NPS-AM-12-205

ACQUISITION RESEARCH SPONSORED REPORT SERIES

Applications of Lexical Link Analysis Web Service for Large-Scale Automation, Validation, Discovery, Visualization, and Real-Time Program Awareness

23 October 2012

by

Dr. Ying Zhao, Research Associate Professor,

Dr. Shelley P. Gallup, Research Associate Professor, and

Dr. Douglas J. MacKinnon, Research Associate Professor

Graduate School of Operational & Information Sciences

Naval Postgraduate School

Prepared for: Naval Postgraduate School, Monterey, California 93943

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Abstract

DoD acquisition is an extremely complex system, composed of myriad stakeholders, processes, people, activities, and organizational structures. Processes within this complex system are encumbered by continuous creation of large amounts of unstructured and unformatted acquisition program data. Acquisition analysts and decision-makers must analyze this available data to obtain a complete and understandable picture. For those embedded within the complexities of the acquisition community, this effort represents a daunting, if not impossible, task. We apply a data-driven automation system, namely, Lexical Link Analysis (LLA), to help acquisition researchers and decision-makers recognize important connections (concepts) that form patterns derived from dynamic, ongoing data collection. This year we have built two use cases of the LLA web service to develop focused practice and theory. In practice, we have been examining both LLA and System Self-awareness (SSA) as knowledge management tools for scoring/ranking interesting information and for visualizing/reporting correlations among categories of information. In theory, we have shown how to optimize the overall fitness of the system by considering the trade-off between a node's authority and expertise. This work has advanced the DoD-wide effort of integrating and maintaining authoritative and accurate acquisition data services in both legacy and new platforms.

Keywords: Lexical Link Analysis, Text Mining, Data Mining, Program Elements, Major DoD Acquisition Programs, Universal Joint Task Lists, Resource Allocation, Warfighters' Requirement, Urgent Need Statements, Unstructured Data, Data-Driven Automation, System Self-Awareness, Knowledge Network Theory, Authority Centrality, Expertise Centrality

THIS PAGE INTENTIONALLY LEFT BLANK

About the Authors

Dr. Ying Zhao is a research associate professor at the Naval Postgraduate School (NPS). Dr. Zhao joined NPS in May 2009. Her research is focused on knowledge management approaches such as data/text mining, Lexical Link Analysis, search and visualization for system self-awareness, decision-making, and collaboration. She received her PhD in mathematics from MIT and co-founded Quantum Intelligence, Inc. She was principal investigator (PI) for six contracts awarded by the DoD Small Business Innovation Research (SBIR) Program. She was the co-author of two U.S. patents in knowledge pattern search from networked agents and in fusion and visualization for multiple anomaly detection systems.

Dr. Ying Zhao
Information Sciences Department
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-3789
Fax: (831) 656-3679
E-mail: **yzhao@nps.edu**

Dr. Shelley Gallup is a research associate professor at the Naval Postgraduate School's Department of Information Sciences, and the director of Distributed Information and Systems Experimentation (DISE). Dr. Gallup has a multidisciplinary science, engineering, and analysis background, including microbiology, biochemistry, space systems, international relations, strategy and policy, and systems analysis. He returned to academia after retiring from naval service in 1994 and received his PhD in engineering management from Old Dominion University in 1998. Dr. Gallup joined NPS in 1999, bringing his background in systems analysis, naval operations, military systems, and experimental methods first to the Fleet Battle Experiment series (1999–2002) and then to the FORCEnet experimentation in the Trident Warrior series (2003–present).

Dr. Shelley P. Gallup
Information Sciences Department
Naval Postgraduate School
Monterey, CA 93943-5000

Tel: 831-656-1040
Fax: (831) 656-3679
E-mail: spgallup@nps.edu

Dr. Doug MacKinnon is a research associate professor at the Naval Postgraduate School (NPS). Dr. MacKinnon is the deputy director of the Distributed Information and Systems Experimentation (DISE) research group where he leads multi-disciplinary studies ranging from Maritime Domain Awareness (MDA), to Knowledge Management (KM) and Lexical Link Analysis (LLA). He also led the assessment for the Tasking, Planning, Exploitation, and Dissemination (TPED) process during the Empire Challenge 2008 and 2009 (EC08/09) field experiments and for numerous other field experiments of new technologies during Trident Warrior 2012 (TW12). He holds a PhD from Stanford University, conducting successful theoretic and field research in Knowledge Management (KM). He has served as the program manager for two major government projects of over $50 million each, implementing new technologies while reducing manpower requirements. He has served over 20 years as a Naval Surface Warfare Officer, amassing over eight years at sea and serving in four U.S. Navy warships with five major, underway deployments.

Dr. Douglas J. MacKinnon
Information Sciences Department and Graduate School of Operational and Information Sciences
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-1005
Fax: (831) 656-3679
E-mail: djmackin@nps.edu

NPS-AM-12-205

ACQUISITION RESEARCH SPONSORED REPORT SERIES

Applications of Lexical Link Analysis Web Service for Large-Scale Automation, Validation, Discovery, Visualization, and Real-Time Program Awareness

23 October 2012

by

Dr. Ying Zhao, Research Associate Professor,

Dr. Shelley P. Gallup, Research Associate Professor, and

Dr. Douglas J. MacKinnon, Research Associate Professor

Graduate School of Operational & Information Sciences

Naval Postgraduate School

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

DoD acquisition is an extremely complex system composed of myriad stakeholders, processes, people, activities, and organizational structures. Processes within this complex system are encumbered by the development of large amounts of unstructured and unformatted acquisition program data, which, due to their enormity and complexity, are narrowly useful and difficult to aggregate across the enterprise. Acquisition analysts and decision-makers must, however, analyze all types and spectrums of the available data in order to obtain a complete and understandable picture. Considering the work that acquisitions systems must accomplish, there is a lack of internal congruence between multiple points at which the system should have knowledge of itself and of decision-makers who depend on aggregate information. Current information and decision support systems may not readily help overcome this difficulty, and they present users within the acquisition community with information overload and limited situational awareness. We believe that the application of a data-driven automation system—namely, Lexical Link Analysis (LLA)—can facilitate a resolution of acquisition researchers' data sense-making dilemma and help reveal important connections (concepts) and patterns derived from dynamic, voluminous, and on-going data collection.

In the past two years, we have utilized the LLA method to discover valid associations among disparate, unstructured data sets that would otherwise have required lengthy and expensive man-hours to analyze. The LLA technology and methodology were used to uncover and graphically display relationships among competing programs and to compare their features with Navy-driven requirements. In the past year, we tested our method for visualization and validity using samples of acquisition data.

During the research period begun in 2012, we achieved the following goals:

1. Conceptual and Focused Development

We continued using the LLA web service, hosted in the NPS Distributed Information and Systems Experimentation (DISE) lab with the link http://firedev2.ern.nps.edu:8080/ARP, to further assist the DoD-wide effort of integrating and maintaining authoritative and accurate acquisition data services in both legacy and new platforms.

We communicated with the Office of Secretary of Defense (OSD) contacts to identify the data sources from the Acquisition Visibility Portal for the Systems Engineering Plan (SEP), the Test & Evaluation Master Plan (TEMP), and the Acquisition Strategy Report (ASR). This provided opportunities to apply LLA to examine consistency, gaps, and data quality and to address the OSD requirements and understand program dependencies in ensuing research next year.

2. Theory and Methodology Development

The core technologies, LLA together with SSA, used throughout this project, have been examined thoroughly, in practice and in theory. In practice, we have examined both LLA and SSA as knowledge management tools for scoring/ranking interesting information and for visualizing/reporting correlations among categories/layers/systems of information, including lexical, semantic, and social links using the use cases. In theory, to take advantage of both concepts, we have demonstrated that it may be possible to stand outside a self-organizing system and optimize the overall fitness of the system by considering the centrality measures of authority and expertise scores. We have summarized the research in a journal paper entitled “Lexical Link Analysis (LLA) and System Self-Awareness (SSA): Theory and Practice,” planned for submission to the journal *ACM Transactions on Information Systems* (TOIS).

I. Significance of the Research

Acquisition research has increased in component, organizational, technical, and management complexity. It is difficult for acquisition professionals to remain continuously aware of their decision-making domains because information is overwhelming and dynamic. According to the *Chairman of the Joint Chiefs of Staff Instruction for Joint Capabilities Integration and Development System (JCIDS*; CJCS, 2009), there are three key processes in the DoD that must work in concert to deliver the capabilities required by warfighters: the requirements process; the acquisition process; and the Planning, Programming, Budget, and Execution (PPBE) process.

Each process produces a large amount of unstructured data; for example, the warfighters' requirements are documented in Universal Joint Task Lists (UJTLs), Joint Capability Areas (JCAs), and urgent need statements (UNSs). These requirements are processed in the JCIDS to become projects and programs, which should result in products such as weapon systems that meet warfighters' needs. Program data are stored in the Defense Acquisition System (DAS). Programs are divided into Major Defense Acquisition Programs (MDAPs), and Acquisition Category II (ACATII), and so forth. Program Elements (PEs) are the documents used to fund programs yearly through the congressional budget justification process. All the data is too voluminous, too unformatted, and too unstructured to be easily digested and understood—even by a team of acquisition professionals. There is a critical need for automation to help reveal to decision-makers and researchers the interrelationships within these processes (see Figure 1).

We have attempted to develop and frame our research efforts around research questions in the following categories: conceptual, focused, theory development, and methodology.

Conceptual

- How can the information that emerges from the acquisition process be used to produce overall awareness of the fit between programs, projects, and systems, and of the needs for which they were intended?
- If a higher level of awareness is possible, how will that enable system-level regulation of programs, projects, and systems for improvement of the acquisition systems?

Focused

- Based on the normal evolution of documentation and current data-based program information, how can requirements (needs) be connected to system capabilities via automation of analysis?
- Can requirement gaps be revealed?

Theory Development

- How can a correlation between system interdependency (links/relationships) and development costs be shown, if present?

Methodology

- How can we use natural language and other documentation (roughly, unformatted data) to produce visualization of the internal constructs useful for management through Lexical Link Analysis (LLA)?

Lexical analysis is a form of text mining in which word meanings are developed from the context from which they are derived. Link analysis, a subset of network analysis that explores associations among objects, reveals the crucial relationships between objects when collected data may not be complete. LLA is an extended lexical analysis and link analysis. LLA can also be used in a learning mode in which such features and contextual associations are initially unknown and are constantly being learned, discovered, updated, and improved as more data become available.

We consider that the cognitive interface between decision-makers and a complex system may be expressed in a range of terms or features (i.e., a specific

vocabulary or lexicon) to describe attributes and the surrounding environment of a system. Here, system self-awareness, or program awareness (Gallup, MacKinnon, Zhao, Robey, & Odell, 2009) allows decision-makers to be aware of what systems, programs, and products are available for acquisition; to understand how the systems match warfighters' needs and requirements; to recognize relationships among them; to improve efficiency of available collaboration; to reduce duplication of effort; and to reuse components to support cost-effective management with greater immediacy, possibly in real-time.

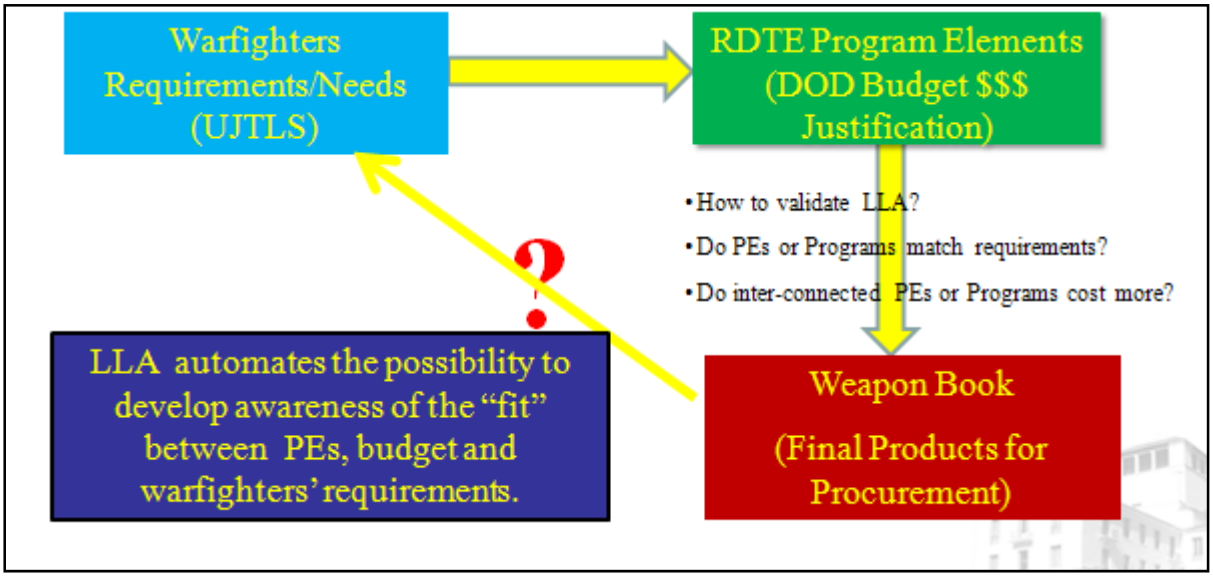

Figure 1. LLA Seeks to Inform the Business Processes Links (e.g., From Requirements to DoD Budget Justification to Final Products) That are Critical for DoD Acquisition Research

In precise terms, we observed three important processes that seem fundamentally disconnected. They are the congressional budgeting justification process (such as information contained within the PEs), the acquisition process (such as information in the MDAPs and ACATII programs), and the warfighters' requirements (such as information in UNSs and UJTLs). They were not analyzed and compared to each other in a dynamic, holistic methodology that could keep up with changes and reflect patterns of relationships.

