DARPA-BAA-11-64
SOCIAL MEDIA IN STRATEGIC COMMUNICATION (SMISC)
TECHNICAL AREA 1 (TA 1): ALGORITHM/SOFTWARE DEVELOPMENT

VOLUME 1: Technical and Management Proposal (includes Appendix A)

PROPOSAL TITLE: Memetics and Media Enhanced Tracking System (METSYS)

LEAD ORGANIZATION: Center for Advanced Defense Studies, Inc.

TYPE OF BUSINESS: Other Nonprofit

TEAM MEMBERS:
Psydex Corporation (Other Small Business)
Behavioral Media Networks (Other Small Business)
Boston University (Other Educational)

POINTS OF CONTACT:

TECHNICAL	*ADMINISTRATIVE*
LTC (Ret.) David E.A. Johnson, USA 10 G St NE, STE 610 Washington, DC 20002 P: 202-289-3332 F: 202-789-2786 David.johnson@c4ads.org	Mr. Farley Mesko 10 G St NE, STE 610 Washington, DC 20002 P: 202-289-3332 F: 202-789-2786 Farley.mesko@c4ads.org

PROPOSAL SUBMITTED ON: 12 SEP 2011

PROPOSAL VALIDITY PERIOD: 120 days

DUNS: 192561012

TIN: 73-1681366

CAGE: 4MT61

AWARD INSTRUMENT REQUESTED: Grant

COST OF PROJECT: $11,262,992

PLACES AND PERIODS OF PERFORMANCE:

Center for Advanced Defense Studies, Inc.
Washington, DC (15 DEC 11-15 JAN 15)
Cambridge, MA (15 DEC 11-15 JAN 15)
Chicago, IL (15 DEC 11-15 JAN 15)

Psydex Corporation
Atlanta, GA (15 DEC 11-15 JAN 15)

Behavioral Media Networks
Cambridge, MA (15 DEC 11-15 JAN 15)

Boston University
Boston, MA (15 DEC 11-15 JAN 15)

Contents

1. Executive Summary

Social media has rapidly emerged as the primary medium for communication and dissemination of news and information, and events of strategic and tactical importance to our Armed Forces and civilian agencies are increasingly taking place in cyberspace. Disparate events such as the Arab Spring, the recent London riots, and BART flash mobs are linked by the significance of social media in organizing and inciting violence or disorder within large public groups. Armed Forces, economic regulatory agencies, and federal, state and local law enforcement would all benefit from, but currently lack, the ability to quickly and effectively detect social media patterns and to prevent the adverse impact of those patterns.

Humans communicate in an estimated 6,800 languages and an ever expanding list of semantic variations to compress more thoughts into smaller messages using shorthand (e.g. emoticons, slang, emerging or lacking grammatical rules). Analyzing patterns in social media and social networks can be a powerful predictor of human behavior and intent. Doing this at speed and scale is an enormous challenge because of the volume, velocity and variety of the social media data generated today.

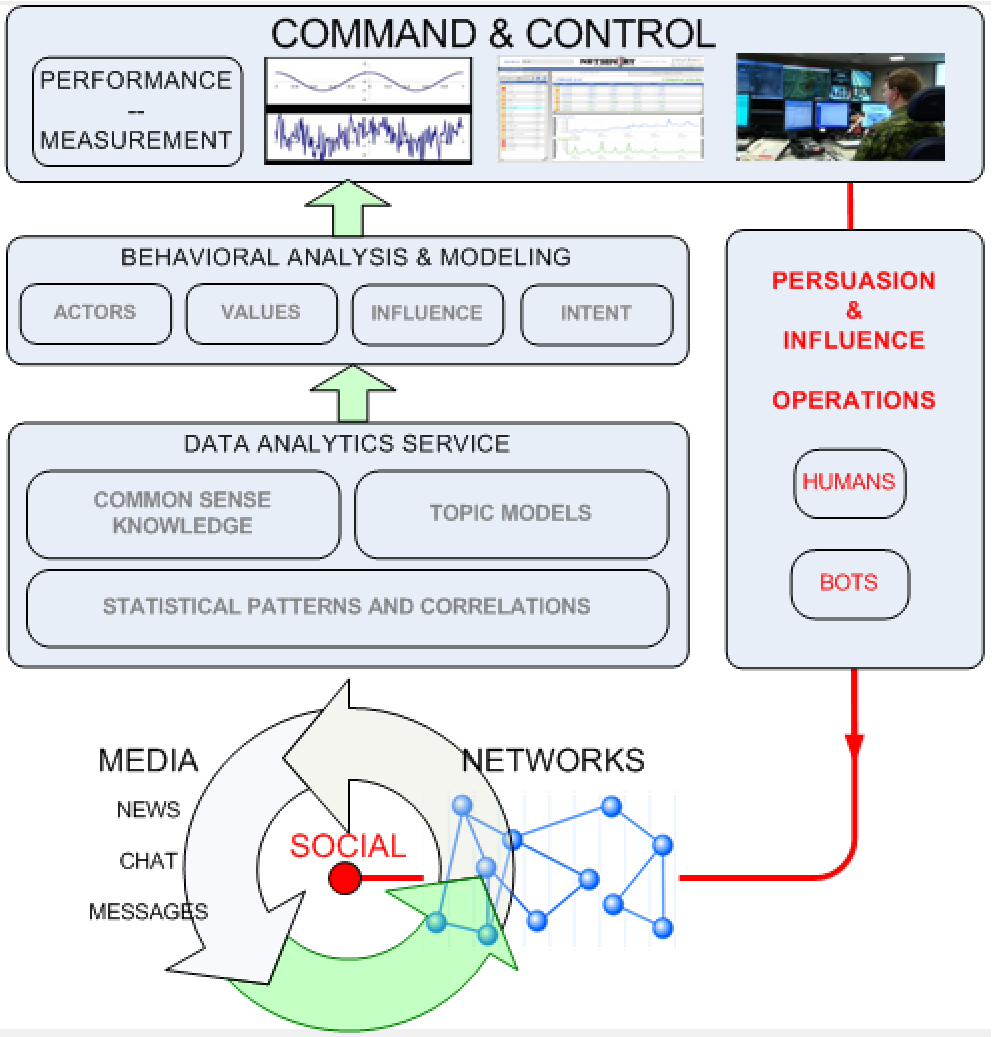
Our goal is to develop a Memetics and Media Enhanced Tracking System ("METSYS") that will enable effective and timely analysis of emerging memes, influence operations, and persuasion campaigns in social media, via systematic detection and classification of actors, values, influence, and intent correlated to evolution of social network structure and interaction patterns over time. Although the scope of the current project is a test comprising 5,000 nodes, our proposed approach is capable in principle of performing analysis of virtually all the world's social media.

METSYS will be designed and developed with massive scalability in mind that will enable humans and algorithms to mathematically model and project human behavior through the analysis of changes in social network structure and semantic patterns. METSYS will leverage massively parallel processing (MPP), distributed computing and specialized temporal indexes and algorithms to facilitate real time integrated analysis of graph structures and conversations over time and space.

While analyzing static graph structures is a fairly well-understood problem, social media conversations, by comparison, are very difficult to analyze because humans do not adhere to a simple data model when communicating. Language is a constantly evolving system, and the science for mapping it to cognition is underdeveloped. Unlike conventional approaches that use standard statistical linguistic methods, our team will take a novel, domain-independent approach that makes no prior assumptions about the language or format of social media conversations. Our approach also incorporates mainstream media to correlate actions and conversations with real-world events and social behavior.

In contrast, current approaches to temporal analysis, as done by some social media sites and monitoring services, generally use trending volume of simple keywords, pre-defined topics, or tags (e.g. Twitter hash tags). Separately, other services (e.g. Klout) analyze relationships between social networkers (e.g. Followers) to determine influence based on relationships and explicit actions taken by fellow social networkers (e.g. re-tweet, like, comment). Only primitive semantic analysis is ever performed to analyze content of messages, and it is rare to see such analysis in the context of network evolution or with a focus on intent and behavior. Moreover, no such work that we are aware of includes cognitive studies to confirm or deny findings. Historical context will be integral to our approach, making it possible to better understand how revolutions / uprisings and changes in societal behavior are impacted by social media and vice-versa.

We propose a three-year project, cost-feasible within budget estimates provided at $11,262,992. The diagram below shows the basic components and flow of METSYS. Each component of this diagram is explained and detailed technical activities required to accomplish this are provided in the Goals and Impact and Technical Plan sections of this document.



The key components of our approach are as follows:

- *We will dimensionalize discourse in real time based on relative and temporal proximity of characters, words and phrase patterns to identify topic models.*

- *Topic models and relationships among them will be mapped to human concepts based on commonsense knowledge and rules from databases such as Open Mind Common Sense (OMCS) developed at MIT.*
- *Topic models will be decoupled from indexes, executed continuously, thus requiring no prior knowledge of sources, tagging, re-indexing or reorganization of source data.*
- *This approach will enable high-performance modeling and learning of relationships between discourse and cognition. By graph analytics and mathematical comparative analysis we will be able to better understand emergence and changes in topic model activity, as depicted in the diagram above.*
- *This will be applied to a culturally agnostic persuasion model that will allow adaptation to target cultural environments.*
- *We will incorporate semi-automated human analysis and verification testing to validate the standalone automated system.*

2. Goals and Impact

Our proposed goal for this effort is to research, design, and develop a comprehensive system, or set of integrated system architectures and algorithms, that will allow the identification and analysis of emerging and targeted social media driven topics (or memes). In order to achieve these objectives it is important to reflect on current practices in identifying these social memes and the extent that these emerging topics are currently analyzed, tracked, and used in social trends and projected social behaviors.

Most readily-understood current methods for identifying emerging social topics, or memes, is performed through Bayesian and Boolean stochastic algorithms which identify emergence based on summative formulas of query statements and material presentations. While this is certainly useful in “taking the pulse” of the zeitgeist of social media activity, it fails to allow for second, third, fourth and so on levels of analysis of these topics of interests. Further, these aggregative and summative methodologies largely lack formal architectural frameworks and methods for in-depth analysis of these memes for such things as durability, lifecycle, attributionality, and social network nodal and nodal communities.

As called for in the BAA, our intent is to directly address these research questions in a fundamental and systematic process that enables meme detection and tracking together with social-network analysis for nodal communities, including their emergence. We recognize the need for this fundamental architectural framework for accomplishing effective analysis activities. This will include information pattern flows, trend analysis, trust and intent analysis, as noted in the BAA, and will be accomplished to achieve layered and in-depth analysis of these memes to assess such things as effects of persuasion and influence campaigns and to produce projections of potential behavioral intents and outcomes. For this to be truly successful this system must be designed with the ability to perform narrative, linguistic structure, linguistic cues, and cultural analyses that cross all existing forms of communication.

2.1 Overall Goal

METSYS and sub-system element Project Goals will enable:

1) Complex semantic patterns in communications to be analyzed over time in near real time
2) Changes in Network Structure to be identified over time in near real time

Based on a series of algorithms that make use of a set of analytic capabilities, we will:

- Detect, measure and track emergent memes, through the use of topic models
- Create and continuously update node and participant profiles, based on historical patterns of discourse and network structure
- Track discrete networks influenced by each node
- Detect changes in social network structure
- Analyze the intent of nodes and node groups

METSYS will discover and use powerful rules for social media group dynamics and intent formation, leveraging emerging technologies to manage the massive increases in volume, velocity, and variety of data. If successful, the METSYS could improve public safety savings up to $200 million dollars (London Riots) per incident, help the Military to counter significant emerging threats like sophisticated insurgencies in built-up areas, extremist narratives used to recruit lone wolf terrorists like Major Hassan, and State-led judo warfare (4th Generation Warfare). The project will provide a semi-automated method to enable decision0makers to observe, orient, decide, and act on the social media domain within a threat decision cycle to:

1) Detect, classify, measure and track the formation, development and spread of ideas and memes, and purposeful deceptive messaging and misinformation
2) Recognize persuasion campaign structures and influence operations across social media sites and communities
3) Identify participants and intent, and measure effects of persuasion campaigns
4) Counter messaging of detected adversary influence operations

This project approach will uniquely link emerging approaches to data management, social network analysis, cognitive informatics, cognitive linguistics, behavioral modeling, and Strategic Communications to create a learning system that is high-performance, massively-scalable, and symbolically/linguistically and culturally agnostic.

Initial slide presentations will detail the state of the art for the services layer, the analysis layer, and the command and control layer with research directions. A monthly code-drop will enable design testing and replication. As described in the announcement, the team will submit other regularly required reports, as well as technical papers and reports as defined in the Statement of Work.

