraged to discover trends. Forecasting tools based on open source data are continuing to evolve. For example, the science and technology environment can be characterized by examination of publically available patent data from the United States and China. Discoveries and innovations can be tracked over time and between countries. Out of the data, the understanding of features must be extracted through topic classification and dynamic algorithms. Social network analysis uses the same understanding but also adds the network so we can understand the context of the data: who is attached to the data, where are they, what do they like. Things can be discovered based on the topological structure of the network as well. A problem with many of these tools is the systems do not translate to other domains well. Event coding must change for new contexts. Modality is difficult and has not been solved.

IARPA's interest is to leverage this information to do forecasting. There is a big push to do forecasting on events. Two programs being advanced in this area include Open Source Indicators (OSI) and Foresight and Understanding from Scientific Exposition (FUSE). OSI attempts to predict protest events in Latin America, and has a prototype for the Middle East region coming soon. The estimated accuracy is 80% with 60-70 % precision. It is done at the city level, which is made more difficult due to geographical location issues. FUSE focuses on forecasting science and technology events by leveraging multiple datasets such as patent data, and technical papers. It forecasts the labs and topics that will generate the most patents.

Dr. Brian Tivnen, Senior Principal at The MITRE Corporation and Chief Engineer for Modeling and Simulations Division, has been working on some serious issues relating to conflict and finance, as discussed in his research on machine ecology beyond human response time. However, further context to this work was deemed too sensitive and could not be discussed in the forum. Instead, Tivnen discussed dynamics that relate to large-scale, socio-technical systems where outcomes have critical and societal importance. Previously, Tivnen looked at 2005 counter insurgency operations dynamics in Iraq and Afghanistan. The data was very effective at counting direct observables in the supporting establishment that would help infer the dynamics that were unfolding in the counter insurgency operations. Allowing for analysis to discover other areas where the dynamics showed strategic interactions among autonomous and adaptive actors in a competitive context is what moved him to finance. He found high signal to noise ratio in terms of the granularity in the data, took lessons learned and used it in the significant activities database that they used for analysis. The nature of defense for insurgency versus counter insurgency is move-counter move. Tivnen looks at this problem as a Red Queen Dance, as per Alice in Wonderland: do all that you can do not to fall behind the adversary. He completed an analysis of fatal attacks in Afghanistan and market fractures were discovered. Move vs. counter move showed a robust pattern. The comparison of the data is statistically the same between the Iraq data and the Afghanistan. Statistically they are drawn from the same distributions. This matters because of unauthorized leaks of classified data that have come available. Claims were made that that the Pentagon was manipulating the data as it was being reported to mainstream media. MITRE did not touch the data, but universities they collaborated with did and they analyzed the data using the MITRE algorithms. Findings showed that the results were statistically the same as the unclassified data that was being reported. When US troops are killed in an attack, the reporting is accurate, hence the analysis focused here. Data for indigenous people killed or wounded data was noisy. Interval times on the red team attack on the blue team from the red team perspective were constructed. When looking at the data from this perspective of moves and counter moves, it fits a robust pattern. Power Law analysis shows that the pattern exhibits moves and counter moves in time and space, underpinning this analysis.

Tivnan then took this lesson learned in conflict to the financial world. The stability of a nation state is reliant on a stable economy and markets know when there is stability, as discussed previously by Brinkley. Anecdotal evidence supports this claim because of the financial virtual space where virtual actors compete. Transaction

data show dramatic price movement up and down over a very small timeframe. All spikes also reverse themselves in a short period of time by more than a percentage point which is a large amount of money. The redistribution of wealth was very large in these small periods of time. Where were these events occurring? He found that in the lead up to September 2008 Lehman Brothers filing bankruptcy, these fractures happened at an increasing rate right up to that point. The dozen most fractured stocks in the US economy were all international financial institutions. This can be a potential signal that can shine a light on a structure that this may or may not be a good thing to global market stability. Current data of this source is not open for free, but there is a set of data coming that will be free. Consolidated audit trail will give the ground truth regularity to all market transactions so it will not be just the privileged and wealthy that will have access to this data.

Content Understanding and Information Management

Harry Cooper is Chief of the Classification Management and Collaboration Group at the Central Intelligence Agency. As a founding member of the Joint Interactive Content Understanding Forum, he offered context on prior meetings and a direction for interagency workgroups.

The first Content Understanding Workshop was held January 8-9, 2013 and 90 representatives from 15 organizations attended. The group determined that in order to effectively manage content, a joint effort is needed across multiple organizations, including government, industry, and academia. The goals were formulated as follows:

- Pool Resources for Strategic Advancement - different groups need to be able to work together to achieve a better shared-awareness of content understanding.
- Provide Awareness- Keep senior leadership in the know about current methods and techniques to solve cyber related problems.
- Promote Adoption and Use of Content Understanding Technologies - Once a new tool or method has shown promise, we should be advocates of its use.

Cooper pointed out that the amount of data received is too large for humans to analyze themselves and there is a need to put scientists and analysts in the same room together to come up with ways to automatically conduct an effective analysis. He asserts that applications that we can put on desktops are needed, especially since organizations are moving to the cloud. Websites have been exploited by outside entities, generating the requirement to go through terabytes of data to determine the cost of damage to the enterprise both now and in the future. Putting eyeballs on the most important data is critical, as there could be tens of millions of pages in some of these websites. Cooper asserts the need to find a way to identify concepts not keywords. If Boolean searches are the only method, there will be too many false positives. Visual analytics may work to identify concepts, and further identification of concepts that are out there is needed. In the end, human ability must be effectively merged with machine intelligence to maximize the impact of these massive data flows on decision making. Support of organizations like IARPA, DARPA, and In-Q-Tel, in addition to this group, are needed to identify and fund the research that can build the requirements and task the researchers and scientist to solve the problem.