There had been little previous effort to integrate the data in these three components. We analyzed in detail samples in the three components, validated the LLA method using large-scale data sets, and also successfully applied the method to discover the patterns in the data that were interesting and previously unknown to many acquisition professionals (Zhao, Gallup, & MacKinnon, 2010, 2011a, 2011b, 2011c, 2012a).

II. Research Results

The work, begun in 2012, has the following objectives:

- Build at least two use cases of applications of Lexical Link Analysis web service for large-scale automation, validation, discovery, visualization, and real-time program awareness.
- Demonstrate the methodology for assisting the DoD-wide effort of integrating and maintaining authoritative and accurate acquisition data services in both legacy and new platforms.

A. Use Case 1: Acquisition Program Awareness

We have conducted two research projects to date on this subject, namely "Towards Real-Time Program-Awareness via Lexical Link Analysis" (2010) and "A Web Service Implementation for Large-Scale Automation, Visualization and Real-Time Program-Awareness via Lexical Link Analysis" (2011b). This follow-up research (Phase III) extended the work to the previous two projects. We used this use case to answer the research questions stated previously regarding the levels of conceptual and focused theory development and methodology.

1. Conceptual Development

To realize the potential of the LLA method, we first established the validity of the method in the context of realistic, large-scale data sets, which include the budgeting process through PEs to the acquisition process via acquisition programs (MDAPs, ACATIIs) to the warfighters' requirements (UNS, UJTL, etc).

1) Congressional budget process (i.e., Program Elements [PEs]):
http://www.dtic.mil/descriptivesum/

2) Programs and products (MDAPs and ACATIIs):
http://comptroller.defense.gov/defbudget/fy2008/fy2008_weabook.pdf
http://www.fas.org/man/dod-101/sys/land/wsh2007/index.html
http://www.acq.osd.mil/ara/am/sar/

3) Requirements (i.e., UJTLs): http://www.dtic.mil/doctrine/jel/cjcsd/cjcsm/m350004d.pdf

Result 1: We found that the Pearson correlation between the links identified by human analysts and by the LLA method was 0.57 with a *p*-value = 10e-7 (Zhao et al., 2010, 2011b). LLA was used to predict correctly 80% of the links identified by the human analysts.

The high correlation of LLA results with the link analysis done by human analysts makes it possible for automation, saving human effort, and improving responsiveness. Automation is achieved via computer program or software *agent(s)* to perform LLA frequently—and in near real-time. Agent learning makes it possible to reach real-time; visualization correlates lexical links to core measures; features and patterns are discovered over time for the system as a whole. We can take advantage of the data in motion (social media data) and RSS feed data to build a better picture of real-time program awareness.

An accurate text analysis requires a thorough initial search of the resources available on the Internet. At this point, our efforts are sometimes compared to those of a typical search engine. One of the disadvantages of conventional search engines is that they typically sort documents based on the popularity of documents, e.g. the frequency with which they're linked to other documents, not based on semantics. Therefore, it does not satisfy completely the frequency with which they're linked to other documents search needs nor determine relevance if the links among the documents are not available. For example, the content in the forum is not cross-linked; therefore, if conventional search engines are used, the discovered or *revealed* topics or themes cannot be found as prioritized results.

2. Focused Development

Result 2: We took a detailed look at RDT&E budget modification practices from 2008–2009, specifically, the observed percentage change of funding for the approximately 500 PEs from 2008–2009. For the PEs whose number of LLA links to

other PEs was larger than 10, the funding change was 14%, compared to 40% for those whose number of LLA links to other PEs was fewer than 10. This indicated the current practice tended to increase the budget *less* for the PEs with more links to other PEs and to increase the budget *more* for the ones with fewer links, an effort to allocate resources to avoid overlapping efforts. In a different perspective, the overall numbers of LLA links to the UJTLs were much fewer. The PEs that had at least one LLA match to UJTLs, had an average percentage cost increase of 10%, compared to 29% for the PEs that had no matches. This indicates a need to consider gaps and the warfighters' requirements as priorities in the RDT&E investment (Zhao et al., 2011a, 2011b).

This demonstrated that our approach "discovers" and displays semantic and social networks of programs and PEs. It discovers blind spots on the part of human analysts that are caused by overwhelming data. These findings can be useful as validation and guidance for implementing the DoD's budget reduction planning. Patterns revealed by LLA create an opportunity to reduce the overall inefficiency of cost cutting by linking programs with warfighters' requirements, as opposed to cost cutting, which focuses mainly on the big ticket items such as MDAPs.

Result 3: We used the LLA method to generate semantic networks for the PEs, in which two PEs are connected if they are discovered to use similar lexical terms from the LLA method shown in Figure 2. The size of a node in Figure 2 shows the percentage of the budget increase from a current year to the following year. The network shows that the more connected programs tend to be in the middle with smaller nodes, while the less connected programs tend to be on the outside with larger nodes. This pattern indicates the correlation between independencies of programs (i.e., the connections among nodes in Figure 2), and cost increases (i.e., the size of the nodes in Figure 2). The social network links marked by human analysts, in contrast, do not reveal this pattern.

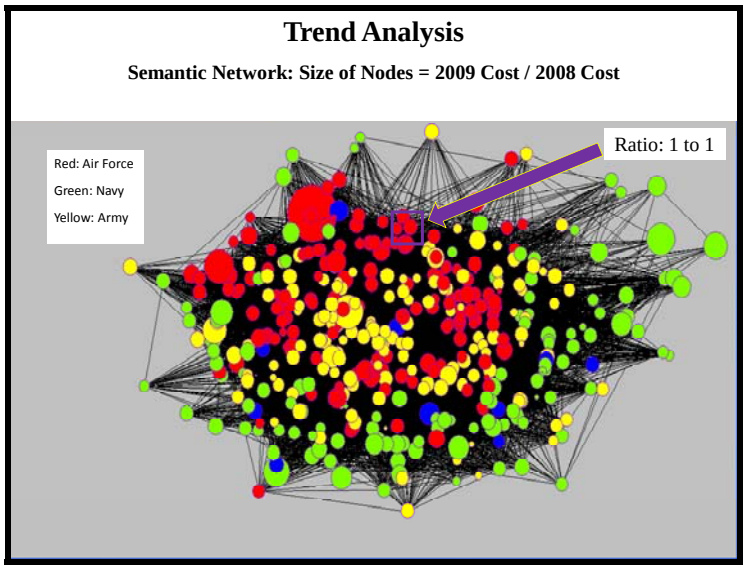

Figure 2. A 3-D View of PEs Identified by the LLA Semantic Network

Result 4: We developed a web service to link the available public data in the budgeting process through PEs to the acquisition process via acquisition programs (MDAPs) responding to warfighters' requirements (UJTLs). The web service is currently hosted at the NPS DISE lab and the link is http://firedev2.ern.nps.edu:8080/ARP as shown in Figure 3.

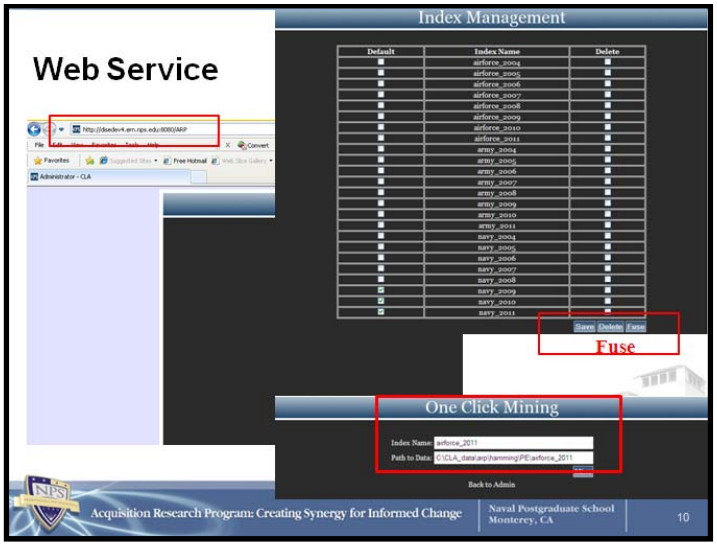

Figure 3. Acquisition Web Service Hosted in DISE

Figure 4 shows a summary of the approximately 30 top themes for the PEs for all three services (2004–2011) using the web service result in Figure 3. "LLA Counts" are the number of lexical links (word pairs) that were categorized into one group as a theme characterized by keywords. "Max Source" shows which data source (e.g., navy_2011 in the first row) shows the maximum number of LLA counts compared to the other sources for that theme.

Theme Id	LLA Counts	Max Source	Theme Keywords
1817	1876	navy_2011	COST,SUPPORT,TOTAL
1726	1533	navy_2008	TECHNOLOGY,BASED,MATERIALS
1555	743	navy_2006	PROGRAM,TEST,FLIGHT
1851	679	navy_2006	SYSTEM,TRAINING,MANAGEMENT
793	627	navy_2008	DEVELOPMENT,UPGRADE,NAVY
741	599	navy_2011	DATA,ANALYSIS,MODELS
1864	565	navy_2011	SYSTEMS,DESIGN,HARDWARE
1659	427	airforce_2008	RESEARCH,TECHNOLOGIES,SPACE
555	411	navy_2007	TESTING,CAPABILITY,ARCHITECTURE
1005	388	navy_2007	CONGRESSIONAL,IOC,MS
1477	348	navy_2009	PERFORMANCE,AIR,CONTROL
1767	313	navy_2005	SOFTWARE,FLEET,SW
1170	306	navy_2004	JOINT,COMBAT,MILITARY
1105	303	airforce_2010	INFORMATION,COMMON,FUTURE
1125	299	navy_2008	INTEGRATION,DT,WAVE
871	277	navy_2006	CAPABILITIES,EFFORTS,ASSESSMENT
431	270	navy_2004	REQUIREMENTS,INITIATIVE,AREAS,DEVELOPED
1124	266	navy_2004	INTEGRATED,PRODUCTION,FORCE
1426	259	navy_2011	OPERATIONAL,FUNDING,STUDY
768	256	navy_2010	OPERATIONS,MISSION,ACTIVITIES
374	238	navy_2005	ADVANCED,IMPROVED,ADDITIONAL
1872	219	navy_2006	TACTICAL,UNMANNED,PRECISION
1040	205	navy_2008	ACCOMPLISHMENTS,HIGH,OPN
750	201	navy_2008	VEHICLE,DEMONSTRATION,PLANNING
638	189	navy_2006	PHASE,COMPLETE,FULL
763	187	navy_2004	EFFORT,DELIVERY,MEDICAL
630	182	navy_2010	COMMUNICATIONS,EQUIPMENT,COMMUNICATION,EXISTING
906	180	navy_2011	ENGINEERING,ENVIRONMENTAL,NATIONAL
2029	175	navy_2010	INCREMENT,WEAPON,FCS,PLATFORM
1134	168	navy_2004	ENGINE,INTERFACE,GROUND
1364	160	navy_2011	NETWORK,CONCEPT,THREAT
1068	158	navy_2008	II,III,BLK
1571	155	army_2008	PROTOTYPE,SUPPORTS,CENTER
387	153	navy_2004	AIRCRAFT,GPS,MODULE

Figure 4. Summary of Approximately 30 Top Themes for the PEs for all Three Services (2004–2011)

Observations are summarized in the bullet list for Figure 4.