This project will both build upon the existing research led by the team members and define directions for future research. The Center for Advanced Defense Studies team currently leads projects in Culture and Conflict Studies, Strategic Understanding, and Computational Approaches to National Security. This research will contribute to the objective of linking the qualitative and quantitative programs to find measures of effectiveness and power rules for social programs in the cognitive space. The project should result in advances in the understanding of intent formation and intention awareness that lead to creation of Behavioral Media Networks (BMN) and enhance efforts of the MIT Mind Machine Project (MMP). Finally, enhancements in data management and analytics across volume, variety, and velocity will refine Psydex products being developed for the Financial Services Industry and the US Intelligence Community.

The team proposes to design and develop METSYS, a massively scalable system that will enable humans and algorithms to mathematically model human behavior and provide probabilistic predictions of intent through the analysis of changes in social network structure and the analysis of semantic patterns in communications over time. METSYS will leverage massively parallel processing (MPP), distributed computing and highly-specialized temporal indexes and algorithms to facilitate real time analysis of graph structures and conversations over time. The system will run on commodity hardware, primarily making use of Open Source software and Commercial-Off-the-Shelf (COTS) software where required.

The following set of specific system goals and impact components proposed in this section are provided to describe specific research areas proposed and to provide amplification and clarification of our proposed project. Each of these identified goals is specifically explicated and actions proposed for meeting these goals are provided in the next section, the Technical Plan proposal. We have identified broad goal categories of our proposed effort. Under each of these broad project goals, we have identified several component activity areas with discrete goals for each component. In addition, for each component goal identified we have purposefully included '*Why'* each of these component goals are seen as critical to our project. A summary is provided for each set of component goals that identifies how each component set of activities (taken together) provide the unique and innovative contribution to the success of this effort and the subsequent impacts to be garnered by this effort in achieving the objectives stated in the BAA. Finally, in each of the component summaries we identify expected deliverables that will be produced for these component activities.

2.2 Technical and Architectural Goals

2.2.1 System Design

A. Support concrete and abstract topic models. *Why?* METSYS must be capable of representing concrete entities, such as people, places, events, and things, as well as abstract concepts, such as fear, anger, and empathy. Topic models represent the thoughts, ideas and concepts which form the basis for knowledge and cognition, and must be robust and capable of representing abstract concepts and real world entities.

B. Flexible semantic expressions. *Why?* The complexity of a topic model varies, and our semantic expressions must be capable of representing all topic models along this spectrum. There are often thousands of semantic variations for a single topic model. These variations are typically based on context and the various ways a topic is referenced or expressed in social media and discourse in general.

C. Language-agnostic. *Why?* Topic models must represent the various ways a topic is expressed in different natural languages. As well, social media encourages the creation of new tags, emoticons and other ways of indicating emotions, subject matter, etc.

D. Topic models must be dynamic and current. *Why?* Topics models will be learned and evolved by both machines and humans. We must be able to create or modify a topic model and have the system provide an accurate measure of the topic model for both the present as well as the past.

E. Topic model statistics, time series and correlations must be generated in near real time. *Why?* METSYS must continuously monitor and measure existing topic models in order to determine when memes emerge and fade based on how unusual a topic model is relative to the past. METSYS must provide an accurate measure of all topic models all the time.
F. Topic model time series and statistics incorporated into rules processing. *Why?* Incorporating temporal and spatial dimensions into understanding conversations is similar to the way this is accomplished by human brains.

To accomplish the objective of identifying and measuring memes we must first develop a systematic analytical methodology and common framework for analyzing social media, identified in our approach as the topic model. The unique aspect and innovation proposed is recognizing the need for an integrated set of identified goal activities (described in detail in the Technical Plan) for this analytical topic model framework comprised of the set of items A through F above. Associated deliverables for integration of these topic model elements will be provided under this proposed effort to include source code, algorithms, performance measurements, etc.

2.2.2 Performance and Scale

A. Linear & horizontal scale. *Why?* Must be able to adapt to a changing operating environment by simply adding more servers or virtual machines (VM).
B. Near-real-time performance. *Why?* Reduce time to discovery of emerging threats, achieve near real time situational awareness, quickly back-test hypothesis.
C. Massively parallel processing (MPP) based. *Why?* Reduce processing / execution time by running in parallel across many servers. "Divide and Conquer".
D. Commodity hardware. *Why?* Less risk since not dependent on specialized hardware, technology, or company.
E. We will identify the scalability limits of software and algorithms based on general purpose processors (GPP). We will identify areas where the METSYS may benefit from specialized hardware / systems (e.g. FPGA, ASICS and RALA). *Why?* To improve performance at scale.

To accomplish the objective of continuously monitoring, measuring and reasoning over complex topic models in near real time, we must first develop an infrastructure that is capable of elastic scalability as well as high-performance computing. The unique nature of our system is in the algorithms and processes that utilize this infrastructure, as well as the near real time performance of processes which are typically run as batch jobs on large clusters.

2.2.3 Flexible-Modular Design Elements

A. Loose coupling of components. *Why?* Makes it easy to adapt to new requirements
B. Domain knowledge and topic models will be separate from the organization of the data (Indexing). *Why?* Easy adaption to new knowledge and discoveries without impacting indexing or collection of data. As well, fewer constraints and restrictions on how data is collected and pre-processed for indexing reduces the time and complexity to bring on new sources

C. Rapid integration of new, diverse sources. *Why?* Fusing many sources is key to measuring topics in a broad context. The ability to quickly react and deploy new sources in short order will be key to supporting many of the targeted missions.

D. Standards-based APIs. *Why?* Build applications and clients with no lock-in to a particular language, vendor or technology base. We will support the prevailing methods and approaches used by most developers, such as REST and TCP/IP Sockets.

E. Custom queries and execution parameters. *Why?* Give users and systems, total control over what they can query and how it gets executed. We will be able to support back-testing and hypothesis analysis, with customize weighting of individual sources, topics, etc.

To accomplish the objective of being adaptive to the mission and emerging threats, METSYS must not be a rigid structure. It must be capable of morphing as the threats evolve over time. By enabling users and systems to customize the inputs, outputs and behavior of the METSYS, we will be capable of adjusting or repurposing quickly.

2.2.4 Indexes and Algorithms

A. Streaming real time and historical information seamlessly integrated and fused. *Why?* Most systems today deal with either streaming real time (e.g. Yahoo S4) or batch processing of historical information (e.g. Apache Hadoop). These systems were initially developed without the requirement to effectively support both real time and historical processing. We will integrate streaming real-time (new) information with historical information on a continuous basis, thus providing all information in a seamless, fused manner.

B. Conversations (unstructured / semi-Structured) fused with graph structures. *Why?* Most systems today deal with either unstructured/semi-structured (e.g. Google, Lucene) or structured (e.g. MySQL, Oracle, Netezza), but most fail to do both well. We will dimensionalize unstructured data to create structured data that can be easily fused with graph / network data (e.g. communication participants).

C. Index information temporally. *Why?* So we can monitor and measure conversations, relationships and attributes over time. By tracking over time, we can determine if something is unusual or not. As well, this will give us a new dimension for measuring relationships between things by looking at correlations.

D. Natural-language-agnostic. *Why?* The system must not be limited to any language since the threats emerge and morph over time.

E. Domain-agnostic. *Why?* System must not be limited to a particular domain or subject area and must be capable of adapting to emerging threats.

F. Ability to execute self-generated complex semantic expressions against seamless fusion of streaming and historical data in real time. *Why?* Threats emerge and change over time, so flexible semantic expressions will be required to support the needs of the mission.

G. Discover new topics and relationships. *Why?* We need a force multiplier. The System must learn new topics and relationships over time and allow users to provide feedback. This machine learning will be based on our ability to index information temporally, statistical analysis of the time series, and correlation of time series with each other as well as external sources (real world events).

To accomplish the objective of detecting, analyzing and reasoning related to memes, we need indexes and algorithms that are not limited by our current knowledge or understanding. As new techniques and discoveries are made in the fields of cognitive science, neuroscience, linguistics, and computer science, we must be capable of integrating these new sources, techniques and algorithms without having to start from scratch.

2.3 Behavioral Modeling and Analysis Goals

2.3.1 Cognitive Analytics Goals

A. Create *concepts* from parsing natural language statements of commonsense knowledge. *Why?* To allow analysis of subjects (memes/topic models), related noun phrases (actors), verb phrases (actions), and modifiers or adverbs (directionality/values), including individual words.

B. Use analysis based on the use of semantic primitives such as semantic primes. *Why?* To develop a common linguistic grammar, thus allowing the deconstruction of communicated thoughts into standard component form, which will help develop mood and mind state / intent analytics capability.

C. Create matrices of features or assertions that are attributes of nodes and communal nodes. *Why?* To enhance the speed and efficiency of the analytical process.

D. Perform *blending* for *cross-domain inference* and reason jointly across more than one independently created knowledge base. *Why?* To make our approach applicable to realistic logic patterns and easy integration of multiple knowledge databases.

E. Organize the memes according to their most significant patterns, corresponding to semantically-meaningful distinctions, and extract this information as input into attributional profiles. *Why?* To measure a meme's volume relative to all other topic models.

F. Recognize unusual versus usual information, based on commonsense knowledge bases. *Why?* To mitigate deliberate misdirection or misinformation.

G. Create historical trust profiles for individuals and topics, and build social network participant profiles. *Why?* To measure credibility and influence.

H. Build attribute profiles for the nodes in the network. *Why?* To determine the intent of that node or a community of nodes and enhance the behavior projection capability.

I. Build the capacity for topic trend analysis that incorporates narrative-structure analysis, sentiment detection and opinion mining. Why? To add the unique dynamic of intent and behavioral projections.

We will greatly enhance the quality of METSYS analysis by exploiting current and state of the art research on how the human brain produces and processes linguistic information, and what information is available for commonsense processing. To achieve that, we will incorporate the latest research on discourse analysis and behavioral modeling into the suite of tools developed within the project. We will also take advantage of existing commonsense knowledge basis to help the system 'mimic' the processing done by a human analyst.

2.3.2 Persuasion and Influence Goals

A. Develop and apply a culturally-agnostic persuasion models. *Why?* To allow identification of persuasion campaigns across a variety of linguistic and cultural domains.
B. Identify, measure and perform persuasion processes based on attributes of nodes and nodal communities. *Why?* To counter persuasion campaigns.
C. Identify, measure and perform influence operations based on attributes of nodes and nodal communities. *Why?* To coopt persuasion memes.
D. Identify any correlation, in phase or out of phase, between two or more memes or nodal communities. *Why?* To identify cross-domain persuasion efforts.
E. Identify propagation delay, hijacking of memes, replacement memes. *Why?* To define the behavior of competing memes.
F. Measure the propagation rate of memes and impact on trust of nodes and nodal communities. *Why?* To enable assessment of campaign impact.
G. Measure how much can be propagated through a given medium *Why?* To measure a meme's relative volume.
H. Identify optimal centricity for meme injection. *Why?* To facilitate effective influence operations.
I. Learn and develop memory of historical changes to the graph structure based on persuasion and influence experimentation. *Why?* To allow the system to become more adept at identifying and analyzing persuasion campaigns.

Achieving these persuasion goals will enable METSYS to orient decision-makers to the social media environment and enhance the effectiveness of Strategic Communications decisions. Our experimental approach will be unique in its integration with cognitive dynamics. However, since social media is only one aspect of communications and credibility comes not only with accuracy and timeliness, but also consistency, any Strategic Communications plan must be coordinated across all forms of media and consistent with actions in the physical world. This will require any complete decision cycle to be semi-automated and concerned with planning processes and human-computer information integration and visualization.

The Persuasion and Influence effort will enhance research progress in the study of intent-based analysis in the *Culture and Conflict Studies* and *Computational Approach to National Security Programs* at the Center for Advanced Defense Studies. We expect to produce white papers and studies on semantic correlation to deception, social media Strategic Communications strategies for information superiority, and case studies to support tasks in our statement of work.

2.4 Performance and Measurement Goals

A. Validate measurements inside graph analytics (e.g., frequency of presentations of topic model within period intervals and duration intervals for occurrences of individual presentation with a single topic model instance, edge to edge presentations, edge to nodal community edge presentations, and nodal generated occurrences for attributional power analysis. *Why?* To ensure reliability of mathematical calculations and consistency of measurement presentation.
B. Perform human sampling comparative analysis of randomly selected topic model graph analytical outputs across a given interval for assessment of automated

presentation graph output measurements. *Why?* To ensure validity of automated recognition of specific topic model occurrence and to ensure that presentation calculations are discrete between sampling groups (i.e., edge to edge, edge to nodal community, etc.)

