THE DEPARTMENT OF DEFENSE (DOD)

SMALL BUSINESS INNOVATION RESEARCH (SBIR) PROGRAM

PROPOSAL TITLE: LAYERED DATA TO AREAS OF INTEREST

SUB-TITLE: GUN CONTROL NETWORK—CONTEXT ADAPTIVE PATTERN RECOGNITION WITH MULTI-SOURCE DATA

PROPOSING FIRM: AQUERRE TECHNOLOGIES LLC

PRINCIPAL INVESTIGATOR: Noah B. Jacobsen, Ph.D.

TECHNOLOGY AREAS: Information Systems

KEYWORDS: Pattern Recognition, Spatial Grid Analysis, Activity Detection, Data Features, Prediction, Cognitive Sciences

1 IDENTIFICATION AND SIGNIFICANCE OF THE PROBLEM OR OPPORTUNITY

"The objective of this research is to use spatial, temporal and graph analysis techniques to take very large data streams over wide areas and autonomously highlight areas of interest for a decision maker without a priori knowledge of the area and/or location of high value." –OSD13-LD3 Topic Description

Predictive information systems are an active area of research interest of the Department of Defense and have applications to U.S. government commercial interests and the interests of broader commercial markets [DOD, 2012], [U.S. Navy Fleet Cyber Command, 2013]. This research proposal addresses the objective statement quoted above for the test application of gun control in the United States. Gun violence in the U.S. is a severe problem. The ongoing situation contributes to a diverse set of open data (crime statistics historical data) published by federal, state, and local agencies, making gun control an attractive test application for the predictive information system developed in this SBIR project. We believe that the selected test application in no way limits the range of potential applications of our system and we anticipate a broad degree of flexibility pending sponsor need. A key target customer of the commercial product in development by Aquerre Technologies LLC is the U.S. Government and the DOD.

Aquerre Technologies is a Los Angeles based LLC, organized in May 2013. Our business objective is to advance the state of the art of communication and control systems technologies. Presently, the sole employee is the Principal Investigator. In addition to this SBIR proposal, we have pursued two other Phase I proposals on the topic of gun control applicable to the year 2014 (DHS, NSF). These proposals, each different, are variations on a central theme of developing and commercializing a new gun control technology.

1.1 BACKGROUND

A detailed development of the multi-source, multi-structure data fusion problem is given by [Antony and Karakowski, 2007]. Adaptation to dynamic environments based on natural and social-science methods is discussed in [Sagarin, *et al.*, 2010]. Statistical models are used to perform pattern recognition with military fatality data in [Johnson, *et al.*, 2011]. And [McGirr and L. Keenan, 2011] identify a need for context specific adaptation based on location and time.[1]

1.2 DESIGN APPROACH

In this proposal, we develop a predictive information system, or more generally, a query processing system, for the target application of gun control in the United States. To evaluate system performance we will introduce an error metric—*e.g.* the average squared error between an estimated and measured variable. The actual metric(s) of value is a question that we address in the Phase I project, and we propose to study a "Quality of Information" metric for identifying context relevant data. An "ideal system" is defined to be a system that is optimal with respect to a given metric. We further define an "efficient system" to be a system that reasonably approximates the performance of an ideal system while also satisfying cost (*e.g.* complexity) constraints. Our proposal is highlighted by the following features: (i) an adaptive factor graph approach for modeling system dynamics, (ii) application of the sum-product algorithm for statistical inference processing on graphs, (iii) a modified web-crawler for automated data-mining and information processing with graph-based data, and (iv) a 4-D multi-player virtual environment for pattern recognition and prediction applications. The details are discussed in the following sections.

1.2.1 INCLUSION OF MULTI-SOURCE DATA

We propose to develop a predictive information processing system based on multi-source and multi-structure data inputs. We claim that the inclusion of multiple data sources can only improve the performance of an ideal predictive information system. Where the addition of new data yields statistically-relevant information (for a given query and its context), we can expect greater performance gains than when the addition of new data yields no statistically-relevant information. This argument has implications to big-data set management in the sense that we may disregard "low-quality" data in order to

1 For brevity, we mention only a few of the many relevant background papers on the topic. Subsequent sections of this paper highlight additional details.

focus on the most relevant data for a given instance of a query. In this paper, we propose to adopt a layered data management strategy, in which different management layers represent different levels of data aggregation. We claim that the layered strategy developed in this proposal is advantageous for complexity and functionality simplification and is consistent with the framework developed in reference paper [Antony and Karakowski, 2007].

Table 1 lists local, state, and federal crime statistics statistics data reporting sources we will use as primary data sources for this Phase I project. In addition to the use of crime data as system inputs, this proposal includes the use of text data inputs from online news and information sources (*e.g.* Los Angeles Times archived articles), location and maps data from open sources (*e.g.* www.openstreetmap.org), and human cognitive input data (*e.g.* operator input data).