- The PEs' content seems dominated by the Navy. This might be due to the fact that the Navy provides better (e.g., more specific) PE descriptions.
- The Navy could also, indeed, provide the leadership for the overall RDT&E effort for the DoD, as evidenced by the highlighted (yellow) six

themes in 2011 which are dominant by the Navy (see the Max Source column). These new themes are consistent with the findings from the research categories and reports for the Acquisition Research Program hosted at NPS (Zhao et al, 2012a, 2012b), as well as the new trends for the defense industry.

3. Plans for 2013

Our goal is to demonstrate the LLA web service can further assist a DoD-wide effort to integrate and maintain authoritative and accurate acquisition data services in both legacy and new platforms.

Specifically, we will examine data sources from the Acquisition Visibility Porta l (AVP) for the Systems Engineering Plan (SEP), the Test & Evaluation Master Plan (TEMP), and the Acquisition Strategy Report (ASR).These data will be used to examine consistency, gaps, and data quality, and explore LLA visualizations and reports. Results will be validated against the samples of completed/approved Information Support Plans (ISPs; http://jitc.fhu.disa.mil/jitc_dri/pdfs/interim_guide_interoperability_nss_mar_12.pdf, https://gtg.csd.disa.mil), and associated milestone artifacts, where program managers identified program dependencies.

B. Task 1: Work With the AVP Data Source and Ongoing Requirements

This year, we will focus on more specific challenges, questions, and data sources found in the following communications:

1. Data Sources

We will use the data from the AV Portal (https://portal.acq.osd.mil & https://portal.acq.osd.mil/portal/server.pt/community/acquisition_visibility/1427), the Kaleidoscope analysis tool, DAMIR, and AIR (Acquisition Information Repository).

2. Analysis Directions and Questions

We communicated with the OSD contact, Mr. Robert Flowe regarding the AVP requirements and how LLA might help AVP during the 2012 symposium. The following is the discussion between Mr. Flowe and Ms. Zhao

Flowe: "I wonder if I could 'prime the pump' regarding potential applications of LLA to the issues we're grappling with at OSD? I hope you don't mind if I indulge in a little 'stream of consciousness' musing about where LLA could really add value. One of the biggest risk factors we're facing in defense acquisition is the unanticipated effects of program interactions. ASD(SE) and Dahmann work on identifying interdependence among programs within SoS as a risk driver. More generically, you can call it the result of joint capabilities, portfolios, program interdependencies, system-of-systems effects, or whatever, the bottom line is that our 'program centric' acquisition paradigm is increasingly ill suited to identify and address program risks that arise outside of the program boundary. I think LLA could help us isolate these issues from the body of information we currently collect, but have yet to effectively utilize."

Zhao: "Yes, we would love to work with you to work on these specific requirements."
Flowe: "Part of the problem is that very little of the information generated for program oversight is amenable to effective analysis. Every major acquisition program's milestone review generates volumes of information, which the OSD staff is supposed to review to determine if the program is properly prepared for the next milestone. Although we are beginning to compile these artifacts centrally to facilitate review/analysis, the fact remains that the only way to analyze the information in these artifacts is to read them. With limitations on staffing, little time is available to thoroughly review the artifacts. Moreover, each functional community is required to review only the particular document it is responsible for. So the Systems Engineering community looks at the Systems Engineering Plan (SEP), the Test and Evaluation community looks at the Test & Evaluation Master Plan (TEMP), the Acquisition community looks at the Acquisition Strategy Report (ASR), but rarely do

any of these stakeholders review multiple reports or jointly discuss them to determine if they are mutually consistent and flag inconsistencies that might indicate programmatic risk. There is even less incentive/opportunity to look for external factors that would potentially invalidate the assumptions that underpin the basic cost/schedule/performance targets the program is executing to. I think that LLA might help cue our attention to these issues, by examining the various milestone artifacts, extracting the lexical links, and portraying a map of linked concepts for each artifact."

Zhao: "To achieve this, we can do by examining SEP, TEMP and ASR, and, thereby, discover inconsistencies and gaps. We can begin by studying the overall patterns first, then go deeper, searching for data gaps and inconsistencies."

Flowe: "Overlaying the concept maps for each of the major artifacts to do a pair-wise comparison might expose significant disconnects between, say, the acquisition strategy and the systems engineering plan, or the SEP and the TEMP. Consider a situation where the SEP identifies a critical dependency between the program and an external system, but the TEMP doesn't have a corresponding reference to testing that interdependency. If LLA could highlight these inconsistencies for further scrutiny, it would help the staff identify significant risks that might otherwise go undetected until later in the program, when opportunities for recovery are limited."

Zhao: "When examining SEP, TEMP, and ASR, we can further examine critical dependencies and report them as themes, concepts, and word pairs, thereby offering specific and productive directions for further scrutiny."

Flowe: "A similar application of LLA would be to compare lexical link maps of the same artifact from one milestone to another. If the lexical link map for the SEP at M/S-B is significantly different from the SEP at M/S-C, that might indicate a reduction in system functionality resulting from cost increases elsewhere.

Zhao: "This is interesting as an initial study to see if the correlation is there"

Flowe: “For identifying external dependencies, I wonder if LLA could be “trained” to distinguish program names or organizational names from other nouns? This would be helpful in identifying external dependencies. We could provide a list of program names that could be used as a training aid or reference table.

Zhao: “We have a context-dependent entity extraction tool that might be able to perform this task.”

Flowe: “Recalling my question at the symposium, I also wanted to explore whether LLA could be used to extract the implicit semantic layer from program documentation, in the form of subject-object-predicate triples. We’re hoping to leverage ontological formalisms to facilitate the alignment of disparate data sources at the enterprise level. We’ve found that each functional domain within AT&L has unique meanings for the data they utilize, so mapping from one domain to another is problematic. We can create significant confusion if we don’t properly account for these domain-unique semantics—the term “program” doesn't mean the same for a program manager and a CAPE resource analyst. We believe that constructing reference ontologies for each domain will help us identify and mitigate these differences, and become more efficient in providing consistent enterprise-wide acquisition data.”

Zhao: “LLA is based on statistical bi-gram text analysis. The subject-object-predicate triples are usually generated by linguistics-based text analysis. Stanford Lexical Parser is one of these tools. We may combine the two approaches (when language is known, e.g., English) to get better results.”

C. Task 2: Explore New Visualization and Big Data Technologies for the Acquisition Research Web Service

There are some significant industry trends in recent years regarding data scale-up and visualization. In this task, we want to investigate new big data technologies such as HDFS (Hadoop Distributed File System) and MapReduce as

an alternative to high performance computing to parallel process the acquisition research web service. We will also investigate how to benefit from using InfoVis, a web visualization tool in JavaScript to replace Organizational Risk Analyzer (ORA) software (Reminga & Carley, 2003) that is currently used to visualize LLA results.

1. Methodology

We have further developed the LLA methodology used throughout this project which is summarized in detail in Appendix A.

D. Use Case 2: Analysis of the Acquisition Research Program (ARP) Data

We applied LLA to eight years of research report data for the NPS Acquisition Research Program. We downloaded about 740 publications (from 2003–2010) from the website http://www.acquisitionresearch.net.

1. Pre-Defined Categories

Each report was labeled manually with a category, for example, “Acquisition Strategy” or “Costing.” Approximately 160 categories were created for the years 2003–2010. Table 1 shows the number for each category and year. By observing the bubble chart derived from LLA, we found three categories:

- Steady categories in which the number of reports increased from 2003 to 2010 as shown in Figure 5.
- New and emerging categories in which there were relatively new information from 2006–2010 compared with 2003–2005 as shown in Figure 6. These categories attracted more research attention in the years that followed.
- *Sunsetting* categories in which the number of reports reduced from 2006–2010 compared with 2003–2005 as shown in Figure 7.

Table 1. ARP Reports From 2003–2010

Year	# of Reports	# of Categories
2003	8	6
2004	27	17
2005	61	34
2006	62	29
2007	143	63
2008	144	68
2009	127	61
2010	184	65

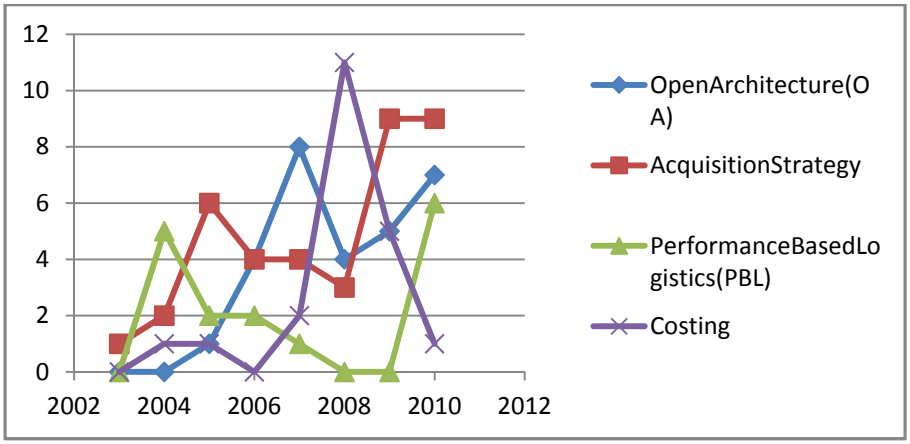

Figure 5. Steady Categories

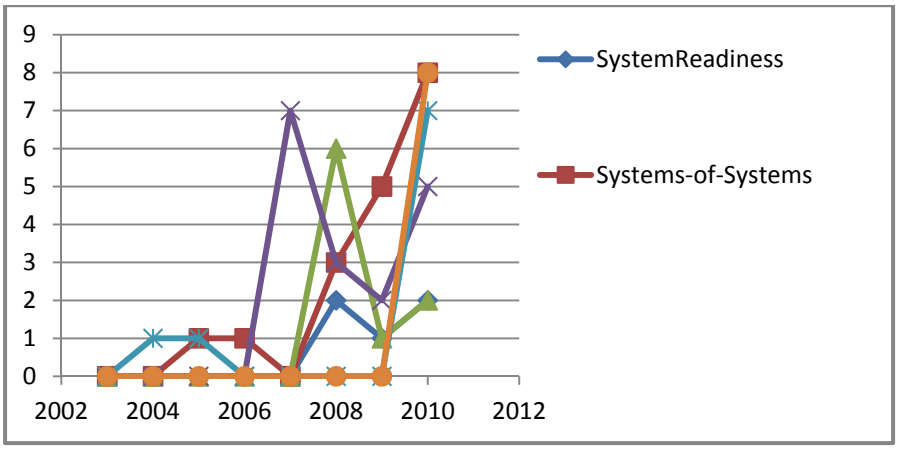

Figure 6.New and Emerging Categories

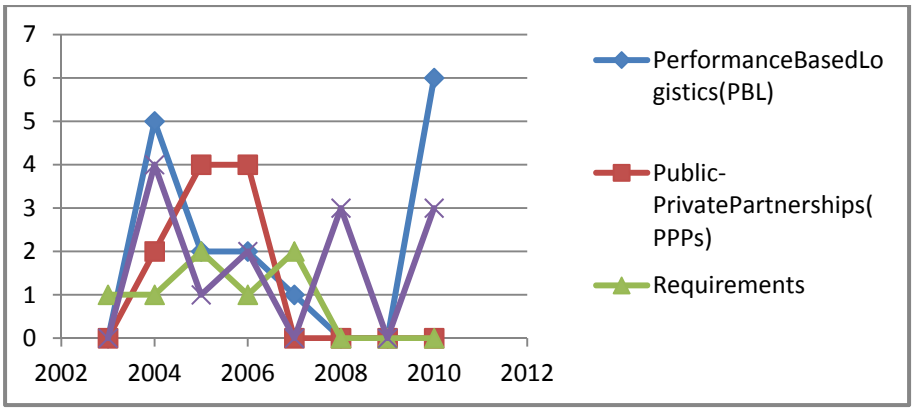

Figure 7. Sunsetting Categories

The question that arises is, what are the characteristics of the three categories? We first sorted the existing combinations of year (2003–2009) and 160 categories (e.g., 2003-AcquisitionStrategy and 2004-Outsourcing, etc.). There are a total of 245 such combinations. For each of these combinations, we labeled it 1 (*kept*), if the associated category was continued in the following year (e.g., 2003-AcquisitionStrategy is an existing category and 2004-AcquisitionStrategy is also a category); 0 (*deleted*), if the associated category was not continued in the next year (e.g., 2003-ContractCloseout is an existing category, but 2004-ContractCloseout is not—no reports were classified in the ContractCloseout category in 2004).