C. Perform graph to graph analytical outputs for each discrete sampling group (i.e., edge to edge, edge to nodal community, etc.). *Why?* To perform longitudinal analysis of graph analytical output power for threshold verification for each topic model to reach meme classification both in interval and duration periods.
D. Perform lifecycle analysis of discrete topic models that have reached threshold over more than one interval or duration period to identify frequency and total lifecycle of recurrences of individual topic memes. *Why?* To add the unique capability to understand how memes emerge, fade below threshold, and reemerge as memes of concern to allow comparison of the responsiveness of a meme based on attention behaviors of nodal and nodal community groups for awareness, attention, and sensitivity of the network to particular memes. We will measure the delta between those memes that remain at threshold, those that shift below threshold, and those that hold significance across multiple durations (e.g., raise, fall, and reemerge to threshold).
E. Validate rules process for building node or nodal community properties and attributes (e.g., age, sex, education, areas of interest, activity, centrality, influence, trust level, span of reach, etc.) or edge properties (e.g., weighting of relationships, temporal relationships, initiation or response activity levels, frequency of transmissions ratio, etc.) of targeted matrix profiles of individual nodes or nodal communities. *Why?* To enhance attribute discovery and clarity in centrality of node or nodal communities, attitudinal dispositions of these profiles, and to allow for information process flows among and between nodes (i.e., source or recipient), activity levels of information generation or information user power per node or nodal community, measurement of persuasion or influence impact analysis from persuasion or influence campaigns.
F. Validate automated detection of persuasion or influence campaign using semi-automated human support through Subject Matter Expert input for discrete persuasion or influence campaign (campaign being operationalized as a new topic model meme actor that can be analyzed across behavioral and intentional analytical parameters). *Why?* To ensure accurate measurement of specific persuasion or influence actions.
G. Validate ability to affect transmission capacity in context of denial of service for transmission or receipt of the routine or other content through the network from deflective persuasion or influence campaigns targeted at an individual nodes or nodal communities. *Why?* To develop the capacity to interrupt or modify memes.
H. Measure the impact of deflective campaigns through actor (node) attention and impulsivity through impacts on node or nodal community matrix profiles. *Why?* To develop the ability to change or disrupt patterns of behavior.

To ensure that METSYS accurately performs noted objectives of detecting, analyzing and reasoning related to memes, we need to validate the processes for node and nodal community matrix profile data collection and retrieval. We will perform testing and validation of software and algorithmic code prior to insertion in “test environment” and subsequent validation of these utilities through random sampling and human random comparative sampling activities. We will provide reports for each tested code or algorithm as well as statistical assessment of these tool sets and code with the intent to achieve a minimum p value of .05. Finally, validated code will be provided in regular intervals during code drop schedules.

As noted above, the following section of this proposal follows the same logical structure and presentation identified in proposed goals with the detailed Technical Plan activities and actions proposed to achieve these.

3. Technical Plan

3.1 Technical and Architectural Plan

3.1.1 System Design

The team will design and develop METSYS to enable humans and algorithms to mathematically model and provide probabilistic projections of human behavior through the analysis of changes in social network structure (nodes and edges) and the analysis of semantic patterns in communications over time. This analysis will be augmented by providing capabilities for in-depth analysis of discourse by individuals and groups identified as the nodes in the structure, and modeling their behavior and intent, to provide better predictive analytics.

Risk is inherent in the development of any large-scale system. METSYS will be developed using an Agile / Scrum development methodology, and as such, will benefit from frequent developer interaction and incremental software deliverables which help reduce risk. As well, the team consists of members with deep experience and knowledge developing components and services similar to those anticipated in the proposed system. We shall benefit from the team's experience, know-how, and lessons-learned from previous endeavors.

3.1.1.1 Topic Models

Topic models will represent any object (e.g. people, companies, things) and even abstract concepts, such as 'fear' or 'civil unrest'. Topic models will provide the link between discourse and cognition. Topic model expressions will be comprised of words, phrases, symbols and logical operators that define a high order concept (e.g. person, place, thing, and event) as well as more abstract concepts (e.g. fear, sentiment). (See Topic Model Expressions section below). Multiple languages (e.g. English, Chinese, and Arabic) can be combined into a single expression. In addition to simple word, phrase expressions, the following operators may be combined to produce high-precision queries. Topic model expressions can be changed dynamically and will be processed ad hoc in real time or near real time across all data indexed in the METSYS. Topic model expressions will be used as criteria in API Calls to retrieve content, statistics, time series, etc.

The semantic indexes in METSYS will have no knowledge of topic models, and topic models will be processed in an ad hoc fashion enabling models to evolve independent of data and indexes. METSYS will consist of distributed indexes that will be primarily stored in main memory. The goal is to have a dynamic system where topic models can be created and evolve over time. Traditional approaches extract, identify and tag just prior to indexing. Our approach will index tokens, metadata and attributes without making assumptions about how this data will be queried in the future and as a result may require more computations at run-time, but will deliver on the goal of being dynamic. By partitioning indexes logically by source and time [6-

13] across many nodes, and by storing parts or all of the indexes in main memory, we can offset the cost of performing more work at runtime, and achieve the desired query performance. This approach of isolating topic models from indexing allows models to be constructed at run-time to evolve as language, taxonomies, and knowledge evolves. Versioning and security will be supported for topic models, and they can be updated and viewed in a historical context without having to re-index or re-organize indexes.

3.1.1.2 Topic Model Attributes

Each topic model will have the following attributes:

SYMBOL: Will represent a unique moniker for a topic model in the context of a user's name space. (e.g. civil.unrest, Bill.Gates)

DESCRIPTION: A description of the subject that represents that discreet moniker or topic area of interest.

PERMISSIONS: Defines the users (or systems) that can identify, view, update or execute a topic model (e.g. public, private, organization)

VERSION: Number assigned by METSYS or user to a specific instance of a topic model

EXPRESSION: words, phrases, symbols and logical operators that define a high order concept or *topic*

3.1.1.3 Topic Model Expressions

*Topic model expressions will be defined using the following operators***:**

( ) **Sets**

Sets define a list of words, phrases or symbols separated by commas where the comma denotes disjunction (OR)

ex. (Apple Computer, شركة ل اب ر يوت ب كم,苹果电脑, iPhone, iPod, iPad, imac, itablet**)**

+ **Conjunction (AND)**

Conjunction defines words, phrases, symbols or sets that are co-referenced within relative proximity in terms of time or position.

ex. (IBM,إم ب ي آي,IBM公司, big blue, International Business Machines,ال تجارية آلات
ية الدول,国际商业机器) + (acquired,مك ت سب,收购, is acquiring, will acquire)

{n}**Unordered Proximity**

Tokens are adjacent to each other and within a specified # of tokens apart

ex. (IBM,آي بي إم,IBM公司, big blue, ال آلات ال تجارية ال دول ية,国际商业机器,International Business Machines) **{5}** (acquired,مك تسب,收购, is acquiring, will acquire)

[n] **Ordered Proximity**

Tokens are adjacent to each other and ordered within a specified # of tokens apart

ex. (IBM,آي بي إم,IBM公司, big blue,ال آلات ال تجارية ال دول ية,国际商业机器, International Business Machines) **[5]** (acquired,مك تسب,收购, is acquiring, will acquire)

$ **Semantic Expansion**

Expand topics based on symbol and morph expression to include related terms.

ex. ($IBM) {10} ($Acquisition)

- **Negation**

Match where include terms are present and excluded are not

ex. (Iraq,Iran) – **(**nuclear,nukes)

3.1.1.4 Memes

A meme will be defined as a topic model which (a) reaches a certain threshold of adoption (as a percentage of social media nodes) over a given period of time, and also (b) crosses an inflection point representing a threshold of presentation relative to its past. Such a topic model continues to be defined as a meme until it no longer satisfies this threshold of nodal connectivity. It may later re-present as a meme if it re-crosses its original threshold of presentation.

Every meme will have a duration or lifetime – the total period in which the presentation of the topic model meets or exceeds its average presentation for X interval or time relative to the broader timeline. Since memes can present, fall below the threshold, and re-present, one of the challenges of analysis is to represent what time interval defines a valid meme.

3.1.1.5 Time Series (aka Topic Ticker)

Time series will represent source-weighted frequency counts for topic models across varying period & interval combinations (e.g. 1 day/1minute, 30 day/daily). Time Series counts will be based on custom source weightings that can be defined at execution time. Aggregation of interval slots will occur at execution time. Time series can be analyzed like stocks using stochastic, Fibonacci and other statistical algorithms.

3.1.1.6 Statistics

Statistics will represent measures of topic models (e.g. mean, standard deviation, Z-Score, exponential moving average) across varying period & interval combinations (e.g. 1 day/1minute/ 30 day/daily). Every token that moves through METSYS will cause statistical updates for all matching topic models. Statistics will be based on custom source weightings that can be defined by users. For example, when statistics are queried for a particular topic model, placing a higher weighting on Twitter, METSYS may generate high Z-scores if that source is weighted higher. Executing the same query with a low relative source weighting might show the same topic model at normal levels.

3.1.1.7 Correlations

METSYS will enable correlations of topic models with other topic models and also changes in graph structure over time. The METSYS will perform cross-correlations and auto-correlations to discovery patterns within a conversation among groups of nodes or network. Correlations can be calculated in-phase or using various phase shift increments. This is useful in determining cause/effect where lags are involved.

3.1.1.8 Source Weighting

Custom source weightings will be applied at execution time, thus allowing an individual request to weight sources differently based on the application of such query. This allows queries to weigh certain sources more heavily than others.

3.1.1.9 Ad hoc Query Execution

Topic models will be allowed to evolve and defined continuously based on user input or discovery/suggestion resulting from algorithmic processes in METSYS. To guarantee the current representation of a topic model is executed at all times, METSYS will always execute queries in an ad hoc manner. Users and systems may query the METSYS with any semantic query.

3.1.1.10 Machine Learning

Emerging patterns and topic models will be discovered over time based on correlation and proximity of topic models to other characters, words and phrases. Dynamic creation of topic models and ability to classify topic models according to whether they are user-defined or system inferred. Factor in statistics and correlations across time periods with proximity or co-reference. Known algorithms / techniques (e.g. Bayesian Networks, Latent Semantic Indexing, Principal Component Analysis, Singular Value Decomposition, Expectation-Maximization Clustering), as well as novel techniques will be utilized as part of semantic topic discovery.

3.1.1.11 Event Processing

Deductive rules will be supported using customized off the shelf products (e.g. JBoss Rules). Rules will assert inferred events (e.g. Terrorist Attack) and new topic models. Rules will be organized around domain-specific event types and involve relative and temporal proximity of topic models across space and time, statistical thresholds, association with groups and other topic models, and conditional criteria. Satisfied rules result in the extraction of relevant attributes and values that drive downstream processing and decision making.

3.1.2 Performance and Scale

METSYS will be based on an MPP (massively-parallel processing) and Shared-Nothing Architecture. COTS (commercial off-the-shelf) software and hardware products will be leveraged where applicable. METSYS will minimize query execution times by loading indexes primarily in to main memory. METSYS will support serializing indexes to disk for increased indexing capacity and high availability / failover. METSYS will distribute work of all services: query, statistics, correlations, etc. to as many processors and processes/ threads. METSYS will minimize the time from collection to indexed / rest to as close to zero as possible. METSYS will target an average execution time for queries of 50ms or less. METSYS will use techniques to maintain an approximate measure of all statistics on all known topic models at all times. For example, if a new Twitter message matches five topic models, then within 100ms METSYS should have updated statistics across all period/intervals for each topic model.

Many traditional algorithms, especially for NLP, were not originally designed for distributed computing environments, like GRID or Cloud. In particular these traditional approaches did not plan for the real-time analysis of the volume and variety of data generated today. Many of the approaches have inherent limitations, such as their ability to parallelize and/or thread tasks, ability to move from a disk-based or I/O bound approach to an all in memory or hybrid model. Many systems require extensive pre-processing and limitations related to how much of the processing can be distributed, such as Latent Semantic Indexing / LSI. Our approach will not be bound to a single process or single machines and will be designed from the ground up to maximize distribution of data, indexes and processing.

3.1.2.1 Horizontal Scalability

Horizontal scalability is achieved by adding additional servers or virtual machines (VM) with additional processors, memory and storage, resulting in increased processing and index capacity. Horizontal scalability is consistent with Cloud-Based computing. When we need additional processors, RAM or storage, we will be able to quickly and cost-effectively deploy on-demand.