DATA SOURCE	URL
Los Angeles Police Department – Crime Map Tool and Archived Statistics	https://tinyurl.com/88jnveb
State of California – Department of Justice – Crimes/Clearances Statistics	https://tinyurl.com/mfqz3r9
State of New York – Division of Criminal Justice Services – Criminal Justice Statistics	https://tinyurl.com/73vekep
State of New York – Open Data – Index Crimes by County and Agency	https://tinyurl.com/mdmppjq
The District of Columbia – Metropolitan Police Department – Statistics & Data – Crime Map Tool and Archived Statistics	https://tinyurl.com/mhlsehy
U.S. Dept. of Justice – Federal Bureau of Investigation – Summary of Reported Crime by Locality, State, and Nation	http://www.ucrdatatool.gov/

TABLE 1. U.S. crime statistics data published on the internet.

A technical challenge that will be addressed in the Phase I study is the varied levels of aggregation of the available data. For example, each local, state and federal agency has its own standard of data reporting. As a result we have some agencies reporting statistics per locality-month (granular data set) and some agencies reporting statistics per state-year (coarse data set)--and various combinations in between. We address this challenge by adapting a hierarchical data model and an adaptive processing strategy.

The system model and algorithm development framework is described in Section 1.1.2. The strategy for data organization and visualization is described in Section 1.1.3. In Section 1.1.3, we further develop the data-mining and information discovery strategy. Additional details of the context adaptation and data aggregation strategy are provided in Sections 1.1.4-1.1.5.

1.2.2 SYSTEM MODEL AND INFORMATION PROCESSING STRATEGY

We assume that a graph-based model is used to describe the relationship between system inputs and outputs. The information processing approach developed in this proposal uses the underlying graphical model as a framework for developing predictive algorithms.

A “graph” is defined to be a set of nodes and a set of links connecting the nodes. We consider a specific family of graphs, known as “bipartite graphs”. Bipartite graphs are comprised of two sets of nodes, which we shall call “variable nodes” and “check nodes”. Furthermore the links of a bipartite graph may only connect a check node to a variable node. Check nodes have the interpretation of representing statistical relationships between the variable nodes to which they are linked. Both check and variable nodes may be either hidden or observed [Jacobsen and Soni, 2007]. In this paper predictive algorithms are developed to estimate of the values of hidden (unobserved) system variables.

In the reference paper [Antony and Karakowski, 2007], the variable nodes are “(entity, time, location)-nodes” and check nodes may be viewed as relationships between entity nodes. In this paper, we propose to apply a graph-theoretic approach for developing predictive algorithms using the same (entity, time, location) variable-node convention developed in the reference paper. The fundamental design challenge is to fit the model (a bipartite graph) to a (partial and imperfect) observed data set while simultaneously estimating the value of an unobserved variable. To address the challenge, we propose to use our own research experience [Jacobsen, *et al.*, 2007—2011], results in the field of inference processing on graphs [MacKay and Pelo, 1995], and results on factor graphs and the sum-product algorithm [Kschischang, *et al.*, 2001].

To illustrate the proposed graphical development framework, we introduce the following mathematical notation: let X denote an event of gun violence in region A during time period B and let Y denote the observed data (crime data, news reports, etc.). In practice, Y could represent a massive data set. Let $f(X, Y)$ denote the “global” joint Probability Distribution Function (PDF) of the variables X and Y. In general, the variables (X, Y) represent a large multi-variate set in which there exist statistically related and statistically unrelated variables. As a consequence, the global PDF $f(X, Y)$ factors into a product of component PDFs. Let us denote a factorization as:

$$f(X,Y)=\prod_{n=1}^{N} f(Z_n)$$

where Z_n denotes a subset of the variables (X,Y), $f(Z_n)$ denotes the joint PDF of the variables Z_n, and $n=1,\ldots,N$. The factorized global PDF can be represented by a "factor graph" [Kschischang, *et al.*, 2001]. The results summarized in [Kschischang, *et al.*, 2001] show that the global distribution $f(X,Y)$ can be computed using message passing algorithms (*i.e.* the sum-product algorithm) operating on a factor-graph corresponding to the factorized global PDF. For a given number of system variables, a factor-graph framework yields a simplification in processing complexity (less links between nodes) as compared to a general case model in which each system variable is assumed to be statistically related to every other system variable (most links between nodes). In practice, a factor-graph message-passing framework could be used to estimate any statistical quantity relating the variables X and Y. For example, the unconditional probability of the event X,

$$Pr(X)=\int f(X,Y)\,dY,$$

is readily estimated using outputs of the sum-product algorithm. Note that factorizing the global PDF $f(X,Y)$ represents a context adaptive process in the sense that the underlying model should be chosen to yield a reduced complexity implementation while still accurately estimating unknown system variables. Thus, we claim that finding a simple factorization of the global PDF leads to an efficient implementation of a big-data predictive information system.