The combinations and labels represent the two decision-making processes in the Acquisition Research Program, namely

1) Whether or not a research area or project should move forward from one year to another, and

2) How a research area or project might be categorized.

By furthering our understanding of how dynamics of the combinations were *kept* or *deleted* from 2003–2010, we hope to explore how decisions were made in the current process, and, more importantly, to discover the characteristics of research areas (i.e., categories that are emerging from the past to the present, and to the future).

Figure 8 shows a detailed view of the semantic network for 2003 with six categories; the two in red were deleted in the following year; and the four in green were kept. Initially looking at each year seems to indicate the deleted nodes are associated with the “hot” links (i.e., links in red and orange colors). For example, two such nodes (2003-ContractCloseout and 2003-CostasIndependentVariable) are red for 2003. This may indicate one of the characteristics of the emerging and steady growing categories is to have unique content compared to the existing information at the time. Observing Figure 8 and Figure 9, we realize the deleted categories might also be associated with the “cold” links (i.e., links in green and blue). One such node (2004-LogisticsModenizationProgram) is shown in Figure 9 and is shown in the border of the graph, indicating total degree of the node might also be low.

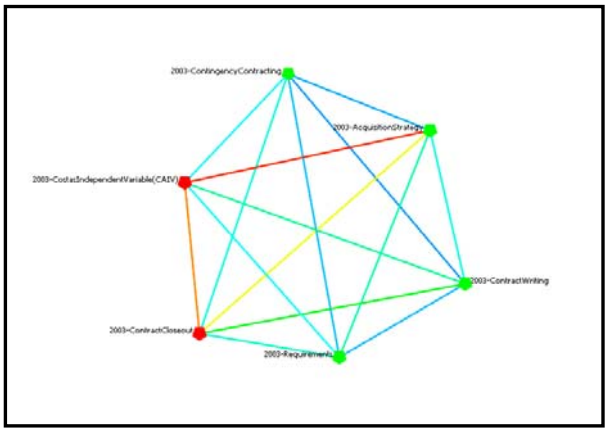

Figure 8.Semantic Network of Year-Category for 2003

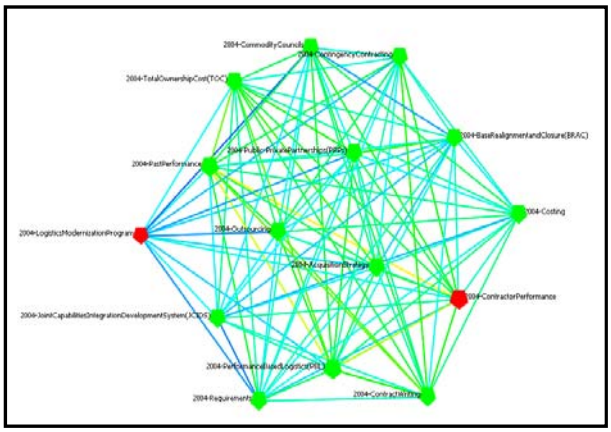

Figure 9. Semantic Network of Year-Category for 2004

Emerging categories tend to form *fewer but stronger* links (i.e., Group B & D, which have higher kept rates). This type of node is likely to reside in the "Ring of Emergence" as shown in Figure10 between the red and green circle. The most central one, Group B & C, represents categories that are well-researched and that the research community are already aware of and, therefore, are less likely to grow. The nodes located at the borders (Groups A & C and A & D) represent categories with weaker connections with others, some are even isolated, and that are, therefore, also less likely to grow.

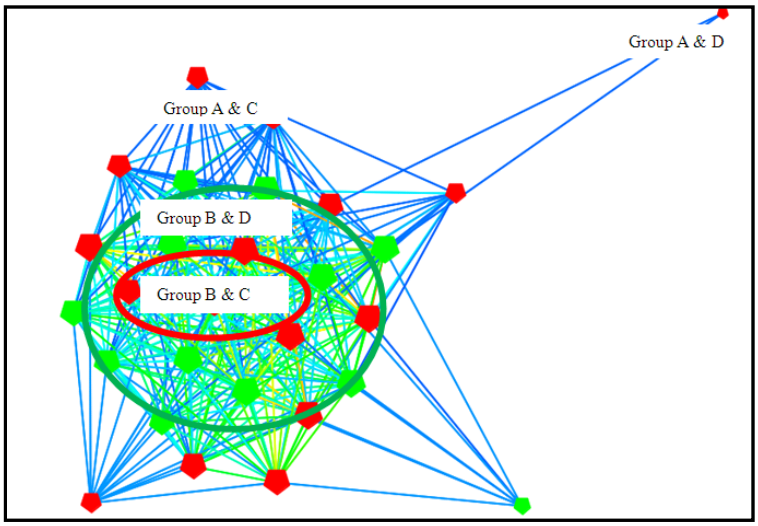

Figure 10. Ring of Emergence

We define *system self-awareness* of a complex system as its ability to assess itself within a global context. This concept is connected to the network theories and self-organizing features of the complex systems we reviewed. In particular, we are interested in the following network measures for SSA, in which a node, representing a document, a concept, a theme, a person, or an object with features of lexical terms, assesses its position in a global context via the connections to other nodes. The terms we develop and use are as follows:

- Eigenvector: The principal eigenvector of a network matrix. The eigenvector centralities capture not only the number of neighbors a node has, but also take into account the importance of each neighbor.
- Authority centrality: in-links are considered in the eigenvector centrality.
- Hub centrality: out-links are considered in the eigenvector centrality.
- Betweenness: How frequently the node is part of the shortest paths between pairs of other nodes in the network.
- Burt constraint (Burt, 1992): The degree to which a node in a network is constrained from acting because of its existing links to other nodes.
- Closeness: The length of the shortest or average path from the node to the rest of the nodes in the network.
- Expertise (Carley, 2002): The degree to which each pair of nodes has complementary links, expressed as a percentage of the links of the first node.
- Simmelian ties (Krackhardt, 1998): The normalized number of Simmelian ties of each node that are often associated with the number of brokers embedded in the fully connected nodes (cliques). For example, if a man has a strong tie to a woman, and both of them share a strong tie to their child, then the tie between the man and woman is considered stronger and is, therefore, *Simmelian*.
- Triad count (Reminga & Carley, 2003): The number of triads centered at each node in a square network. A triad is a relationship amongst three nodes such that they constitute a distinct relationship. This may make the structure of a network more stable.

Many other centrality measures other than LLA-related ones are defined, and are further discussed and computed in the ORA (Organizational Risk Analyzer) software (Reminga & Carley, 2003). These network centralities characterize in various ways a knowledge node's position with other nodes.

Figure 11 shows the relations among many centrality measures in the social network analysis (SNA) context; these measures can be used to evaluate the importance of lexical terms. Notice that the PageRank used by Google uses one of the measures.

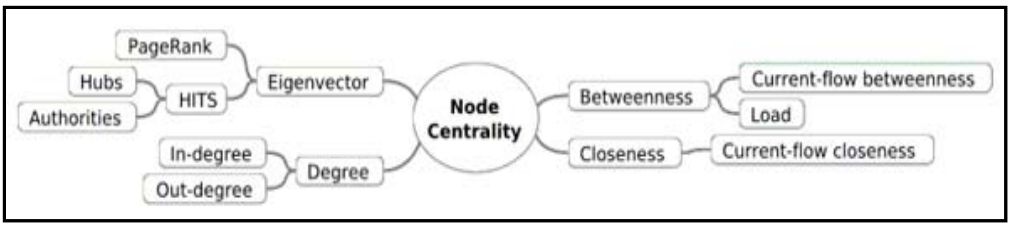

Figure 11. Relations Among Centrality Measures

LLA semantic/statistical techniques revealed certain common characteristics of interesting information (e.g., the growing themes). Eigenvector centrality sorted from low to high can be used as an indication for *interestingness*, which is typically an area for an expert to investigate or invest.

Table 2 shows the results of statistical significance tests for the two groups of pre-defined categories (*kept* and *deleted*) using more centrality measures. The kept nodes have statistically significant higher authority centrality, but lower total degree, Simmelian ties (Krackhardt, 1998), and triad count (Reminga & Carley, 2003) centralities. Simmelian ties and triad counts are traditionally considered as measures of the stability of the social network structures. Along with the total LLA degree, they indicate that human decisions may focus on the information (e.g., ARP pre-defined categories) that exhibits weaker stability as semantic networks and, therefore, possesses the potential to change or grow.

Table 2. Statistical Significance Tests (Pre-Defined Categories)

	Centrality Authority	Simmelian Ties	Total LLA Degree	Triad Counts
Kept	0.732	0.123	0.415	1967.766
Deleted	0.665	0.150	0.478	2646.340
p-value	0.015	<0.0001	0.028	0.0002

2. Using Automatic Discovered Clusters (Self-Organizing)

We also applied the automatically discovered themes as categories to see if the same theory applies (i.e., the automatically generated themes combined with

years as categories, 225 of such automatic categories; e.g., 2003-COST*COSTS*TOTAL and 2004-SYSTEMS*SYSTEM*PROGRAM). We define a value of an automatic category as

> *# of lexical links in the time frame for the theme – # of lexical links in the next time frame for the same theme.*

The goal is to compute the centrality measures for the 225 node semantic network, generated from the 225 automatic categories, in which links are only computed within the same time frame. We also computed the correlation between the centrality measures and "values" of the nodes.

Table 3 shows the results of statistical significance tests for the two groups, and represents values of growing or *sunsetting* themes for the automatically discovered themes using more centrality measures. The growing nodes have statistically significant higher authority and *betweenness* centralities with statistically significant fewer triad counts. The total degree of centrality is lower, but not statistically significant.

Table 3. Statistical Significance Tests for the Growing and Sunsetting Groups for the Automatic Categories

Node ID	Centrality Authority	Centrality Betweenness	Centrality Total Degree	Triad Count
Growing	0.43	41.70	0.18	50.28
Sunsetting	0.35	28.50	0.19	65.79
p-values	0.043	0.086		0.038

3. Theory Development

The core technology, Lexical Link Analysis (LLA), which we applied in theory and in practice to our use case, is a form of text analysis in which words and concepts, and their meanings, are represented as networks. LLA discovers and displays these networks of word pairs (i.e., semantic networks, from large-scale

unstructured data). This type of text-as-networks (TAN) has many advantages over other text analysis tools. When concepts and ideas are represented in lexical terms as if they are in communities of word networks, network theories (e.g., centrality measures that typically measure the position and influence of a node in a social network) and the preferential attachment theory (Barabási & Albert, 1999) of network growth, can be readily applied to evaluate the importance of lexical terms in a global context of interconnected concepts, ideas, and themes. Traditionally, *authority* centrality has been widely used to evaluate the importance of a network node in various applications, from ranking leadership in a social network to ranking a web page on the Internet.

In contrast to authority measures, which are primarily used to represent established values such as power and leadership of nodes in a network, which were rarely examined in the past, are the new types of centralities (e.g., *expertise* measures) which measure a node according to its degree of uniqueness and innovation for a concept, an idea, or an organization. As we will show in this paper, they can be computed from categories of information using LLA. Expertise measures are more interesting because they seem to correlate with real-life values, such as the growth potential of a new or emerging concept, return of investment of a new business idea, and competiveness of an organization.

In a system-of-systems point of view, a category of information, which is stored in unstructured texts, can be represented as a system of semantic networks using LLA. A system “self” can be a node in such a semantic network. System Self-Awareness (SSA) refers to the “self” which is aware of itself in two ways: its relations to others (i.e., its authority and influence in a network) and its expertise (i.e., its innovation and uniqueness in a global context). Each self node can also be independently evaluated and associated with a “value,” which represents the measurable importance of the self node, such as its growth potential and competiveness. The correlations between the value of a self node and its self-

awareness measures will serve as predictions of the emerging properties of information

In practice, we have been examining both LLA and SSA as knowledge management tools for scoring/ranking interesting information and for visualizing/reporting correlations among categories/layers/systems of information, including lexical, semantic, and social links using four use cases. This effort then presents decision-makers with previously unavailable/emerging patterns and themes, as well as unprecedented levels of analysis, thus reducing the workload and overcoming the blind spots of human analysts and providing opportunities for potential automation.