3.1.2.2 Linear Scalability

Linear scalability will be achieved by distributing processing across an optimal number of servers and processes – eliminating data skew and hot spots. METSYS will be designed to distribute processing of queries and analytical processes, leveraging multiple cores per processor per server. Our distributed indexing approach will be similar to database sharding or temporal index sharding [6-13]. This approach will allow us to parallelize execution of queries against many nodes simultaneously.

3.1.2.3 Commodity Hardware & Software

No specialized hardware will be required. In the future, optimizations may be achieved by using specialized hardware, such as Field Programmable Gate Arrays (FPGA), but it is not required. The METSYS stack will be comprised of the Linux Operating System, Java Runtime Environment, Apache Tomcat, MySQL, and other open source technologies and products that are in wide-spread use and have been proven to be reliable and scalable. Maturity and widespread adoption and support for the Linux operating system and Java make them ideal candidates. In addition to the widespread adoption and support tools, Java's "Write Once, Run Anywhere" promise make it suitable given the expectation that METSYS will have different hardware configurations (e.g. Intel and AMD processors, x86 and AMD64 architectures), especially over time. Different hardware configurations make it more difficult to package and deploy with other software languages and environments, such as C/C++. Java will allow us to easily build and deploy updates to METSYS will very little knowledge of the physical hardware. In addition, most Cloud-Based services, like Amazon EC2 (Elastic Cloud Computing), have support for Linux and Java, allowing us to focus on what METSYS is going to do rather than how. METSYS will be mission ready with high availability / failover, replication, and load-balancing principles integral to its design.

3.1.3 Flexible-Modular Design Elements

3.1.3.1 Service Oriented Architecture

We will develop and/or extend existing service frameworks. Services will be developed that process on-demand requests, as well as optimized processing of other services and data streams. For example, topic models will be compared to every data message / stream ingested in to the METSYS. When topic models match the data message / stream, services will be able consume this stream of matching topics in a data-driven manner to perform operations based on the detection of these topic models. Each of the following classes of services will be extensible:

Source Adapters – transformation, encoding, source-specific processing
Ingest Services – Processing and enrichment prior to indexing
Statistical Services – Statistics based on semantic topic models and queries.
Correlation Services – Correlations based on semantic topic models and queries, correlated with fact-based time series (e.g. Real-World Event Timelines).
Analytical Services – We will provide a framework and SDK for building and integrating custom analytics. In particular, we will provide/integrate algorithms for intention awareness, sensemaking, etc.

Data ingest, indexing and modeling tiers will be completely independently of each other. That is, the "knowledge" about the data will be maintained separately and can evolve independent of the data. Indexes will load directly from native sources/feeds, message queues, etc., and models/tagging will change in real time with no impact to source data. Source handlers will exist for databases, message queues, proprietary wire feeds, television, and other sources.

3.1.3.2 API

METSYS will support a REST (Representational State Transfer) API. METSYS may support other system interfaces, such as TCP/IP Sockets, Multicast, etc. At a minimum, support for query, topic management, system management, and event listeners will be supported. Basic HTTP Authentication will be supported, and we may support other authentication methods (e.g. OAUTH) if appropriate. Standard output formats, such as XML (Extensible Markup Language) with XSD (XML Schema Definitions), or JSON (JavaScript Object Notation) will be used.

3.1.3.3 Data Fusion and Federation

Diverse data from real time streaming feeds and historical data archives. METSYS will support the seamless integration and distributed indexing of new data folded in with existing historical data. Real-time streaming feeds will be handled with a sustained ingest that has no down time. Near real-time indexing of collected data will occur with near zero latency. Source adapters will handle any transformations or pre-processing prior to indexing.

3.1.3.4 Semantics will be decoupled from Indexing

Indexing in METSYS will make no assumptions about the semantics or subject matter / domain. By keeping semantics separate, we will be able to build topic models that are cultural, behavioral and/or domain specific. Topic models will be able to evolve without having to re-index information when models are developed or enhanced. Language is evolving and proper grammar is often not used in social media and many other forms of communications (e.g. Texting / SMS). Relying on entity extractors (e.g. InXight, SRA NetOwl) and relationship extractors (e.g.

Attensity, Autonomy) will fail in these environments due to their expectation of proper grammar, dependence on supported languages, and in some cases, the reduced size of messages (e.g. Twitter has maximum message size of 140 characters). Semantic relationships will be discovered and confidence determined over time by using correlations, proximity and other techniques. Distribution amongst sources will confirm the degree of contagion and a measure of how common the knowledge is over time – the more unique conversations over time, the higher the validity and potential impact and relationship to real world events and behavior.

3.1.4 Indexes and Algorithms

3.1.4.1 Natural Language independence

Social media has rapidly increased the creation and use of short-hand, emoticons, and variant vocabularies. No assumptions can be made at indexing time about how the data will be queried and analyzed. Indexing will be language independent and all characters will be UTF-8 encoded before indexing. Tokenization will be similar for all languages with white spaces (e.g. English, Spanish, and French). Languages with no word boundaries, such as Chinese, will require each character to be tokenized. By supporting and indexing all languages consistently, we will be able to create multi-lingual semantic queries with multiple languages in a single query. For example, **(**Apple Computer, شركة ل اب ر يوت ب كم,苹果电脑, iPhone, iPod, iPad, iMac**).** Many systems rely on Entity Extraction to identify nouns, proper nouns (e.g. people, places, organizations), and verbs. Entity Extraction is typically slow and limited to a finite set of languages, and pre-defined rules and heuristics specific to each language. METSYS must not make any assumptions about how humans and BOTS will communicate. Our approach will be to separate the semantics from the indexing, and achieve natural language independence.

3.1.4.2 Temporal Indexing

Indexes will organize tokens by time slots, normalized time around the smallest unit of aggregations (e.g. 10 seconds). Tokens will represent words, symbols, numbers, metadata, and nodes. This approach is fundamentally different than a traditional inverted word index. With an inverted word index, time is typically an attribute of the document and is not normalized. Since time is considered secondary and not integral to the index, it requires significantly more processing and can be I/O intensive, since this value is not part of the index and may be on disk, to derive a normalized timeslot. The advantage to storing a normalized time slot in the index is the aggregation of time (e.g. sum all mentions of topic x for the day) can be performed at the smallest unit of work – a thread executing a single segment of a query. This will allow us to quickly aggregate data based on an aggregating interval (e.g. day) from the smallest timeslot interval (e.g. 10 seconds) and finally aggregate all results in to a complete, accurate result set.

A document provides a logical container or space where tokens exist and have ordinal position(s). This logical space can be extended to represent a physical location. This will allow us to answer questions like, did multiple actors communicate regarding a particular set of topics during a particular time frame in a particular region of the world (e.g. city/province/country).

METSYS will handle messages / document-centric content as well as streaming data sets. Most systems primarily index documents (e.g. Lucene, Google) and don't necessarily handle streaming data sets that have no document per se. The main difference between the two types of inputs is the messages / document-centric inputs will have a well-defined beginning and end, and the streaming data will not. The streaming data boundaries are implicitly defined by time. Messages and streams of information will be indexed using a similar and compatible index structure. Streams will be segmented in to time windows and will not have a document. For example, a television broadcast contains discourse that is not a message or document with a beginning and ending position, but it can be segmented in to 10 second windows. A topic model can be developed that detects a topic that requires multiple words to be mentioned within a 30 second window of each other or within the same document. METSYS will be capable of executing a semantic query against both streams and messages.

3.1.4.3 Partitioning

Indexes will be partitioned by source and time. Multiple processes and threads will execute in parallel across multiple servers and process spaces. Intermediate results will be aggregated incrementally as processes have completed processing individual segments of a query against all appropriate indexes.

3.1.4.4 Storage

Metadata will be stored as name / value pairs. Names will correspond to entities defined in a taxonomy (e.g. XSD, RDF/S, OWL, Protégé). Values will be normalized, transformed, and enriched where appropriate. A large-scale, distributed name / value storage engine (e.g. MongoDB, Cassandra) will be used. Name / value pairs will be indexed using the temporal indexing engine.

Directional graphs, with timeslots and metadata associated with the edges. Reification will be utilized to provide relationships to relationships. Graphs may be exported and shared using common representations, such as RDF (Resource Description Framework). Support for graph operators, such as centroid, nearest-neighbor, shortest-path will be supported. Graphs will be stored in a distributed, scalable graph database engine.

Nodes and edges in graph will be indexed temporally with content and metadata, which will allow us to ask content, metadata and graph related questions in a single statement.

Content will be stored in either a distributed file system (e.g. Hadoop Distributed File System / HDFS) or NAS (Network Attached Storage) or SAN (Storage Area Network). Content will be retrieved with a UUID (Universally Unique Identifier) which can be quickly translated to a physical reference (e.g. File Path) via a content catalog.

METSYS will be capable of storing Indexes to long term storage (e.g. File System, NAS/SAN). The index files will likely be comprised of multiple dimensions (e.g. token, timeslot) and storage should be optimized for multi-gigabyte file sizes and random access (e.g. read from a specific byte offset in file). Processing will be moved as close to data as possible and in general, movement of data will be minimized.

3.2 Behavioral Modeling and Analysis Plan

Statistical and time-series based analysis of social networks and media as a whole will identify and provide a wealth of usable information about the network edges (such as topic models) and nodes (individuals and groups). To provide deeper analysis capabilities, METSYS will also employ a number of techniques focusing on the nodes in the networks. This capability will complement the network-level analysis by identifying the drives behind nodes' behavior, and helping predict the possible changes in the network structure and other characteristics.

3.2.1 Discourse and Mood State Analysis

The mode of network analysis we propose integrates several levels of analysis in order to capture and usefully analyze only the relevant components of discourse. The structure of human thought (with key elements of intention and cognition), and general properties of language all converge and contribute to the analytical picture we seek to paint with respect to social networks.

As the foundation of interaction within social networks of all kinds, individual discourse must be analyzed effectively in order to discern intent and make credible behavioral predictions. In particular, it is necessary to deconstruct thoughts communicated through language (spoken or otherwise) into an elemental form that can be parsed and evaluated. From a computational perspective, an intuitive tool to use is the notion of the semantic primitive (or prime). Semantic primes are word-concepts that are innate to humans, regardless of native language. They need not be defined in terms of other linguistic constructs, so at some level, all individual discourse is composed of semantic primes. Semantic primes offer an insight into the development of the human brain, but equally significant is the fact that these fundamental concepts are shared among all humans, and are thus part of every instance of linguistic communication, independent of language.

Prior research has made significant relevant strides in conceptual discourse analysis. In particular, the work by Howard et al. [16] lays out a means for determining and predicting patients' mood states by algorithmic means. In particular, this means linking mind axiology, or conceptual beliefs common to particular cognitive conditions, to behavioral trends expressed by patients being diagnosed. Although past research had a clinical focus, the results are generalizable to the predictive analysis of a much wider range of human interactions [15]. We will also incorporate computational linguistic work on functional-stylistic analysis of natural language [3], which has been used to determine cognitive modes in scientific thinking [2] and personality typing [1].

The methodology allows for determining value for each individual word, and using the values to determine the mind states of an individual. This analysis takes into consideration the 'intrinsic' value of the word (based on semantic prime-derived value), as well as temporal, consequent, and contextual values of the word, which vary depending on the language, culture, and experiences of a specific individuals. Specialized knowledge bases will be developed or adapted for use for this purpose within the system, such as the commonsense logic. The pluggable architecture of the core system will allow for integration of additional layers of knowledge [14].

In addition, statistical methods will be used to find metaphorical usages to detect *cognitive frames*, which show how different communicators cognitively represent ideas and issues differently [5]. As a simple example, whether the issue of legalized abortion is talked about as a question of "life" or of "choice" tells a great deal about the speaker's political views. Metaphors will be detected by finding usages of verbs and nouns that depart from normal usage patterns; when a number of related deviant usages are consistent in a subset of the data, metaphorical cognitive frames will be detected, which will give important clues as to the cognitive structures being expressed [4].

3.2.2 Commonsense Logic

Commonsense logic provides a higher, more generalized level of analysis. In essence, commonsense logic seeks to pre-fetch common conceptual associations and axioms shared by humans, and employ them for a more expedient analysis of raw communication data.

Recently, fundamental advances have been made in Artificial Intelligence in representing and reasoning with knowledge about people's everyday experience, which is referred to as *commonsense knowledge*. The Open Mind Common Sense (OMCS) knowledge base, developed at MIT, contains over a million assertions in English, which we estimate represents 1% of an average person's commonsense knowledge. OMCS constructs a semantic network from assertions in natural language by using a variety of linguistic patterns. It has about 20 native relations that generally express relationships between concepts, such as "X is a kind of Y" or "You use X to Y". The relation set was extended with a few relations specifically for the neuroscience domain, e.g. "Connected To", "Activates", "Regulates" and "Inhibits".