In theory, to take advantage of both concepts, we have been showing that a decision maker may want to stand outside a self-organizing system and optimize the overall fitness of the system by considering the trade-off between nodes' self-awareness of their own authority and expertise. When computing the fitness of a system, a larger weight toward authority over expertise will result in a network growing with preferential attachment theory. Conversely, a larger weight towards expertise over authority will result in a network which is more competitive and gain a bigger return on investment.

A different correlation (i.e., Pearson correlation coefficients) was computed for the various centralities and values as defined in the use cases for the nodes. For the automatically discovered clusters, the authority, betweenness, and correlation expertise centralities had positive correlations of 0.23, 0.24, and 0.19 ($p < 0.05$) with the defined node values. For the pre-defined ARP categories, the correlation expertise centrality had a positive correlation 0.15 ($p < 0.05$) with the node values, while the total LLA degree and triad count centralities had negative correlations of -0.12 and -0.17 ($p < 0.05$), respectively. These results seem to suggest that the interplay of authority and expertise centrality measures is important for the growth of a self-organizing system, and expertise measured via correlation expertise and LLA

total degree could provide indications of a system self-organizing into a network of experts.

Recall that expertise centrality is a measure of the degree to which each pair of agents has complementary knowledge, expressed as a percentage of the knowledge of the first agent (Carley, 2002),

$$\frac{\sum_{k=1}^{K}(1-TD\_LLA_{ik})*TD\_LLA_{kj}}{\sum_{k=1}^{K}TD\_LLA_{ik}}, \quad (1)$$

where TD_LLA_{ik} represents the total degree for agents *i* and *k* from a semantic network. Such a semantic network is generated from the content stored in the memory of agent *i* or *k* using LLA. The agent learning theory is explained in the next paragraph

As illustrated in Figure 12, to automate human cognitive tasks (e.g., to separate and extract information automatically from the documents) we *train* synthetic learning agents to perform human tasks. Modern agent-based modeling and simulation systems originated using concepts such as cellular automation from the game of *Life* invented by John Conway in 1970. This began the development and implementation of genetic algorithms (Goldberg, 1989) and other artificial intelligence techniques to improve the ability of one agent acting alone. Synthetic, multi-agent, distributed networks were then developed to provide for an integrated community of heterogeneous software agents, capable of analyzing and categorizing large amounts of information and, thus, supporting complex decision-making processes (see Figure 13). At present, self-managing (Hinchey & Sterritt, 2006), self-healing (Dashofy, van der Hoek, & Taylor, 2002), self-optimizing, self-configuring, and self-adapting software agents are desirable to automate ongoing human cognitive tasks in a complex network environment.

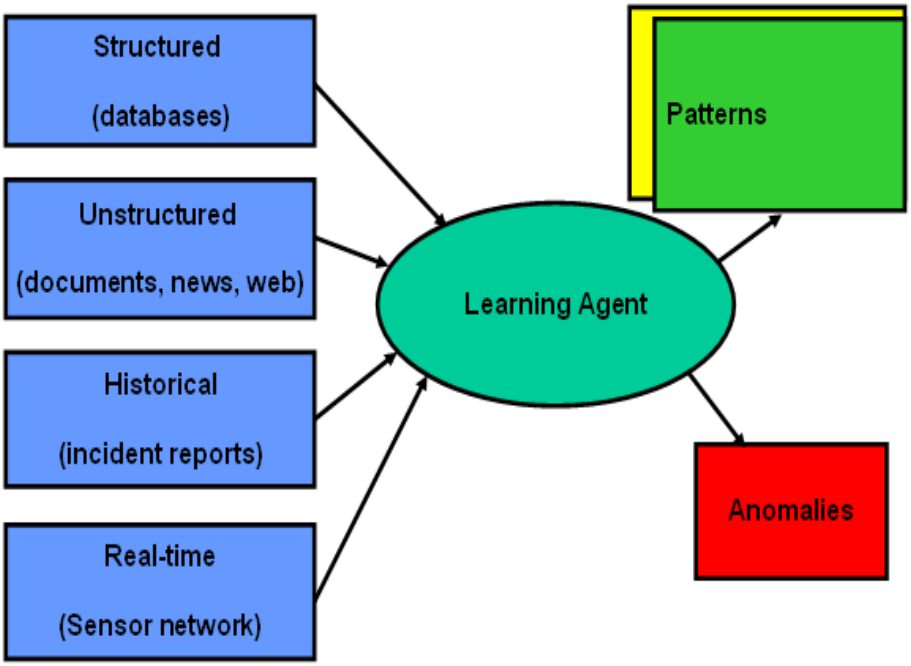

Figure 12. A Single Learning Agent Ingests Structured, Unstructured, Historical, or Real-Time Data and Separates Patterns and Anomalies

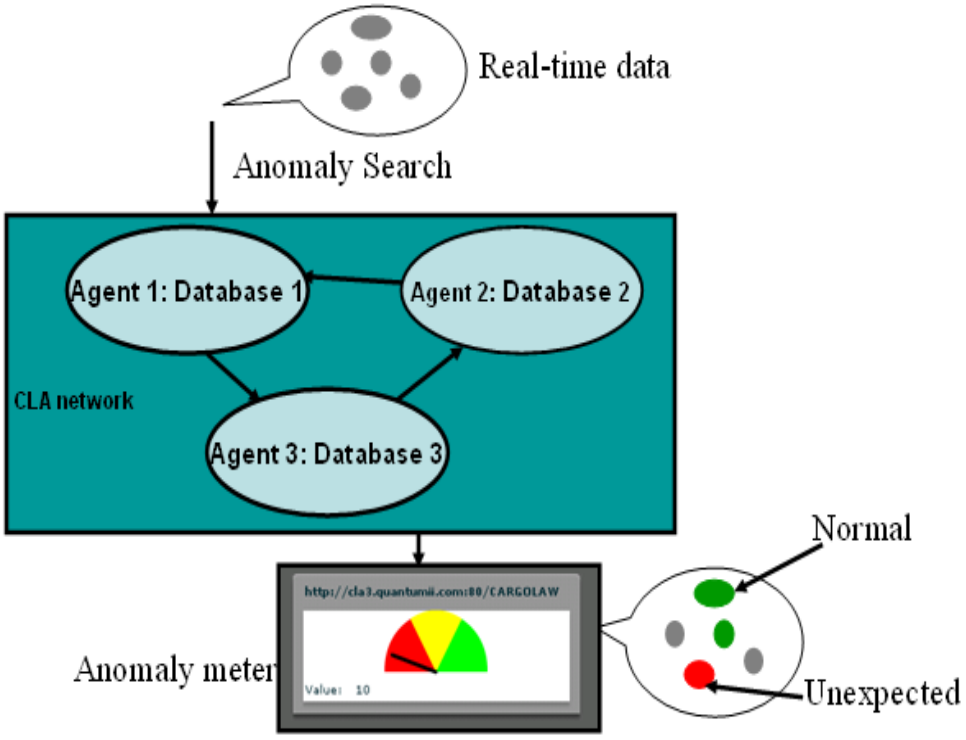

Figure 13. Agent Collaboration: Multiple Agents Work Together for Anomaly Search

Our research creates and develops a computer-based learning agent capable of ingesting and comparing a wide range of data sources, while employing a process that separates patterns and anomalies within the data. Multiple agents can work collaboratively in a peer-to-peer network as shown in Figure 13. The Collaborative Learning Agents (CLA) were first invented and implemented by Quantum Intelligence, Inc. (QI, 2001–2012). The unique contribution of this architecture is to leverage a peer-to-peer agent network in which each agent can be self-aware of its position in a global knowledge network in order to be competitive as well as collaborative. We show here that the overall fitness of a network of agents can be accomplished through a learning framework, in which each agent contributes to the overall effect through a trade-off between its authority and expertise measured as centralities in the global knowledge network as follows.

At any given time, we are able to rank a knowledge artifact (a document, concept, theme, or category) based on its predicted future importance, and distribute knowledge among collaborative agents (organizations, stakeholders). On a theoretic level, we will use a time series model as follows:

Observation x_t: Data for a single agent that is observable (e.g., measures of a single agent's awareness of information [or expertise] using lexical links extracted from its stored content).

State *j*, *j = 1 ,… J:* For different types of expertise, a transition matrix r_{ij} is used to describe how one type of expertise, *i,* is transited to another type of expertise, *j*.

An agent can have one or multiple types of expertise. For simplicity, we assume one agent only focuses on its best type of expertise, though, when it needs other types of expertise, it may collaborate with other agents through lexical links via semantic networks or social connections through a peer list of friends. We also model the relation as a probability density function *bj[O(t)]* between lexical links of a single agent's content input, *O(t)*, and states as different types of expertise *j* that are observed from the whole network.

This approach is related to the Expectation and Maximization (EM) method in statistics. It is a statistical method used to compute maximum likelihood estimates given incomplete samples **(**Dempster, Laird, & Rubin, 1977). Here we describe how to use a tied-mixture EM algorithm to compute the correlation or affinity between an input content *x* and a type of expertise *j*

Let $b_j(x)$ be a likelihood function $P(x \mid j)$, where *j* represents an expertise *j*. $P(x \mid j)$ represents the likelihood of producing content *x* if an agent possesses an expertise *j.* For a joint likelihood of multiple agents, given all the parameters associated with a model λ ,

$$f(x \mid \lambda) = \sum_{all\ \ s} \prod_{t=1}^{T-1} r_{s_{t-1}s_t} b_{s_t}(x_t), \tag{2}$$

where *t = 1, … T* are samples.

In mathematical terms, the learning part of a collaborative learning agent system refers to finding the model parameter λ to maximize the likelihood $f(x \mid \lambda)$. It is difficult to maximize $f(x \mid \lambda)$ directly due to the interlocking of the parameters of all the agents. By introducing a *Q*-function (e.g., Kullbuk-Leibler statistic method), this problem can be transformed into two relatively simple problems, rather than maximizing $f(x \mid \lambda)$ directly.

Let $Q(\lambda, \overline{\lambda})$ be a utility function when

$$Q(\lambda, \overline{\lambda}) \geq Q(\lambda, \lambda) \Rightarrow f(x \mid \overline{\lambda}) \geq f(x \mid \lambda) \tag{3}$$

$$Q(\lambda, \overline{\lambda}) = \sum_{o \in n^T} \frac{f(x, o \mid \lambda)}{f(x \mid \lambda)} \log f(x, o \mid \overline{\lambda}) \tag{4}$$

.

Maximizing $Q(\lambda, \overline{\lambda})$ with respect to $\overline{\lambda}$, referred to as the expectation step of the EM methodology, leads to the maximization of $f(x, \lambda)$.The maximization step of the EM methodology involves looking for a joint set of states, in this case, a set of expertise types that maximize the joint likelihood function, which can be viewed as the total fitness for a multi-agent system. The overall fitness *R(t,j)* is the total fitness of a multi-agent system up to time *t* if the ending expertise is *j*. The overall fitness function can be computed recursively, as shown in Figure 14. The overall fitness comes from a combination of the accumulative authority from the past *R(t-1, i)* and individual expertise at time *t* (i.e., $b_j[x_t]$).

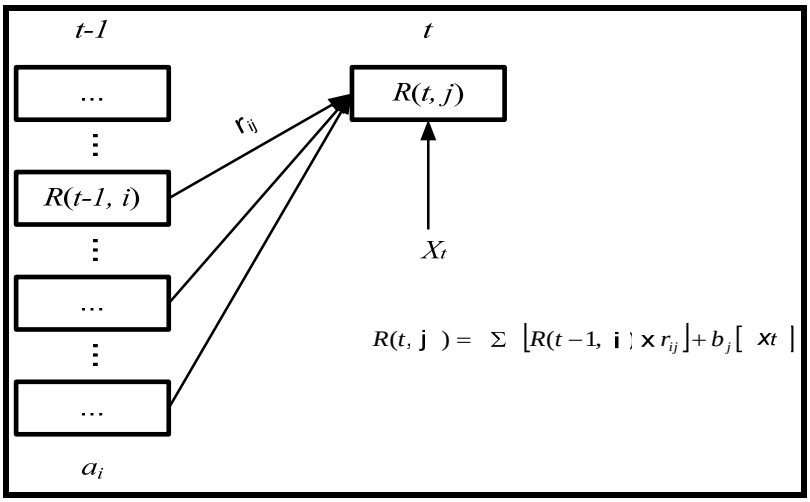

Figure 14. Recursion to Compute the Overall Fitness of a System *R*(*t*, *j*)

It is evident that traditional swarm intelligence system or PageRank-like algorithms only consider the accumulative authority part of the recursion. We introduce the expertise or competitiveness part of the recursion as the total fitness of a collaborative learning system.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A. Overview of the Lexical Link Analysis (LLA) Method

As in military operations, where the term *situational awareness* was coined, we note that that our efforts can inform *awareness* of analyzed data, in a unique way that helps improve decision-makers' understanding or awareness of the data content. We therefore define *awareness* as the cognitive interface between decision-makers and a complex system, expressed in a range of terms or "features," or a specific vocabulary or "lexicon," to describe the attributes and surrounding environment of the system. Specifically, LLA is a form of text mining in which word meanings represented in lexical terms (e.g., word pairs) can be represented as if they are in a community of a word network. Link analysis "discovers" and displays a network of word pairs. These word-pair networks are characterized by one-, two-, or three-word themes. The weight of each theme is determined by its frequency of occurrence.

Figure 15 shows a visualization of lexical links for Systems 1 and 2 of two systems, which are shown in the red box. Unlinked, outer vectors (outside the red box) indicate unique system features.

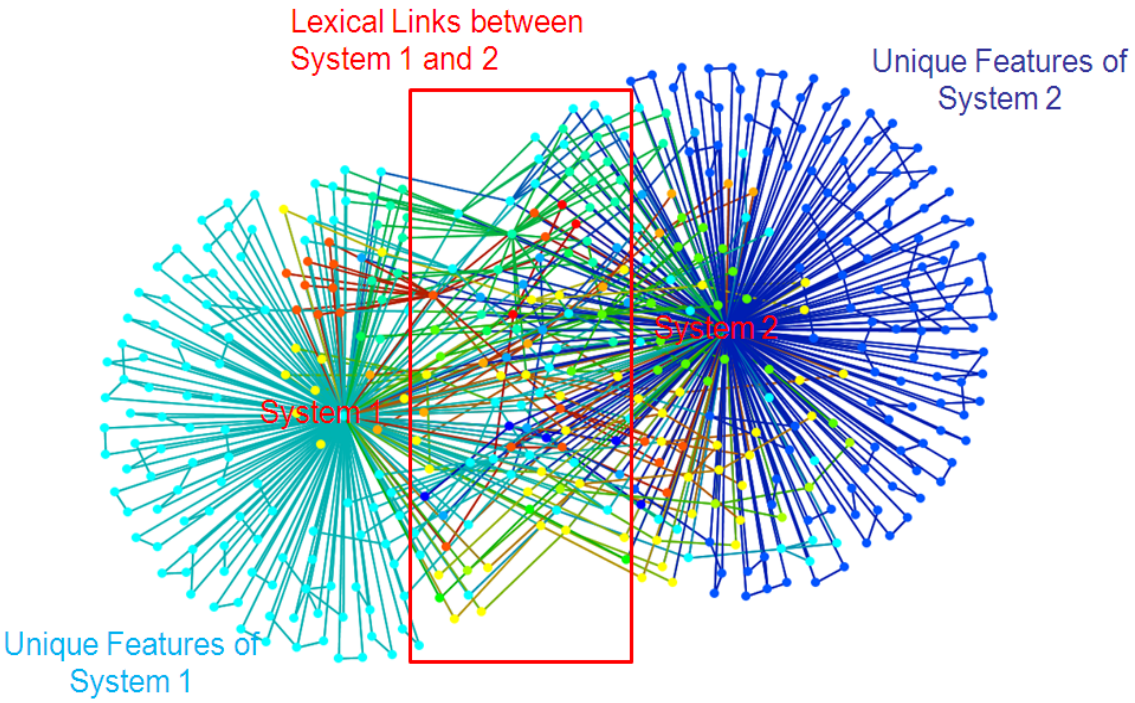

Figure 15. Comparing Two Systems Using LLA

Figure 16 shows how the information from three categories can be compared.

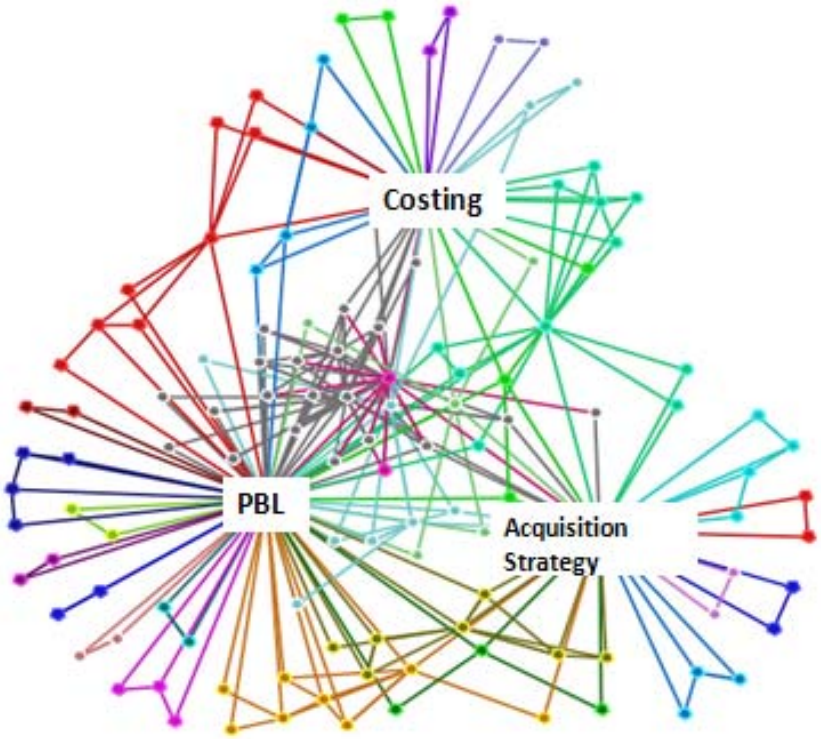

Figure 16. Comparing Three Systems Using LLA

Figure 17 shows how the information from two time periods can be compared.

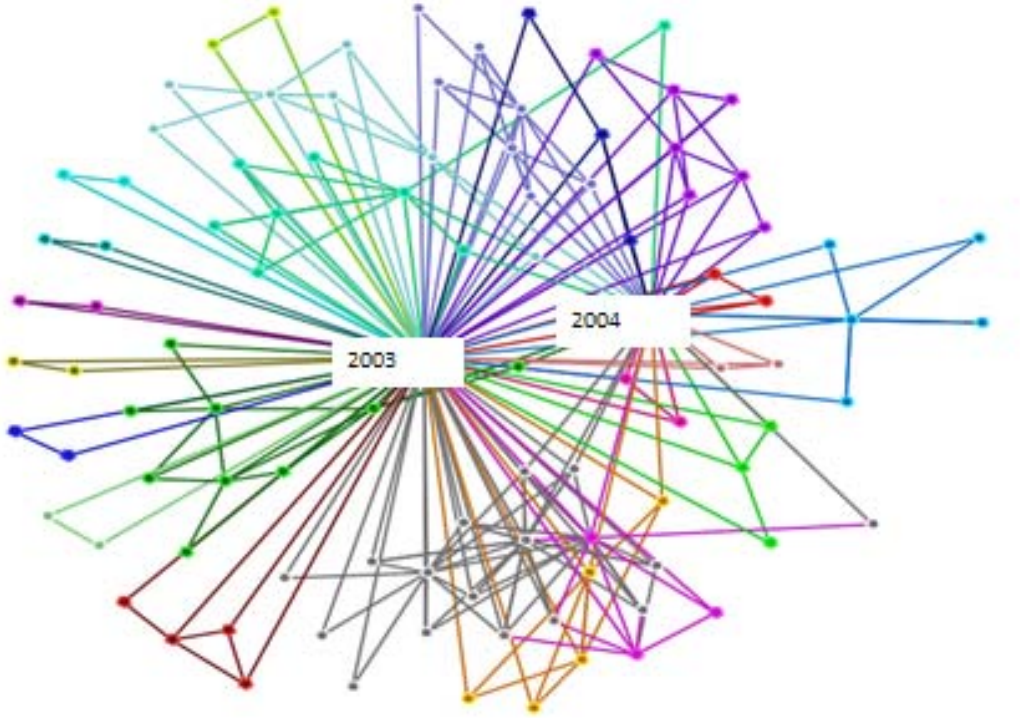

Figure 17. Comparing Two Time Periods

The closeness of the systems in comparison can be visually examined or quantitatively examined using the Quadratic Assignment Procedure (QAP; Hubert & Schultz, 1976; e.g., in UCINET, Borgatti, Everett, & Freeman, 2002) to compute the correlation and analyze the structural differences in the two systems as shown in Figure 18. Figure 19 shows word and term themes discovered and shown in colored groups.

QAP Correlations

		1 lla_n	2 lla_n	3 lla_n	4 lla_n	5 lla_n	6 lla_n	7 lla_n	8 lla_n
1	lla_network_1_2010-AcquisitionStrategy	1.000	0.174	0.156	0.155	0.036	0.111	0.020	0.062
2	lla_network_1_2003-AcquisitionStrategy	0.174	1.000	0.447	0.149	0.052	0.119	0.043	0.089
3	lla_network_1_2004-AcquisitionStrategy	0.156	0.447	1.000	0.111	0.047	0.119	0.051	0.080
4	lla_network_1_2005-AcquisitionStrategy	0.155	0.149	0.111	1.000	0.156	0.084	0.034	0.088
5	lla_network_1_2006-AcquisitionStrategy	0.036	0.052	0.047	0.156	1.000	0.067	0.036	0.056
6	lla_network_1_2007-AcquisitionStrategy	0.111	0.119	0.119	0.084	0.067	1.000	0.097	0.123
7	lla_network_1_2008-AcquisitionStrategy	0.020	0.043	0.051	0.034	0.036	0.097	1.000	0.286
8	lla_network_1_2009-AcquisitionStrategy	0.062	0.089	0.080	0.088	0.056	0.123	0.286	1.000

QAP P-Values

		1 lla_n	2 lla_n	3 lla_n	4 lla_n	5 lla_n	6 lla_n	7 lla_n	8 lla_n
1	lla_network_1_2010-AcquisitionStrategy	0.000	0.020	0.020	0.020	0.020	0.020	0.020	0.020
2	lla_network_1_2003-AcquisitionStrategy	0.020	0.000	0.020	0.020	0.020	0.020	0.020	0.020
3	lla_network_1_2004-AcquisitionStrategy	0.020	0.020	0.000	0.020	0.020	0.020	0.020	0.020
4	lla_network_1_2005-AcquisitionStrategy	0.020	0.020	0.020	0.000	0.020	0.020	0.020	0.020
5	lla_network_1_2006-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.000	0.020	0.020	0.020
6	lla_network_1_2007-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.020	0.000	0.020	0.020
7	lla_network_1_2008-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.020	0.020	0.000	0.020
8	lla_network_1_2009-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.000

QAP statistics saved as datafile QAP Correlation Results

Figure 18. QAP Correlation via UCINET

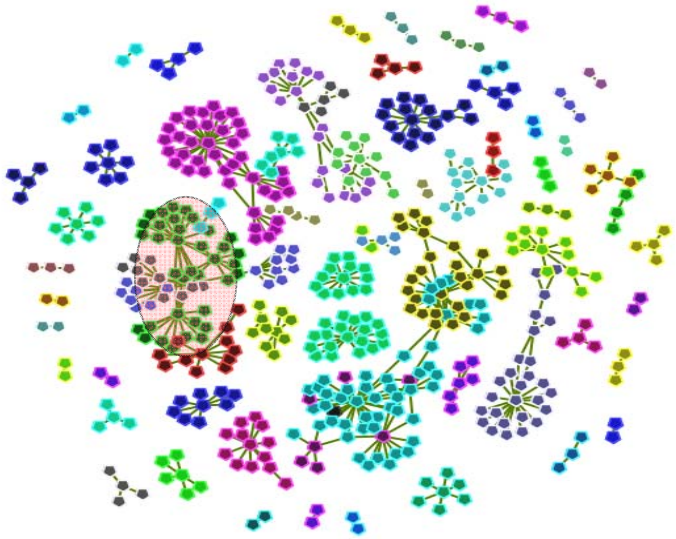

Figure 19. Word and Term Themes Discovered and Shown in Colored Groups

The detailed steps of LLA processing include applying collaborative learning agents (CLA) and generating visualizations, including a lexical network visualization via AutoMap (Center for Computational Analysis of Social and Organizational Systems [CASOS], 2009), radar visualization, and matrix visualization (Zhao et al., 2010). The following are the steps for performing an LLA:

- Read each set of documents.
- Select feature-like word pairs.
- Apply a social network community-finding algorithm (e.g., Newman grouping method (Girvan & Newman, 2002) to group the word pairs into themes. A theme includes a collection of lexical word pairs connected with each other.
- Compute a "weight" for a theme for the information of a time period; that is, how many word pairs belong to a theme for that time period and for all the time periods?
- Sort theme weights by time, and study the distributions of the themes by time.