Rather than relying on a fixed hierarchical ontology, our approach creates *concepts* from parsing natural language statements of commonsense knowledge, allowing noun phrases and verb phrases as well as individual words to be concepts. Redundancy and contradiction are permitted. A matrix of concepts vs. *features* (an assertion minus a concept) is created, so the cells of the matrix represent all possible assertions about the concepts and features being studied. Assertions that actually appear in the knowledge base fill in the corresponding cell in the matrix with a value indicating confidence in that assertion, if known. This matrix is usually very sparse, since any given knowledge base only contains a small fraction of possible assertions. [18]

Reasoning is performed by *reducing the dimensionality* of that space, using the mathematical techniques of Principal Component Analysis. This has the effect of "filling in" the missing cells in the matrix. It organizes the space according to its most significant patterns, which can often correspond to semantically meaningful distinctions such as "good vs. bad" or "easy vs. hard". The technique is more tolerant of noise, imprecision, contradiction, context and point of view, than traditional logical inference. A related technique, *blending*, can perform *cross-domain inference*, reasoning jointly across more than one independently created knowledge base, even if the knowledge bases may not initially completely agree on their vocabulary, or may not completely agree on certain assertions. [4]

We will employ these reasoning techniques in a unified organizational system. This component of our project will enable analysts and researchers to ask general questions about what is relevant to a given concept, and to navigate through dimensions that represent relevance or interestingness computed dynamically and specified by example.

Successful implementation of our project will require the mitigation of deliberate misdirection and misinformation. One can recognize an *unusual* condition by contrasting it with the *usual* case. We can determine what is usual, and mitigate misinformation, by referencing OMCS.

We will create historical profiles for individuals and topics, and build social network participant profiles to measure credibility, influence, and other characteristics. Based on these participant profiles, we will build a value system for the node to determine and track the intent of that node.

3.2.3 Intention awareness

The inclusion of intent processing is an important feature. Systems that incorporate situational awareness or enhance that of human operators have found successful applications, especially in scenarios dealing with high volumes of data. However, such applications may give the impression that situation awareness is complete when only a few parameters are known. These parameters typically include raw quantitative data, and by themselves, tell us little about situations when they are governed by human actors, and hence by the complex human mind. By integrating concepts of intention awareness, or taking into account the internal processes of actors / nodes themselves, it is possible to significantly improve this analysis.

The fundamental utility of intention awareness in social network analysis is that, as semantic primes do for language, intention provides a basic grammar for describing the relationship between cognition and action. In this way, we differentiate ourselves from other approaches to network analysis not only by improving the speed or efficiency with which it is done, but redefining the results to be more useful. That is, there are interpretable intentional structures that can be provided to a computer [17] such that it is able to better prioritize the incoming information and produce quick, useful results to the operator. In addition, these results, being based on the cognitive architecture of the human mind to begin with, are more easily assimilated by those who need to use and interpret this analysis. This approach has been used successfully in battlefield situation awareness and information sharing applications [24], and METSYS will expand the area of it applicability to social media and social networks.

3.2.5 Persuasion and Influence

The study of persuasion and influence is rooted in two major fields, advertising and military information support operations (MISO). Most persuasion models are adapted to advertising. Advertising seeks to generate a measurable effect (increased sales) in the physical world. The advertising approach was not designed to address a Malthusian battle of the narrative or decrease behavior. MISO, on the other hand, may seek to influence audience behaviors relative to campaign objectives in order to win a battle of the narrative as part of a strategic communications campaign and often lacks quantifiable metrics to determine relative success. There is no current model bridging both domains. Existing persuasion models have well identified flaws [19-21]. MISO persuasion campaigns follow bureaucratic processes and while recognizing the impact of accuracy, timeliness, security, privacy, delegation of voice, and unity of voice on the relationship between credibility and trust, they rely more on art than science [22].

Persuasion and influence efforts face the following basic challenges:

- The effectiveness of persuasion and influence campaigns are tied to the audience and its existing attitudes and cognitive biases.
- Influence and persuasion exist in an open, complex, adaptive system.

Risks and mitigation: The project risks the creation of an algorithm for persuasion that is specifically adapted only to the very narrow and likely self-selected audience chosen for

Technical Area 2. Our team hopes to mitigate this by conducting our own testing and validation against a cross-cultural body of individuals to ensure that the basic algorithm is broad enough that with couplings for target unique cognitive biases METSYS can dynamically adapt to evolving group perspectives.

Further, the closed environment may not replicate the real-world closely enough to produce useful results. Our team has described a system that will allow the introduction of other media sources in future versions to allow for better granularity of indirect inputs.

The young science of cognitive informatics may not be mature enough to provide automated cognitive attributes that usefully model intent and sentiment analysis. While advances in automated mood analysis from written media, based on axiology, exist, many of the commercial claims for sentiment analysis are unsubstantiated with rigorous scientific testing. We intend to mitigate this by creating a loosely coupled system that will admit semi-automated engines for intentionality and attitude modeling.

3.3 Performance and Measurement Plan

Testing automated and semi-automated systems requires validation of correctness of mathematical computations performed in rule sets and algorithmic formulas for each computational activity. Further, each of these process actions must be tested individually and collectively in conjunction with overall systems applications. Each level of algorithmic decision gates must be validated to ensure performance outputs meet specifications. Technical evaluation of this project will involve validation, verification and evaluation of the initial parametric outputs for topic model generation to ensure reliability of mathematical calculations and consistency in performance of calculations over time.

As this system performs not only detection and identification of topic model memes based on real or near real time data assessments, but also performs interval and duration temporal assessments of resulting topic models, a second order analysis of threshold parametrics for accuracy and consistency of tracking and relating topic models across time periods will be conducted. This will include assessments to ensure that topic models that fade below threshold and reemerge are accurately associated with the respective initial topic model.

Verification and validation of graph analytic outputs for automated lifecycle analysis of topic model will involve testing of mathematical calculations performed in comparative graph-to-graph analytical outputs with evaluation of effectiveness in power analysis and computational accuracy. Social network analysis of METSYS generated node and nodal community structures will be performed to ensure accuracy in mathematical calculations and to validate node and nodal matrix profiles.

Risks and Mitigation: Automated systems that rely on rule sets and algorithmic detection schemes can introduce bias in detection and valuation of topic model categories. Additionally, aggregation of rule sets and algorithmic formulas can result in data correlation errors. As a means to resolve these potential errors we will perform the aforementioned testing activities in analysis of each algorithm developed for mathematical correctness and collectively with overall system applications. Random samples of discrete topic model output data will be extracted and tested for statistical accuracy with the intent to achieve a minimum p value of .05. Further, as a method of testing efficacy of topic model detection and subsequent second, third, fourth level analysis of these topic models in the areas of values, influence, and intent, we will use Subject Matter Experts. SME's will perform random sample analysis of specific topic model source data for selected time slots to compare METSYS reported output associated with topic model recognition as well as METSYS reported output data on values, influence, and intent associated with the respective topic model. Also, the Performance and Measurement team will consult with the organization selected by DARPA to perform Technical Area 3 (TA 3) Algorithm Integration, Test and Evaluation to ensure consistency in test and evaluation activities.

4. Management Plan

We have sought to develop a management structure that is efficient, dynamic, integrated and multidisciplinary. The day-to-day management of the overall project will be the responsibility of Dr. Newton Howard as the Principal Investigator. Project Management oversight will be accomplished by David Johnson as the Program Manager. Every effort has been made to remain extremely cost-effective with most of our budget allocated to research and not administration.

4.1 Approach

4.1.1 Agile / SCRUM Development

Scrum is an iterative, incremental framework for project management often seen in agile software development, a type of software engineering.

Although the Scrum approach was originally suggested for managing product development projects, its use has focused on the management of software development projects, and it can be used to run software maintenance teams or as a general project/program management approach.

4.1.2 Team Management

- Weekly call concerning milestone achievement
- Bi-weekly budget call
- *Major Deliverables Every 3 months*
- *Minor Deliverables Every 2 Weeks*

- *DARPA Grant Required Reporting*

(1) Quarterly R&D Status Report - This report, due 30 days after the reporting period, shall keep the Government informed of the Center's activity and progress toward accomplishment of Grant objectives and advancement in state-of-the-art on the research and development involved.

(2) Special Technical Report - This report, due as required, shall document the results of a significant task, test, event or symposium.

(3) Final Technical Report - This report, due 90 days after expiration or termination of the Grant, shall document the results of the complete effort. It shall contain brief information on each of the following:

(i) A comparison of actual accomplishments with the goals and objectives established for the grant, the findings of the investigator, or both.

(ii) Reasons why established goals were not met, if appropriate.

(iii) Other pertinent information.

(4) Final Financial Status Report - This report, due 90 days after completion of the Grant, shall be submitted on a Standard Form 425 "Federal Financial Report (FFR)". The report shall be on an accrual basis.

4.2 Principal Investigator

Dr. Newton Howard will lead the overall supervision and integration of this project. Dr. Howard, working directly with the Program Manager, and the respective research project leads will be responsible for the overall coordination of the research effort, for completing deliverables, and for the timely completion of progress reports, and oversee the work within the research teams responsible for the major research goals. Specifically, he will produce:

- Technical papers and reports.
- Annotated slide presentation one (1) month after project start and annually thereafter.
- Implementation and documentation – documentation will describe algorithms, source code, hardware descriptions, language specifications, system diagrams, part numbers, and other data necessary to replicate and test the designs.
- Monthly Progress Reports.
- Final Report that concisely summarizes all project activities.

Dr. Howard is founder and the first chairman of the Center for Advanced Defense Studies. Currently he is the director of the Mind Machine Project and a resident scientist at the Massachusetts Institute of Technology (MIT). Dr. Howard holds a Doctoral degree in Cognitive Informatics from La Sorbonne, France. Internationally, he is a leading researcher on the Physics of Cognition (PoC) and its applications to Defense and International Security. Dr. Howard is a graduate of the Department of Mathematical Sciences at the University of Oxford, where he first proposed the Theory of Intention Awareness (IA).

In addition, Dr. Howard is the Founder and former Director of the Institute for Mathematical Complexity and Cognition (MC2), where he addresses subjects related to cognition, complexity and intentions and the Director of the Descartes Institute (France), where he focuses on behavioral models and codification that can be used to develop new approaches for counter-terrorism. Dr. Howard advises several organizations in the U.S. Special Operations community and has extensive experience working in the defense industry. He has served in the U.S. Armed Forces as a Strategic Intelligence Officer. Additionally, he holds multiple U.S. patents, and is the author of several publications in the areas of military information science, computer systems theory, and strategic thinking.

Dr. Howard's recently published Mood State Indicators research modeled and explained the mental processes involved in human speech and writing, to better predict emotional states. His natural language approaches to systems understanding and design pointed to the multi-dimensional structures embedded and nested within speech-based cognitive systems.

Understanding of these structures allows building more accurate software engines for modeling behavioral and cognitive feedback.

Dr. Howard has built research teams around the world with the best experience in introducing such models into "kinetic" and "non-kinetic" military and battlefield awareness systems such as the Coalition Secure Management & Operations System (COSMOS), and other US DOD research augmenting officers' understanding of the situation, and modeling potential outcomes. It is important to note that cultural intelligence/awareness, which are becoming increasingly important for today's multi-faceted warfare, have been incorporated into these models and have been used in command rooms, classrooms, and operational environments.

4.3 Algorithm and Software Development Team

Rob Usey (CEO Psydex) will lead a team that will provide:

- Vision, Leadership & Innovative Prototyping
- Integration between Data Analytics capabilities and cognitive research.

Rob is a visionary and innovator at the intersection of computational sciences and intelligence systems and has over 20 years of experience designing and developing innovative systems for some of the largest commercial and government entities in the world. IBM acquired his first company KnowledgeX in 1998. KnowledgeX was an early pioneer in social networking, ontology, data mining and visualization. Rob holds multiple patents in semantics and has a passion for game theory and the application of algorithms to understanding and predicting human behavior. For the past 7 years Rob has been CEO/co-founder at Psydex Corporation where the vision is to predict *What the World is Thinking*® in real time. Psydex leverages MPP (Massively Parallel Processing) and a novel temporal indexing approach to analyze data streams in real time to generate statistical patterns, trends and correlations across news, social media, market data, packet data and other sources.