A. Two Steps

Figure 20 and Figure 21 illustrates two steps (iterations) to discover themes as follows:

1st Iteration (Figure 20): Compute word pair clusters using Newman's community finding algorithm—words grouped as in a community (Girvan & Newman, 2002).

2nd Iteration (Figure 21): Select lexical terms linked to the most *central* nodes.

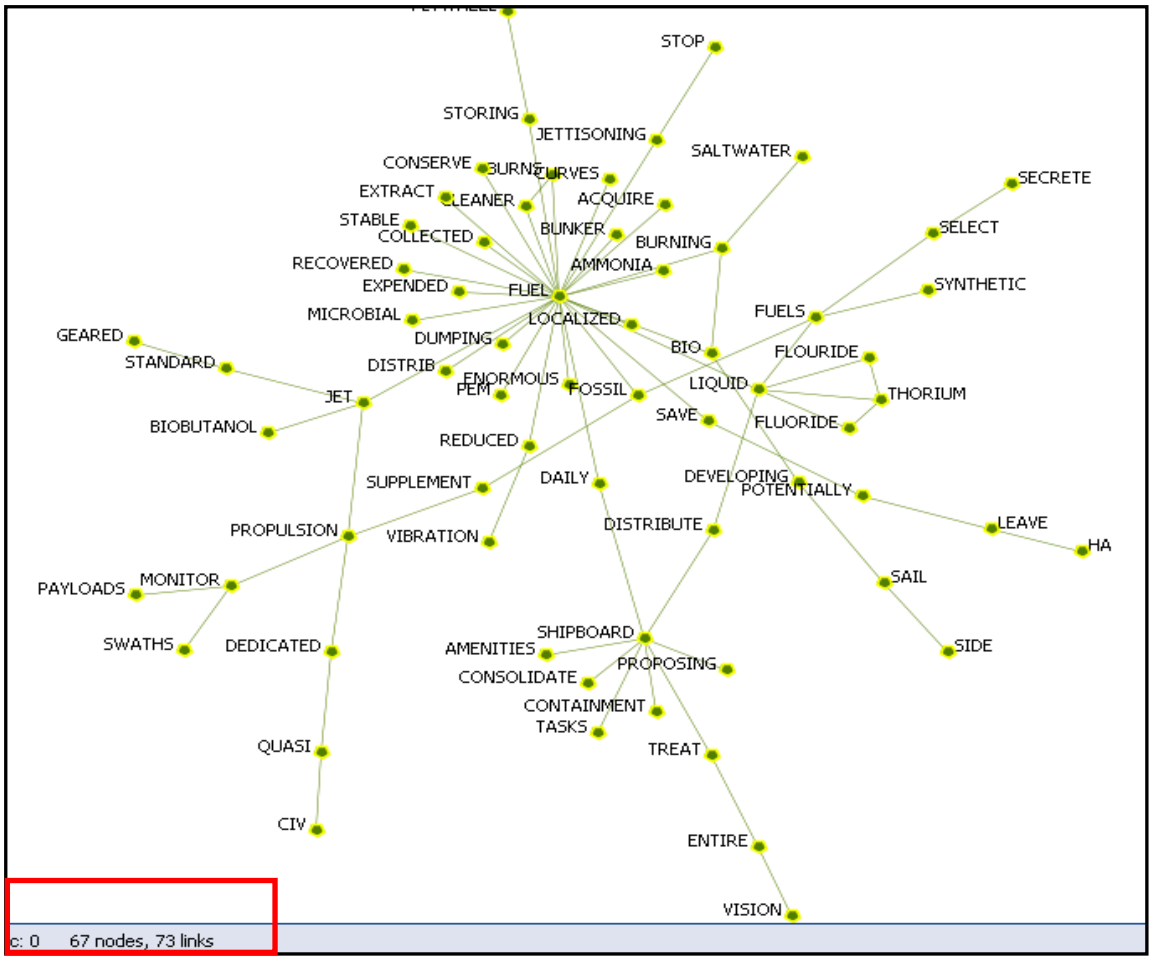

Figure 20. Initial Groups of Word Pairs

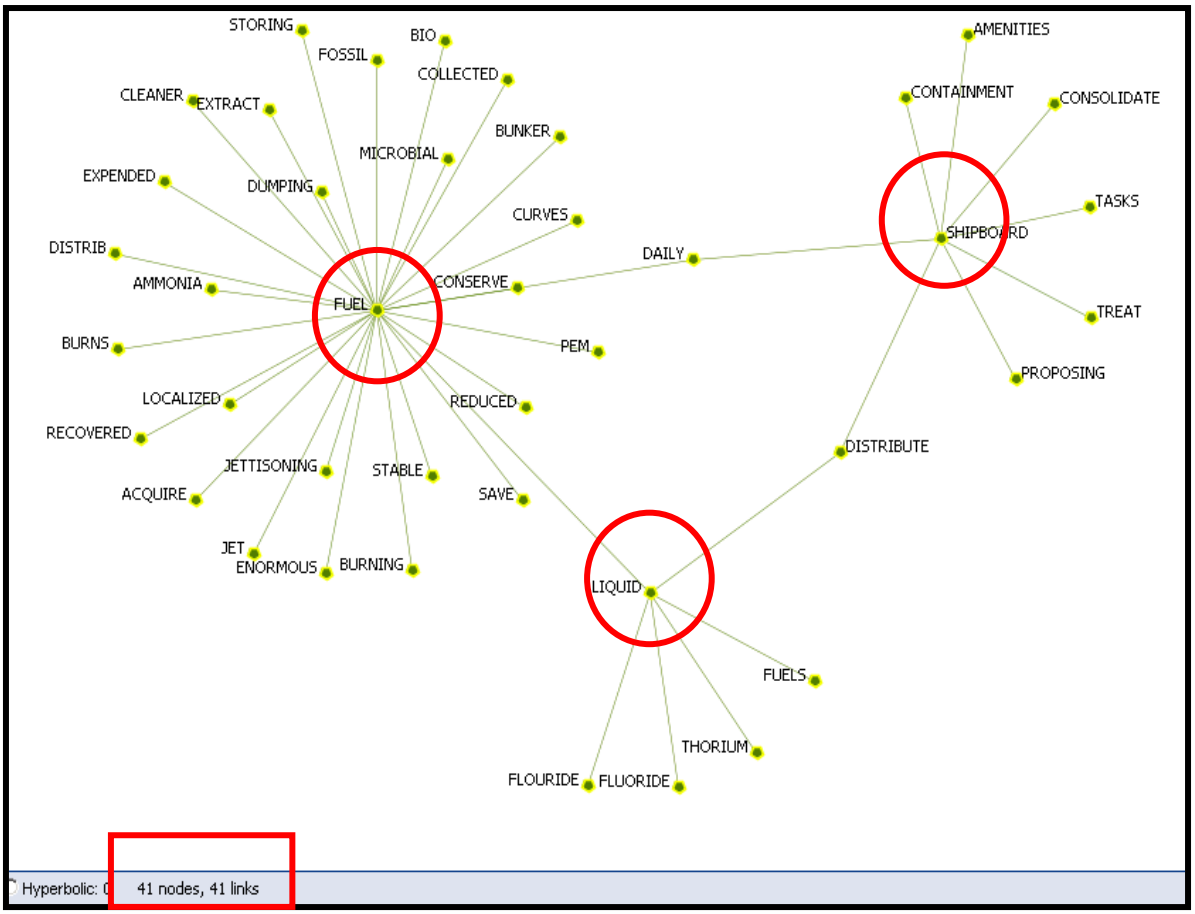

Figure 21. A Theme Includes a Collection of Word Pairs Selected Using Degree Centrality

B. Word Pair Selection Details

Figure 22 shows the computation and selection of word pairs in detail. In the CLA tool, the properties administration page allows various parameters to be set for LLA. For example,

- The "minimum frequency" specifies the minimum frequency a word can remain in the analysis. Words that appear less often than the minimum frequency are filtered out from the analysis.
- The "probability cut" specifies the cutoff probability for a word pair, which is defined as the probability of a word given a context. A context is the word in a word pair with a minimum frequency. For example, "Global Survivability" uses "Survivability" as the context. Since

"Survivability" appears only once (frequency = 1), it is skipped and, therefore, the whole word pair is not used. "Seek Strategies" and "Reduce Necessity" remain as detected word pairs for the following two conditions:

- Prob(Word|Context) > "Probability Cut"
- Frequency(Word) or Frequency(Context) > "Minimum Frequency"

"Seek" and "Reduce" are contexts. "Strategies" and "Necessity" are words that are associated with the contexts. The context is always placed first in a word pair.

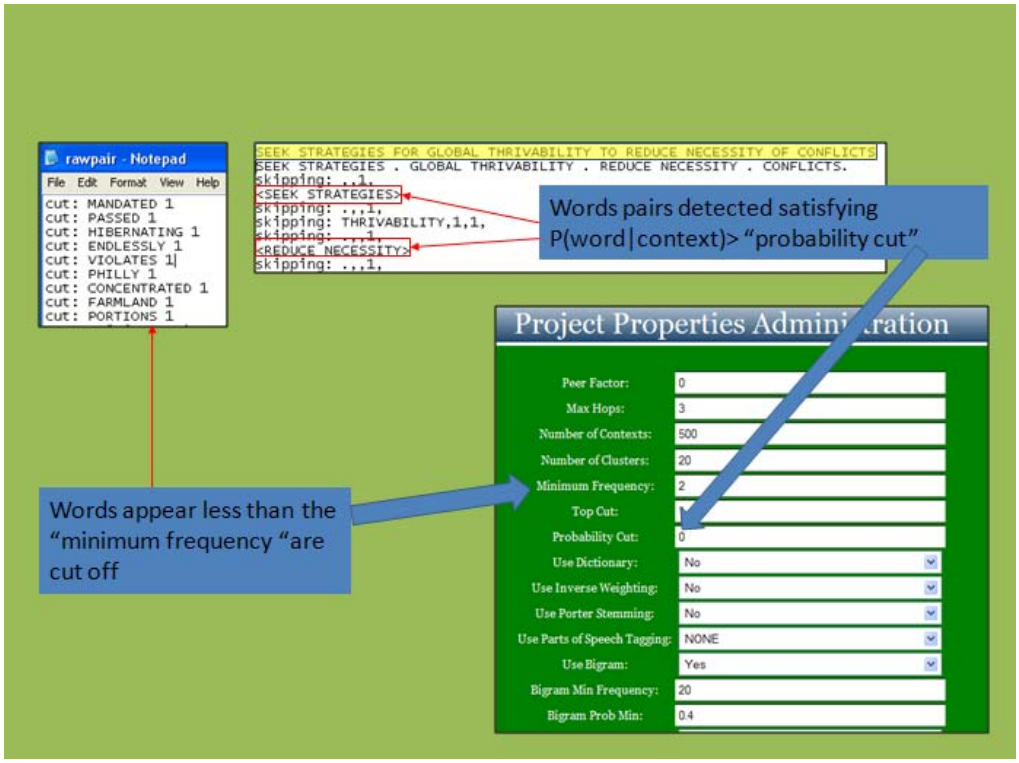

Figure 22. Computation and Selection of Word Pairs in Detail

C. Business Problems That LLA Can Address

General inquiries that LLA usually answers are as follows:

- Discover themes and topics in unstructured documents and sort the importance of the themes;
- Discover social and semantic networks of organizations that are involved and compare the two networks to obtain insights to answer the following questions—
    - identification of the organizations involved in the important themes and
    - comparison of the potential collaboration using semantic networks versus social networks.

D. Social and Semantic Network Analysis

Current research on social network analysis mostly focuses on people or organizations regardless of the contents linked. The so-called study of centrality (Girvan & Newman, 2002; Feldman, 2007) has been a focal point for the study of social network structures. Finding the *centrality* of a network lends insight into the various roles and groupings, such as the connectors (e.g., mavens, leaders, bridges, isolated nodes), the clusters (and who is in them), the network core, and the periphery. We have been working toward three areas of innovations in network analysis:

- Extract social networks based on entity extraction;
- Extract semantic networks based on the contents and word pairs using LLA;
- Apply characteristics and centrality measures from semantic networks and social networks to predict latent properties, such as emerging leadership that might dominate in the future in the social networks. The characteristics are further categorized into themes and time-lined trends for prediction of future events.