Don Simpson (CTO Psydex) will lead a team that will provide:

- Innovation, Architecture, Design & Development
- Technical plan and development of time-critical insight algorithms
- Technical plan and development of software architecture

Don Simpson is a pragmatic innovator and engineering guru in virtually all areas of information management. He co-Founded KnowledgeX with Usey and served in numerous leadership roles in IBM Software Group and Global Services after IBM acquired the company. Simpson has a track record of leadership, execution and delivery of large scale software solutions, holding lead technical positions on numerous Government Intelligence programs and commercial projects.

Simpson's passion for prediction drives his innovation at the intersection of computational sciences, language and financial markets. Simpson holds multiple patents and is an expert in high-performance text analytics and semantic algorithms. Simpson has a degree in Computer Engineering.

4.4 Natural Language Processing Team

Dr. Shlomo Argamon, PhD (Senior Fellow, CADS) will lead a team that will provide:

- Computational Linguistics Subject Matter Expertise
- Opinion, Metaphor, and Stylistics Subject Matter Expertise

Dr. Shlomo Argamon is an internationally known expert on developing effective computational methods to analyze natural language style and authorship. His research focuses on developing computational methods for style-based analysis of natural language using machine learning and shallow lexical semantic representations. His work has included applications in intelligence analysis, forensic linguistics, biomedical informatics, and humanities scholarship. Dr. Argamon's current work includes projects on extracting structured representations of opinions from text, authorship attribution and profiling of anonymous texts, and evidence-based analysis of the medical literature. His group focuses on developing better formalizations of functional linguistic structures, clear semantic representations of knowledge to be extracted from such structures, and rigorous methods of developing annotated corpora to support such research. He is particularly interested in elucidating the relationships among linguistic structures, individual cognition and reasoning, and social context. Dr. Argamon holds degrees from Carnegie-Mellon and Yale Universities, and has been a Fulbright Postdoctoral Fellow and a Fannie and John Hertz Graduate Fellow. His research has been funded by NSF, IC-ARDA, BSF, and IBM.

Sergey Kanareykin (CTO, BMN) will lead a team that will provide:

- Cognitive Modeling
- Natural Language Processing and Cultural Parsing Integration

Sergey Kanareykin is the Chief Technical Officer of Behavioral Media Networks, Inc., responsible for project development and research direction at the company. As the CEO of Make a Mind Company, a tech incubator in Cambridge, MA, he is also involved in early stage financing for cognitive informatics startups. His overall experience includes management, engineering, and investment projects related to information technology solutions in a wide range of areas.

In 2004-2007, he served as the Chief Technical Officer (CTO) of the Center for Advanced Defense Studies (CADS), responsible for all development and was instrumental in establishing the Center's technology development in the fields of information sharing and security, with funding from the Department of Justice, Department of Defense, and intelligence community, as well as private industry, such as Intel Corporation and Harris Corporation. He was instrumental in the development of the Center's Advanced XML Security Lab (AXSL), in collaboration with Sarvega, Inc., and later Intel Corporation.

Sergey received his bachelor's degree in International Relations from the Saint-Petersburg State University (Russia), where he did research on the influence of the Internet on world economics and international affairs. He represented Russian companies during international fairs and technology demonstrations. He received a bachelor's degree in Computer Science from Denison University, and a master's degree in Computer Science with concentration in Computer Security and Information Assurance from The George Washington University. He is currently a doctoral student at the University of Paris 1, Pantheon-Sorbonne (France).

4.5 Validation, Verification and Evaluation Team

Kevin Thomas, PhD, MBA (BUSM, Director Societal and Behavioral Laboratory) will lead a team that will provide:

- Validation, Verification, and Evaluation
- Performance Metrics
- Methodological Review
- Neurological Assessments
- Network Intelligence Subject Matter Expertise

Dr. Kevin Thomas was the founding Research Programs Director for the Critical Infrastructure Protection Program, an over $20 million dollar research program in developing and analyzing methods of critical infrastructure protection and cyber security. He provided research and project oversight for over 50 research activities, conducted throughout a consortium of over 14 universities, into critical infrastructures (e.g., electrical distribution, internet infrastructures, telecommunications, banking, etc.)

Working directly with the Department of Homeland Security, National Capital Region Coordinator, and the Assistant to the Governor for Commonwealth Preparedness for the State of Virginia, and the Director of the Maryland Emergency Management Agency (MEMA), Thomas developed a multi-year three million dollar research program for assessing vulnerability for the National Capital Region. This research activity involved over 40 research faculty from 7 universities evaluating methods of vulnerability assessments and preparedness for this region.

Thomas served as Program Manager and Lead Researcher for certification of the National Infrastructure Simulation and Analysis Center (NISAC) Peer Review, as required under the Patriot Act, for the Department of Homeland Security. This effort involved site visits at Los Alamos and Sandia National Laboratories and review of technical and program management activities of these activities. The subsequent report of findings was used by DHS in restructuring program activities. Thomas also serves as Principle Investigator and Lab Director for the Behavioral and Societal Dynamics Laboratory. This laboratory studies human behavioral and societal dynamics of individuals and groups involved in complex adaptive environments. This Laboratory is partnered with the Neuroscience Interdisciplinary Modeling and Simulations Laboratory in studying human situational awareness and educational neurobiology.

5. Capabilities

5.1 Center for Advanced Defense Studies (CADS)

CADS has conducted academic and sponsored research on: (a) Automated analysis of writing style, including authenticating authorship, profiling document authors by demographics and psychographics, forensic linguistics, and sociolinguistics of written documents; (b) Developing intelligent computer systems that find and use patterns in large document collections, including applications in intelligent text retrieval, categorization, and summarization; (c) Application of machine learning and natural language processing to social media and literary textual analysis and visualization for counterterrorism and humanities scholarship; (d) Automated linguistic analysis of digital documents, particularly systemic functional linguistics, shallow parsing, and discourse structure; (e) Subject Matter Expertise in Strategic Communications, Physical Threats, Cognitive Linguistics, Cognitive Modeling and Analytics, Red Teaming, Cultural Intelligence, and Dynamic Network Analysis.

5.2 Psydex Corporation

Psydex has extensive experience designing and developing massively scalable data mining systems for some of the largest government and commercial entities in the world. Our expertise is in real time data mining and predictive analytics across to high velocity, high variety, and high volume data. Psydex helped pioneer social networking in the mid 90's and holds multiple patents in semantics.

Existing Psydex intellectual property includes AG (Analytics Grid) and related technology assets for indexing, mining and analyzing unstructured data.

5.3 Behavioral Media Networks, Inc. (BMN)

Behavioral Media Networks (BMN) is a spin-off of the Mind Machine Project (MMP) at MIT, and is managed by the team of original MMP founders. It focuses on applications of cognitive modeling and discourse analysis to social / media networks. BMN works closely with research groups at MIT, University of Oxford (UK), University of Cambridge (UK), and Boston University, developing custom software for specialized applications.

Members of the extended team of BMN contributed to such developments as the START parser (used in the Watson computer), Open Mind knowledge base project at MIT (Henry Lieberman), Radicalization Watch Program (Mathieu Guidere, meme tracking on the Web, US-France collaboration project funded by Fulbright). In 2006-2007, a technical team led by CTO Sergey Kanareykin developed models for Diplomatic, Information, Military and Economic (DIME) operations, building predictive capability for key non-military terrain characteristics available to tactical commander.

BMN's existing intellectual property includes licenses to two patent applications in the area of intention awareness and one patent application in the area of cognitive modeling based on discourse analysis.

5.4 Boston University School of Medicine (BU)

As a basic science research institution, and through various previous funded research activities, BU has performed testing and measurements for validation of computational system algorithms, design and performance of research study protocols for validation of automated processing systems.

BU has developed for the NIH road map project methods and practices in Neuroscience Interdisciplinary Modeling and Simulation that translate diverse mathematical, computational, and cognitive dynamics into experimental modeling systems to validate and test hypothesis, calculations and parametric data.

BU has specialized facilities in the form of computing capacity that may be used by this effort to execute system runs.

6. Statement of Work

Period 1: 12/15/11-9/30/12

1. Semantics and OMCS **(CADS)**
    1.1. Perform cross-domain inference between semantic structures and general commonsense knowledge to produce and expand Open Mind Common Sense (OMCS) database. **Deliverable:** Report of updates to OMCS.
    1.2. Use Open Mind Common Sense (OMCS) to construct a semantic network from assertions in natural language by using a variety of linguistic patterns.
2. System Setup **(Psydex)**
    2.1. Network
        2.1.1. Procure Firewall, Router, Switches for 10GigE Network.
    2.2. Servers
        2.2.1. Server Categories: Database, Query, Analytics, Applications.
    2.3. Storage
        2.3.1. High Performance Storage – need to size based on ingest requirements and retention policy.
3. Cognitive Model Development **(BMN)**
    3.1. Implement 3-layer model with situation awareness, perceived, and ontological layers.
    3.2. Implement software agents for population and updates.
4. Validate Graph Analytic Calculations, Algorithms, and Measurements **(BU)**
    4.1. *This task will take place on a university campus.*
    4.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.
5. Validate Rules Process and Rules Sets for Nodes and Nodal Community **(BU)**
    5.1. *This task will take place on a university campus.*
    5.2. Validate Rules Process and Rules Sets for Nodes or Nodal Community Properties and Attributes. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.
6. Data Management **(Psydex)**
    6.1. Unstructured Data Management
        6.1.1. Manage content – messages, conversations, documents/press releases, blog pages. Manage multiple versions, where appropriate, over time. Content management must be extensible, scalable and language independent.
    6.2. Structured Data Management
        6.2.1. Extensible model for building and associating metadata with content. Metadata management must be extensible, scalable and language independent. Relationships/Graphs network data model design and implementation will be part of the Metadata Management task.
    6.3. Taxonomy/Model Development
        6.3.1. Develop taxonomy for metadata. Design / Integrate system and user interfaces for managing taxonomy. Populate taxonomy with types, entities, attributes and relationships.
7. Culturally Agnostic Persuasion Model **(CADS)**

7.1. Study persuasion and influence campaigns to produce a model which identifies common target, approach, trust, and timing factors in successful persuasion campaigns. Verify this qualitative study with targeted assays to produce quantitative evidence. **Deliverable:** White Paper case study and supporting data

7.2. Identify power rules for the relationship between target, source credibility, timeliness, and message accuracy. **Deliverable:** Power Rules.

8. Axiology Development **(BMN)**

8.1. Adapt the Mind Default Axiology research results and model to individual and group discourse in social media.

8.2. Develop efficient structures and browsing and lookup of axiological information.

9. LXIO / Natural language processing integration **(BMN)**

9.1. Set up an individual processing node for START parser (or similar).

9.2. Adapt the START parser output for LXIO input requirements.

9.2.1. Implement time-based / tense-based processing module.

9.2.2. Implement intrinsic value module.

10. Perform Random Sampling Analysis of Graph to Graph Outputs **(BU)**

10.1. *This task will take place on a university campus.*

10.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.

11. Concept Production **(CADS)**

11.1. Create and/or implement mechanism to produce concepts from parsing natural language statements of commonsense knowledge.

11.2. Create and/or implement matrix of concepts vs. features (an assertion minus a concept) so the cells of the matrix represent all possible assertions about the concepts and features being studied.

12. Data Collection **(Psydex)**

12.1. Collector Framework

12.1.1. Develop software framework that provides base functionality for all collectors. Service-oriented approach for distributed collection, job execution, job management (e.g. start, stop, pause).

12.2. Data Feed Procurement and Configuration

12.2.1. Integrate with external interfaces (APIs). Procure data feeds where appropriate. Configure Feeds. Develop custom wrappers/adapters for consumption of 3rd party data dumps or ingest from external interfaces.

12.3. Social Media Collector Development

12.3.1. Develop custom collector for Twitter Stream API. Plan for scaling up to Twitter's Firehose (Full Feed). Develop custom collectors for popular social media sites, such as Facebook, Delicious, Digg, Flickr, Google Buzz, Hulu, MySpace, Tumblr, WordPress, Yahoo, YouTube, etc.

12.4. Web Data Collector Development

12.4.1. Web Site/Page, Blog crawler and collector. Support at a minimum RSS 1.0/2.0, ATOM and Custom Page Handlers. Develop/integrate ability to manage sites to crawl and metadata for each page, such as maximum crawl depth, scheduling, etc.

12.5. Mainstream Media Collector Development

12.5.1. Develop collectors for major news wires, such as AP, Dow Jones, Reuters, Business Wire, PR Newswire. Develop collectors for national TV, such as CNN Headline News, MSNBC, Fox News, etc.

12.6. Pattern Matching

12.6.1. Identify ideas and concepts (memes) that may be in a new message.

12.6.2. Identify topics that may match by determining that part or all of a model is related to content.

12.6.3. Confirm potential matches by executing model against content.

12.7. Transformation and Enrichment

12.7.1. Transform and normalize content and metadata. Add additional metadata and content using enrichment techniques.

12.8. Message Persistence

12.8.1. Barcode and store messages.

12.9. Relationship Persistence

12.9.1. Extract, transform and load relationship oriented data to persistent store. Relationships for social media messages will be the participants from, to, followers, following. Related Hashtags, Emoticons, Links, User Mentions, etc.

13. KB Testing **(BMN)**

13.1. Test the pluggable architecture for multiple knowledge bases.

13.2. Load and test default axiologies.

13.3. Load and test cultural overlays.

14. OMCS Testing **(BMN)**

14.1. Test the APIs for OMCS access (joint work with CADS expected).

14.2. Test the 2-way population mechanisms with OMCS models.

14.3. Test lookup and browsing efficiencies.

Period 2: 10/1/12-9/30/13

15. Indexing **(Psydex)**

15.1. Indexing at token level across Temporal and Geospatial dimensions. Must support ability to change Taxonomy independent of indexing. Must support ability to perform temporal analysis, proximity, Boolean logic, sets or groups, graph-style queries.

15.1.1. Index Source Adapter Framework

15.1.1.1. Develop software framework that provides base functionality for all source adapters. Need core capabilities like parse, normalize, remove stop words, etc.

15.1.2. Index Source Adapter Development

15.1.2.1. Develop source adapters that are capable of transforming and parsing source specific content and metadata.

15.1.3. Message Index Builder

15.1.3.1. Associate tokens with normalized timeslots, metadata, geospatial information, and content. Indexes must support on disk, in memory and hybrid.

15.1.4. Relationship Index Builder

15.1.4.1. Index relationships associated with messages and between metadata. This will include participants (followers, following), hash tags, user mentions, emoticons, referenced links, secondary relationships like URL-GEO, URL-Assignee, etc.

15.1.5. Index Management

15.1.5.1. System for managing distributed indexes. Indexes may be loaded in memory or on disk, or both across many machines. Knowing where shards / indexes are loaded or available is extremely important in a distributed indexing system.

16. Persuasion and Influence Measurement **(CADS)**
    16.1. Identify, measure, and perform persuasion processes based on attributes of nodes and nodal communities to counter persuasion campaigns. **Deliverable:** Case Study.
    16.2. Identify, measure, and perform influence operations based on attributes of nodes and nodal communities to co-opt persuasion memes. **Deliverable:** Case Study.
    16.3. Identify any correlation, in phase or out of phase, between two or more memes or nodal communities to identify cross-domain persuasion efforts. **Deliverable:** Case Study.
17. LXIO/OMCS/KB Integration **(BMN)**
    17.1. Develop and integrate components for information exchange with OMCS modules.
    17.2. Develop and integrate components for information exchange with multiple KB modules.
    17.3. Implement contextual value module.
    17.4. Implement consequent value module.
    17.5. Test the system with merged OMCS and KB data.
18. Validate Rule Sets, Algorithms, and Mathematical Calculations for Detection and Effect of Persuasion or Influence Activities **(BU)**
    18.1. *This task will take place on a university campus.*
    18.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.
19. Perform Random Sampling Topic Model Lifecycle Analysis for Consistency and Reliability and to Verify, Validate, and Evaluate Second Order Parametrics **(BU)**
    19.1. *This task will take place on a university campus.*
    19.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.
20. Perform system to Human Subject Matter Expert Comparative Analysis of Random Sampling of discrete Time Slot Topic Model Raw Source Media Data **(BU)**
    20.1. *This task will take place on a university campus.*
    20.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.
21. Distributed Query Service **(Psydex)**
    21.1. Query Management
        21.1.1. Develop component responsible for building and managing queries.
    21.2. Query Execution
        21.2.1. Develop component responsible for executing queries against indexes. Encapsulates the logic associated with executing query operators: proximity, keyword, conjunction, disjunction, negation, etc.
    21.3. Query Federator
        21.3.1. Federated, parallel execution of queries across many distributed indexes.
    21.4. Search and Retrieval
        21.4.1. Develop components responsible for searching and retrieving content, metadata, time series, relationship graphs, and other result sets.
22. Reasoning and Inference **(CADS)**

22.1. Create and/or implement mechanism to perform reasoning by reducing the dimensionality of a space, using the mathematical techniques of Principal Component Analysis.

22.2. Create and/or implement mechanism to perform cross-domain inference, reasoning jointly across more than one independently created knowledge base, even if the knowledge bases may not initially completely agree on their vocabulary, or may not completely agree on certain assertions.

22.3. Employ these reasoning mechanisms in a unified organizational system.

23. Core Services **(Psydex)**

23.1. Auditing, Authentication & Logging

23.1.1. Basic AAA service. Distributed logging so we can centrally manage system.

23.2. Statistics Service

23.2.1. Develop and integrate components responsible for calculating statistics. Calculate cumulative statistics (e.g. mean, standard deviation) for given time series. Service must operate in near real-time and run in a pool with multiple instances.

23.3. Correlation Service

23.3.1. Develop and integrate components responsible for correlating 2 or more time series. Service must operate in near real-time and run in a pool with multiple instances. Must support phase shifting so we can detect leaders and laggers.

23.4. Messaging Services

23.4.1. Need high performance messaging services. Queues and Topics must be supported. Need high-performance solution.

24. Services Layer Testing and Integration **(BMN)**

24.1. Develop and integrate components for populating the system with real or inferred profiles for individual nodes and groups of nodes, coming from the Services layer (Psydex).

24.2. Develop and integrate components for populating the nodes' cognitive model layers.

24.3. Test the system with the live data coming from Psydex modules.

25. Meme Analysis **(CADS)**

25.1. Identify propagation delay, hijacking of memes, and replacement memes to define the behavior of competing memes. **Deliverable:** Case Study.

25.2. Measure the propagation rate of memes and impact on trust of nodes and nodal communities to enable assessment of campaign impact. **Deliverable:** Case Study.

25.3. Measure how much can be propagated through a given medium to determine impact of medium on persuasion and influence memes. **Deliverable:** Case Study.

26. Core Development I **(BMN)**

26.1. Continued work on the cognitive model

26.2. Continued work on the axiological and KB layers

27. Text Analytics **(Psydex)**

27.1. Natural Language Processing

27.1.1. Integrate MIT Watson Parser. Supplied by Behavioral Media Networks.

27.2. Natural Language Processing

27.2.1. Integrate Illinois Institute of Technologies' NLP. Supplied by Illinois Institute of Technology.

27.3. Natural Language/Dialect Detection

27.4. Parts of Speech Identification

27.5. Named Entity Extraction

27.5.1. Develop/integrate component responsible for named entity extraction. Will be language dependent.

27.6. Named Entity Identification

27.6.1. Develop component responsible for identifying Named Entities in our taxonomy.

27.7. Relationship Extraction

27.8. Summarizer / Gister

27.9. Metadata Extraction

27.9.1. Develop and integrate components responsible for extracting metadata, such as Author, Source, and Time from text.

27.10. Markup Service

27.10.1. Develop and integrate components responsible for annotating text.

28. Offensive Capabilities **(CADS)**

28.1. Identify optimal centricity for meme injection to facilitate effective influence operations. This will impact the algorithms developed for trust, target, and credibility. **Deliverable:** Case Study.

Period 3: 10/1/13-9/30/14

29. Create Misinformation-Direction Mechanism **(CADS)**

29.1. Create and/or implement mechanism to mitigate deliberate misdirection and misinformation.

30. Analytics **(Psydex)**

30.1. Analytics Framework

30.1.1. Develop software framework that provides base functionality for all analytics/algorithms. Should be able to write native algorithms as well as execute external processes.

30.2. Meme Trend Analysis

30.2.1. Known Memes and Topics are analyzed over time and analyzed for patterns and unusual chatter levels.

30.3. Meme Discovery & Analysis / Clustering

30.3.1. Memes may co-occur or be part of a theme.

30.4. Meme Discovery & Analysis / Correlation

30.4.1. Memes may be related to one or more other memes.

30.5. Meme Discovery & Analysis / Intention Awareness

30.5.1. Develop meme discovery method in coordination with CADS and BMN intent models.

30.6. Meme Discovery & Analysis / Axiology

30.6.1. Integrate BMN axiology.

30.7. Meme Actor Discovery

30.7.1. Blogs, RSS Feeds, Twitter Accounts, Facebook Accounts, Youtube Accounts associated with Meme spreading. Look for related actors to identify loosely affiliated groups.

30.8. Meme Target Discovery

30.8.1. Re-Tweets on Twitter; Links to Blogs, RSS; Like/Comments on Facebook. Attempt to codify as sheep/targets and meme propagators.

30.9. Meme Impact Analysis / Real World Event Cause & Effect

30.9.1. Determine real world events that may be the start of a meme or real world events that may be a result of (e.g. uprising in Tunisia).

30.10. Meme Impact Analysis / Social Media Cause & Effect

30.10.1. Track meme propagation over time and networks / groups.

30.11. Meme Impact Analysis / Mainstream Media Cause & Effect

30.11.1. Impact on main stream media coverage and how main stream influences social media. Mainstream media has a limited pipe for distribution, so assume higher-priority items are disseminated.

31. Validate Rule Sets, Algorithms, and Mathematical Calculations of Persuasion or Influence to Measure Node or Nodal Community Capacity and Behaviors **(BU)**

31.1. *This task will take place on a university campus.*

31.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.

32. Perform Random Sampling Topic Model Lifecycle Analysis for Consistency and Reliability and to Verify, Validate, and Evaluate Second Order Parametrics **(BU)**

32.1. *This task will take place on a university campus.*

32.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.

33. Perform system to Human Subject Matter Expert Comparative Analysis of Random Sampling of discrete Time Slot Topic Model Raw Source Media Data **(BU)**

33.1. *This task will take place on a university campus.*

33.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.

34. Testing/Debug I **(BMN)**

34.1. Comprehensive testing of cognitive model output using modules implemented to date:

34.1.1. With recorded datastreams / known expected outputs.

34.1.2. With live manipulated data / unknown sequence.

35. Truthfulness and Values Inference **(CADS)**

35.1. Create and/or implement mechanism to produce historical truthfulness profiles for individuals and topics, and to build social network participant profiles to measure credibility and influence.

35.2. Use historical truthfulness profiles to build a value system for the node to determine the intent of that node.

36. Offensive Capabilities **(Psydex)**

36.1. Meme Development / Offensive

36.1.1. Work with CADS behavioral scientists and information operations SMEs to determine target audience, anticipated reaction, and potential side-effects of memes.

36.2. Propaganda Execution / Offensive

36.2.1. Twitter BOTS, RSS BOTS, BLOG BOTS, Paid Search BOTS.

36.3. Honeypot Development / Offensive

36.3.1. Develop sites to draw in and identify actors and targets.

36.4. SPAM Development / Offensive

36.4.1. Build spam fishing techniques to draw in actors and targets or run offensives that expose members and their identities – potentially leading to email and social media account associations.

37. Core Development II **(BMN)**

37.1. Addressing the issues identified in the testing phase.

38. Machine Learning **(CADS)**

38.1. Learn and develop memory of historical changes to the graph structure based on persuasion and influence experimentation to allow the system to become more adept at identifying and analyzing persuasion campaigns.

39. Visualization and GUI Development **(Psydex)**

39.1. Portal Development

39.2. Timeseries Charts

39.3. Timeline Visualization

39.4. Relationship Visualization

39.5. Geospatial Visualization

39.6. Collaboration

40. Testing/Debug II **(BMN)**

40.1. Test entire system functioning with updates from Core Development II phase using the TA 3 Performer's test environment.

41. Test Automated Persuasion and Influence Detection Results **(CADS)**

41.1. Using Information Operations SMEs, create and/or implement procedure to conduct testing and validation against a cross-cultural body of individuals to ensure that the basic algorithm is sufficiently broad (significant collaboration with BU team and TA 3 Performer is anticipated).

Period 4: 10/1/14-1/15/15

42. Final Integration **(BMN)**

43. Create Mechanism to Provide Context for Test Environment **(CADS)**

43.1. Create and/or implement system that will allow the introduction of other media sources in future versions to allow for better granularity of indirect inputs.

44. Final Integration **(Psydex, CADS)**

45. Final Aggregative System Performance Analysis **(BU)**

45.1. *This task will take place on a university campus.*

45.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.