E. Implementation Details

In the past few years, we began at the Naval Postgraduate School (NPS) by using Collaborative Learning Agents (CLA; QI, 2009) and expanded to other tools, including AutoMap (CASOS, 2009) for improved visualizations. We also set up a

cluster utilizing Linux servers in the NPS High Performance Computing Center (HPC) to handle the large-scale unclassified data and a secure environment in the NPS Secure Technology Battle Laboratory (STBL). We are also in the process of investigating new big data technologies such as HDFS (Hadoop Distributed File System) and MapReduce as alternatives to HPC for parallel processing for the acquisition research web service. We also developed 3-D network views using Pajek (Pajek, 2008) and X3D (X3D, 2011). We also developed our visualizations' Radar view and Match view (Zhao et al., 2010). We are investigating how to benefit from using infovis, a web visualization tool in JavaScript to replace ORA that is currently used to visualize LLA results.

F. Relation to Other Methods

The LLA approach is more properly related to Latent Semantic Analysis (LSA; Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988) and Probabilistic Latent Semantic Analysis (PLSA). In the LSA approach, a term-document matrix is the starting point for analysis. The elements of the term-document or feature-object (term as feature and document as object) matrix are the occurrences of each word in a particular document (i.e., $A = [a_{ij}]$, where a_{ij} denotes the frequency in which term *j* occurs in document *i*). The term-document matrix is usually sparse. LSA uses singular value decomposition (SVD) to reduce the dimensionality of the term-document matrix. SVD cannot be applied to the cases where the vocabulary (the unique number of terms) in the document collection is large. LSA has been widely used to improve information indexing, search/retrieval, and text categorization.

A recent development related to this method is called Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003), which is a generative probabilistic model of a corpus. In LDA, a document is considered to be composed of a collection of words—a "bag of words," in which word order and grammar are not considered important. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a statistical distribution

(Dirichlet distribution) over the corpus. Our theme generation from LLA is different from LDA, in which a collection of lexical terms are connected as text-as-networks instead of a bag-of-words. Our method is easily scaled to analyze a large vocabulary and is generalizable to any sequential data.

G. Anticipated Benefits

The method provides solutions to meet the critical needs of the acquisition research community. The key advantages is to provide an innovative near real-time self-awareness system to transfer diversified data services into strategic decision-making knowledge, detailed as follows:

- Automation: The high correlation of LLA results with the link analysis done by human analysts makes it possible for automation, saving human power and improving responsiveness. Automation is achieved via computer program or software agent(s) to perform LLA frequently—and in near real-time: Agent learning makes it possible to reach real-time; visualization corrects lexical links to core measures; features and patterns are discovered over time for the system as a whole. We can take advantage of the data in motion (Twitter and social media sites), such as RSS feed data to build a better picture of real-time program awareness.
- Discovery: It "discovers" and displays a network of word pairs. These word-pair networks are characterized by one-, two-, or three-word themes. The weight of each theme is determined based on its frequency of occurrence. It may also discover blind spots of human analysis that are caused by the overwhelming data for human analysts to go through.
- Validation: As we continue validating LLA by direct correlation with human analysts' results, new dimensions of using LLA to validate human analysis also show the advantages of our methodology. For instance, LLA may provide different perspectives on links. In the acquisition context, links discovered by human analysts may emphasize component/part connections, but they do not necessarily reflect on the content overlaps; therefore, interdependencies of the programs identified by human analysts (e.g., program managers) might help the programs to stay funded from year to year for the goal of building their importance, and not cost reduction for the government. LLA looks for overlapping content to improve affordability and to meet

the requirements of warfighters. Consequently, it provides better results in terms of trust, quality of association, discovery, and it serves to break through the taxonomy of ignorance (Denby & Gammack, 1999) and organizational boundaries, and to improve organizational reach.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for Windows: Software for social network analysis*. Harvard, MA: Analytic Technologies.

Burt, R. (1992). *Structural holes: The social structures of competition*. Cambridge, MA: Harvard University Press.

Carley, K. M. (2007). Improved data extraction and assessment for dynamic network analysis. In Gade, P. A. (Ed.), *U.S. Army Research Institute Program in Basic Research—FY 2010* (Special Report 69). Retrieved from http://dbaconsultinghq.com/wp-content/uploads/2011/11/Special-Report-69.pdf

Center for Computational Analysis of Social and Organizational Systems (CASOS). (2009). AutoMap: Extract, analyze and represent relational data from texts [Computer Software]. Retrieved from http://www.casos.cs.cmu.edu

Chairman of the Joint Chiefs of Staff (CJCS). (2009). *Chairman of the Joint Chiefs of Staff instruction for joint capabilities integration and development system (JCIDS)* (J-8 CJCSI 3170.01G). Retrieved from http://www.dtic.mil/cjcs_directives/cdata/unlimit/3170_01.pdf

Dashofy, E. M., van der Hoek, A., & Taylor, R. N. (2002). Towards architecture-based self-healing systems. In *Proceedings of the First Workshop on Self-Healing Systems* (pp. 21–26). Retrieved from http://www.antconcepts.com/~edashofy/files/dht-woss-2002.pdf

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B 39, 1–38.

Denby, E., & Gammack, J. (1999). *Modelling ignorance levels in knowledge-based decision support*. Retrieved from http://wawisr01.uwa.edu.au/1999/DenbyGammack.pdf

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing* (pp. 281–285). New York, NY: Association for Computing Machinery.

Feldman, R. (2007). *Link analysis and text mining: Current state of the art and applications for counter terrorism* [Online video]. Retrieved from http://videolectures.net/mmdss07_feldman_latm/

Foltz, P. W. (2002). Quantitative cognitive models of text and discourse processing. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *The handbook of discourse processes*. Mahwah, NJ: Lawrence Erlbaum.

Gallup, S. P., MacKinnon, D. J., Zhao, Y., Robey, J., & Odell, C. (2009, October 6–8). Facilitating decision making, re-use and collaboration: A knowledge management approach for system self-awareness. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management* (IC3K). Madeira, Portugal: INSTICC Press.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *USA*, *99*(12), 7821–7826.

Hinchey, M. G., & Sterritt, R. (2006). Self-managing software. *Computer*, *39*(2), 107–109.

Hubert, L. & Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychology*, *29*, 190–241.

Krackhardt, D. (1998). Simmelian tie: Super strong and sticky. In R. Kramer & M. Neale (Eds), *Power and influence in organizations* (pp. 21–38). Thousand Oaks, CA: Sage.

Letsche, T. A., & Berry, M. W. (1997). Large-scale information retrieval with latent semantic indexing. *Information Sciences*, *100*(1-4), 105–137.

Pajek (2008), Networks/Pajek: Program for large network analysis[computer software]. Retrieved from http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Quantum Intelligence (QI). (2009). Collaborative learning agents (CLA). Retrieved from http://www.quantumii.com/qi/cla.html

Reminga, J., & Carley, K. M. (2003). *Measures for ORA (Organizational Risk Analyzer)*. Retrieved from http://www.casos.cs.cmu.edu/publications/papers/reminga_2003_ora.pdf

X3D. (2011). Open standards for real-time 3D communication. Retrieved from http://www.web3d.org

Zhao, Y., Gallup, S., & MacKinnon, D. (2010). Towards real-time program awareness via Lexical Link Analysis. In *Proceedings of the Seventh Annual Acquisition Research Program*. Retrieved from http://www.acquisitionrearch.net

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011a). *Towards real-time program awareness via Lexical Link Analysis* (NPS-AM-10-174). Retrieved from Naval Postgraduate School, Acquisition Research Program website: http://www.acquisitionresearch.net

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011b, May). A web service implementation for large-scale automation, visualization and real-time program-awareness via Lexical Link Analysis. In *Proceedings of the Eighth Annual Acquisition Research Program*. Retrieved from http://www.acquisitionresearch.net

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011c, September). System self-awareness and related methods for improving the use and understanding of data within DoD. *Software Quality Professional*, *13*(4), 19–31. Retrieved from http://asq.org/pub/sqp/

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2012a, April). Applications of Lexical Link Analysis web service for large-scale automation, validation, discovery, visualization, and real-time program-awareness. In *Proceedings of the Ninth Annual Acquisition Research Program*. Retrieved from http://www.acquisitionresearch.net

Zhao, Y., MacKinnon, D., & Gallup, S. (2012b, June 11–15). *Lexical link analysis and system self-awareness: Theory and practice*. Poster session presented at the Cyber and Information Challenges 2012 Conference, Utica, NY.

Zhao, Y., MacKinnon, D., & Gallup, S. (2012c, June 19–21). Semantic and social networks comparison for the Haiti earthquake relief operations from APAN data sources using lexical link analysis. In *Proceedings of the 17th ICCRTS, International Command and Control, Research and Technology Symposium*. Retrieved from http://www.dodccrp.org/events/17th_iccrts_2012/post_conference/papers/082.pdf

THIS PAGE INTENTIONALLY LEFT BLANK

2003 - 2012 Sponsored Research Topics

Acquisition Management

- Acquiring Combat Capability via Public-Private Partnerships (PPPs)
- BCA: Contractor vs. Organic Growth
- Defense Industry Consolidation
- EU-US Defense Industrial Relationships
- Knowledge Value Added (KVA) + Real Options (RO) Applied to Shipyard Planning Processes
- Managing the Services Supply Chain
- MOSA Contracting Implications
- Portfolio Optimization via KVA + RO
- Private Military Sector
- Software Requirements for OA
- Spiral Development
- Strategy for Defense Acquisition Research
- The Software, Hardware Asset Reuse Enterprise (SHARE) repository

Contract Management

- Commodity Sourcing Strategies
- Contracting Government Procurement Functions
- Contractors in 21st-century Combat Zone
- Joint Contingency Contracting
- Model for Optimizing Contingency Contracting, Planning and Execution
- Navy Contract Writing Guide
- Past Performance in Source Selection
- Strategic Contingency Contracting
- Transforming DoD Contract Closeout
- USAF Energy Savings Performance Contracts
- USAF IT Commodity Council
- USMC Contingency Contracting

Financial Management

- Acquisitions via Leasing: MPS case
- Budget Scoring
- Budgeting for Capabilities-based Planning
- Capital Budgeting for the DoD
- Energy Saving Contracts/DoD Mobile Assets
- Financing DoD Budget via PPPs
- Lessons from Private Sector Capital Budgeting for DoD Acquisition Budgeting Reform
- PPPs and Government Financing
- ROI of Information Warfare Systems
- Special Termination Liability in MDAPs
- Strategic Sourcing
- Transaction Cost Economics (TCE) to Improve Cost Estimates

Human Resources

- Indefinite Reenlistment
- Individual Augmentation
- Learning Management Systems
- Moral Conduct Waivers and First-term Attrition
- Retention
- The Navy's Selective Reenlistment Bonus (SRB) Management System
- Tuition Assistance

Logistics Management

- Analysis of LAV Depot Maintenance
- Army LOG MOD
- ASDS Product Support Analysis
- Cold-chain Logistics
- Contractors Supporting Military Operations
- Diffusion/Variability on Vendor Performance Evaluation
- Evolutionary Acquisition
- Lean Six Sigma to Reduce Costs and Improve Readiness

- Naval Aviation Maintenance and Process Improvement (2)
- Optimizing CIWS Lifecycle Support (LCS)
- Outsourcing the Pearl Harbor MK-48 Intermediate Maintenance Activity
- Pallet Management System
- PBL (4)
- Privatization-NOSL/NAWCI
- RFID (6)
- Risk Analysis for Performance-based Logistics
- R-TOC AEGIS Microwave Power Tubes
- Sense-and-Respond Logistics Network
- Strategic Sourcing

Program Management

- Building Collaborative Capacity
- Business Process Reengineering (BPR) for LCS Mission Module Acquisition
- Collaborative IT Tools Leveraging Competence
- Contractor vs. Organic Support
- Knowledge, Responsibilities and Decision Rights in MDAPs
- KVA Applied to AEGIS and SSDS
- Managing the Service Supply Chain
- Measuring Uncertainty in Earned Value
- Organizational Modeling and Simulation
- Public-Private Partnership
- Terminating Your Own Program
- Utilizing Collaborative and Three-dimensional Imaging Technology

A complete listing and electronic copies of published research are available on our website: www.acquisitionresearch.net

THIS PAGE INTENTIONALLY LEFT BLANK

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CALIFORNIA 93943

www.acquisitionresearch.net