46. Longitudinal Summary and Analysis of Tests and Measurements **(BU)**

46.1. *This task will take place on a university campus.*

46.2. **Deliverable:** Report of Findings, corrective actions feedback to Software Team.

47. Project Presentations **(BMN, Psydex, CADS, BU)**

7. Schedule and Milestones

Figure 7.1: Schedule and Milestones

Phase	Month(s)	Task/Milestone	Key Teams	Cost
Period 1	1-3	Semantics and OMCS	CADS	173,050.27
	1-3	System Setup	Psydex	461,919.15
	1-6	Cognitive Model Development	BMN	137,947.50
	1-7	Validate Graph Analytic Calculations, Algorithms, and Measurements	BU	109,740.67
	3	**KICKOFF**	**ALL**	

	3-10	Validate Rules Process and Rules Sets for Nodes and Nodal Community	BU	88,706.26
	4-6	Data Management	Psydex	506,142.00
	4-6	Culturally Agnostic Persuasion Model Development	CADS	191,574.25
	5-8	Axiology Development	BMN	53,960.50
	5-8	LXIO/Natural Language Processing Integration	BMN	53,960.50
	6-10	Perform Random Sampling Analysis of Graph to Graph Outputs	BU	51,652.11
	7-9	Concept Production	CADS	255,852.33
	7-10	Data Collection	Psydex	672,989.33
	9-10	KB Testing	BMN	28,056.30
	9-10	OMCS Testing	BMN	28,056.30

Period 2	11-12	Indexing	Psydex	339,294.67
	11-13	Persuasion and Influence Measurement	CADS	202,774.25
	11-16	LXIO/OMCS/KB Integration	BMN	172,343.50
	11-22	Validate Rule Sets, Algorithms, and Mathematical Calculations for Detection and Effect of Persuasion or Influence Activities	BU	99,744.40
	11-22	Perform Random Sampling Topic Model Lifecycle Analysis for Consistency and Reliability and to Verify, Validate, and Evaluate Second Order Parametrics	BU	99,744.40
	11-22	Perform METSYS to Human Subject Matter Expert Comparative Analysis of Random Sampling of discrete Time Slot Topic Model Raw Source Media Data	BU	99,744.40
	12	**MILESTONE 1 - VV&T**	**ALL**	
	13-14	Distributed Query Service	Psydex	333,694.67
	14-16	Reasoning and Inference Mechanisms	CADS	191,994.25
	15-17	Core Services	Psydex	506,142.00
	15-19	Services Layer Testing and Integration	BMN	80,651.75
	17-19	Meme Analysis	CADS	191,574.25
	21	**MILESTONE 2 - VV&T**	**ALL**	
	17-24	Core Development I	BMN	216,294.75
	18-22	Text Analytics	Psydex	845,436.67
	20-22	Offensive Capabilities	CADS	197,174.25
	22	**MIDTERM EXAM: FOCUSED SCALING/TRANSITION**	**ALL**	

Period 3	23-25	Create Misinformation/Misdirection Detection Mechanism	CADS	202,262.55
	23-25	Analytics	Psydex	518,026.60
	23-34	Validate Rule Sets, Algorithms, and Mathematical Calculations of Persuasion or Influence to Measure Node or Nodal Community Capacity and Behaviors	BU	101,187.33

	23-34	Perform Random Sampling Topic Model Lifecycle Analysis for Consistency and Reliability and to Verify, Validate, and Evaluate Second Order Parametrics	BU	101,187.33
	23-34	Perform system to Human Subject Matter Expert Comparative Analysis of Random Sampling of discrete Time Slot Topic Model Raw Source Media Data	BU	101,187.33
	25-27	Testing/Debug I	BMN	99,122.40
	26-28	Truthfulness and Values Inference	CADS	196,242.55
	26-28	Offensive Capabilities	Psydex	518,026.60
	28-30	Core Development II	BMN	102,962.40
	29-31	Machine Learning	CADS	201,842.55
	29-34	Visualization and GUI Development	Psydex	1,036,053.20
	30	**MILESTONE 3 - VV&T**	**ALL**	
	31-33	Testing/Debug II	BMN	99,122.40
	32-34	Test Automated Persuasion and Influence Detection Results Using Information Operations SMEs	CADS	196,242.55

Period 4	34-37	Final Integration	BMN	151,263.30
	35-36	Create Mechanism to Provide Context for Test Environment	CADS	147,587.30
	35-38	Final Integration	Psydex, CADS	822,590.18
	35-38	Final Aggregative System Performance Analysis	BU	102,268.37
	35-38	Longitudinal Summary and Analysis of Tests and Measurements	BU	102,268.37
	37-38	Project Presentations	BMN	73,333.50
	39	**MILESTONE 4 - FINAL EVALUATION**	**ALL**	
			TOTAL	**11,262,992.48**

8. Cost Summary

Figure 8.1: Cost Summary

	1st period **12/15/2011** **9/30/2012**	**2nd period** **10/1/2012** **9/30/2013**	**3rd period** **10/1/2013** **9/30/2014**	**4th period** **10/1/2014** **1/15/2015**	**TOTAL**
A. LABOR					
TOTAL LABOR	351,281	443,723	454,817	150,086	**1,399,907**
B. FRINGE BENEFITS					
TOTAL FRINGE	70,927	89,592	91,836	30,303	**282,658**
C. TRAVEL					
TOTAL TRAVEL	14,000	20,000	16,000	8,000	**58,000**
D.					

SUPPLIES					
TOTAL SUPPLIES	1,500	1,800	1,800	600	**5,700**
E. EQUIPMENT					
TOTAL EQUIPMENT	5,000	4,000	4,000	0	**13,000**
F. OTHER DIRECT COSTS					
TOTAL OTHER	490	540	540	370	**1,940**
G. SUBCONTRACTED COSTS					
1. Psydex	1,641,050	2,024,568	2,072,106	705,075	**6,442,799**
2. BMN	301,981	402,008	401,530	191,556	**1,297,075**
3. Boston University	250,099	299,233	303,562	204,537	**1,057,431**
TOTAL SUBCONTRACTED COSTS	2,193,131	2,725,810	2,777,198	1,101,167	**8,797,305**
TOTAL DIRECT COSTS	2,636,328	3,285,465	3,346,191	1,290,527	**10,558,510**
H. INDIRECT COSTS					
40% of Modified Total Direct Costs	177,279	223,862	227,597	75,744	**704,482**
TOTAL COSTS	**2,813,607**	**3,509,327**	**3,573,788**	**1,366,270**	**11,262,992**

Percent of Cost Sharing 0%

Budget Prepared 9/10/11

9. References

[1] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, “Lexical Predictors of Personality Type,” In *Proc. 2005 Conference Classification Society of North America*, St. Louis, MO, June 2005.

[2] S. Argamon, J. Dodick, and P. Chase. Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics*, vol. 75 no. 2, pp. 203-238, 2008.

[3] S. Argamon, et al. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Sciences and Technology*, vol. 58 no. 6, April 2007.

[4] J. A. Goguen, D. F. Harrell, “Style: A Computational and Conceptual Blending- Based Approach,” In: Argamon, Burns, Dubnov, eds. *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*. Springer (2010).

[5] D. Casasanto, When is a linguistic metaphor a conceptual metaphor? *New directions in cognitive linguistics*, vol. 24, pp. 127-146, 2009.

[6] A. Anand, S. Bedathur, K. Berberich and R. Schenkel. Temporal Index Sharding for Space-Time Efficiency in Archive Search. Technical Report MPI-I-2011-5- 001, Max-Planck Institute for Informatics, 2010.

[7] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. SIGIR Forum, 41(2):35--41, 2007.

[8] A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. Efficient Temporal Keyword Search over Versioned Text. In CIKM, 2010.

[9] Benyah Shaparenko, Rich Caruana, Johannes Gehrke, and Thorsten Joachims. "Identifying Temporal Patterns and Key Players in Document Collections". *In Proc. of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05)*, 165--174, 2005.

[10] Klaus Berberich , Srikanta Bedathur , Thomas Neumann, Gerhard Weikum, A time machine for text search, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, July 23-27, 2007, Amsterdam, The Netherlands [doi>10.1145/1277741.1277831]

[11] Klaus Berberich , Srikanta Bedathur, Thomas Neumann , Gerhard Weikum, FluxCapacitor: efficient time-travel text search, Proceedings of the 33rd international conference on Very large data bases, September 23-27, 2007, Vienna, Austria

[12] Rob Usey, Don Simpson, “SYSTEMS AND METHODS FOR PERFORMING SEMANTIC ANALYSIS OF INFORMATION OVER TIME AND SPACE”, U.S. Patent 20080208820, Filed Aug. 28, 2008.

[13] Rob Usey, Don Simpson, “Psydex Time Critical Insight”, white paper, Psydex Corp., 2010

[14] Newton Howard, Mathieu Guidere “Lexical Input/output System (LXIO) design and implementation”, MMP technical report, MIT 2011

[15] Newton Howard, Mathieu Guidere “LXIO: The Mood Detection Robopsych”, MMP technical report, MIT 2011

[16] Newton Howard, “Computational Methods for Clinical Applications: An Introduction”, Journal of Functional Neurology, Rehabilitation, and Ergonomics, Spring 2011

[17] Newton Howard, "The Cognitive Construct of Intention: A Promising New Computational Methodology" Paper presented at the 31st annual meeting of the ISPP, May 23, 2009, Paris, France

[18] Newton Howard, Henry Lieberman, “BrainSpace: Relating Neuroscience to Knowledge About Everyday Life”, MMP technical report, MIT 2011

[19] D.J. O’Keefe, "Theories of Persuasion" in *The SAGE Handbook Of Media Processes and Effects*, Los Angeles, CA, SAGE, 2009, pp. 277-78

[20] R.E Petty and D.T. Wegener, “The Elaboration Likelihood Model: Current Status and Controversies” in *Dual Process Theories in Social Psychology,* S. Chaiken and Y. Trope eds., New York, NY, Guilford Press, 1999, pp. 41–72.

[21] L.McGaan (2011, Feb 16), *Persuasion Theory Review* [power point], Available: http://department.monm.edu/cata/mcgaan/classes/cata339/persuasion_theory_review.htm, accessed Sept 10, 2011.

[22] *Commander’s Handbook for Strategic Communication and Communication Strategy*, Version 3.0, US Joint Forces Command, Joint Warfighting Center, Suffolk, VA 2010.

[23] Iyengar, Sheena S. and Emir Kamenica. *Choice Overload and Simplicity Seeking.* February 6, 2007.

[24] *COSMOS project summary,* Technical report, DISA website accessed on Sept 10, 2011, http://www.disa.mil/news/pressresources/factsheets/aco.html

Appendix A

Team Member Identification
Center for Advanced Defense Studies (CADS)
Psydex Corporation
Behavioral Media Network (BMN)
Boston University (BU)

Government or FFRDC Team Member
NONE

Organizational Conflict of Interest Affirmations Disclosure
NONE

Intellectual Property

COMMERCIAL				
Technical Data Computer Software to be Furnished with Restrictions	Summary of Intended Use in Conduct of the Research	Basis for Assertion	Asserted Rights Category	Name of Person Asserting Restrictions
PSYDEX AG	Base platform for distributed indexing, query, statistics, and correlation. Collection and Ingest for many social media sites, including Twitter and Facebook, RSS and web feeds. Trend Analysis and UI capabilities. Offensive BOT capabilities.	High Performance Temporal Analysis of Topic Models	No Cost, Limited Use License for Program. Psydex maintains rights to intellectual property	Rob Usey

Human Use
NONE

Animal Use
NONE

Subcontractor Plan
Letter in lieu of Subcontractor Plan attached (pg. 42). In the event that a Subcontractor Plan is determined to be required, we will submit one at that time.

Center for Advanced
Defense Studies

Center for Advanced Defense Studies
10 G St. NE, Ste. 610
Washington, D.C. 20002

September 12, 2011

DARPA/I2O
ATTN: DARPA-BAA-11-64
3701 North Fairfax Drive
Arlington, VA 22203-1714

Dear Dr. Waltzman,

I am submitting this letter along with our DARPA-BAA-11-64 proposal in lieu of a Subcontractor Plan. The Center for Advanced Defense Studies is exempt from the Subcontractor Plan requirement outlined in DARPA-BAA-11-64 for the following reasons:

(1) The Center is submitting a grant proposal.

(2) The Center is not a large business concerns.

Please let me know if you require any additional information or verification.

Sincerely,

Farley Mesko
Operations Officer
Center for Advanced Defense Studies