A key assumption is that we can determine a simple factorized PDF. In practice, this must be accomplished by appealing to physical/logical arguments (*e.g.* [Johnson, *et al.*, 2011]), hypothesis evaluation (*e.g.* [Antony and Karakowski, 2007], assuming a statistical distribution over unknown system variables [MacKay and Pelo, 1995], or some other means. One approach, discussed in [Antony and Karakowski, 2007], is hypothesis evaluation. A human operator or an autonomous process might guess that some hypothesis is true or not based on the observation of a subset of the global data. The system might then evaluate the consistency of the hypothesis against other known or estimated variables. The "conjectured" data might then be included in an "observed" data set when determining a new predictive estimate. Thus, hypothesis evaluation is a possible means of discovering links (statistical relationships) between variables, and therefore is a possible approach to discovering a factorization for a global PDF. In addition, hypothesis evaluation is a means of discovering hidden variables—*e.g.* we might conjecture the existence of an *a priori* un-modeled system variable and then use message passing algorithms to evaluate the validity of the conjectured hypothesis.

In summary, we propose to apply research results in the field of statistical inference processing on graphs

[MacKay and Pelo, 1995], [Kschischang, *et al.*, 2001] and our own research experience in the field error control codes [Jacobsen, *et al.*, 2007—2011]. It is known that the application of the sum-product algorithm has enabled significant advances in the field of coding theory (error control codes) and the resulting technology is widely deployed in modern data and communication systems.

Context adaptive interpretation of factor graphs—Optimization of the graphical model for obtaining a context specific adaptation, and therefore an efficient algorithm implementation, is an important component of this proposal. In the proposed framework, information processing (*e.g.* prediction) is based on both an estimate of the system factorization as well as an estimate of the unobserved system variables. If a given factorization strategy yields a good predictor than we have implicitly adapted the system to a context by finding an efficient implementation. We further claim that the proposed dynamic factor graph approach offers a potential advantage of instance specific query adaptation.

We propose to evaluate the effectiveness of the proposed algorithms developed in Phase I based on a multi-source set of historical training data obtained from the internet. The Phase I project will focus on addressing a simple predictive query. Borrowing ideas from the reference paper [McGirr and Keenan, 2011], we suggest addressing the prediction of historical gun violence rates per U.S. City. We plan to demonstrate that context specific adaptation yields performance gains with the predictive algorithms developed in this project. In summary, by re-formulating the underlying graphical model as a function the context of interest (*e.g.* location-time of interest), we expect to obtain superior estimates of the predicted variables.

1.2.3 DATA ORGANIZATION AND VISUALIZATION STRATEGY

We propose to tackle the challenge of information processing with large multi-source/multi-structure data sets using a 4-dimension geo-spatial (three spatial coordinates plus a temporal coordinate) organization strategy. We claim that all data is intrinsically related to points or regions in space and time and therefore we can intuitively map any data to a 4-D space-time grid based on a location-time attribute. In this proposal, we leverage a geo-spatial data organization strategy to support a 4-D virtual environment for data display and information processing. In particular, we develop multi-player games to recognize patterns and discover information (data-mining) using a 4-D virtual environment representation of location-time dependent data.

We propose to develop multi-player games in which human and autonomous players traverse the dimensions of space and time. In the games, players might operate as individual agents or in teams, and in the same locality or across multiple localities. The abstract goal of the game is learning and specific goals depend on the application, such as prediction of events of gun violence based on a collection of multi-source data. We claim that a three-dimensional (location variables) simulation environment with a fourth-dimension component (time-navigation) is well-suited for information processing with location and time dependent data. We further claim that a multi-player simulation environment offers potential advantages for (a.) data organizing and visualization, (b.) data-mining and information discovery, and (c.) inclusion of human intelligence data. These claimed advantages, elaborated in the following subsections, will be evaluated as part of the research and development work plan for this SBIR proposal.

1.2.3.1 Data Organization and Visualization

Data that is characterized by a point in space and time is readily mapped to a 4-D grid representing the location and time coordinates. Multi-time and multi-location data are also readily accommodated by a 4-D grid framework, for example a 4-D cube defined by a set of points in space-time might correspond to an individual data point, or a collection of such cubes might correspond to an individual data point. Complicating issues of location and time ambiguity can be managed with a soft-hierarchical approach, as described in [Antony and Karakowski, 2007].

Once a collection of data is mapped to the 4-D grid, navigating interrelated data can be implemented intuitively by means of a continuous path across space and time, zooming in and out of the space and time coordinates, jump discontinuities in space-time, etc. Moreover, a geo-spatial data organization strategy directly supports the rendering of a virtual environment interface for human operator data visualization. Data visualization is an important component for the human operator. We claim that a geo-spatial data organization strategy paired with a 4-D virtual environment interface will lead to an efficient system implementation, and we plan to demonstrate the validity of our approach with research results from the Phase I project. In summary, the data organization and visualization strategy is based on geo-spatial organization of time-location dependent data and a 4-D virtual environment data rendering interface.

1.2.3.2 Data-Mining and Information Discovery

We claim the process of data-mining can be efficiently implemented from point of view of a multi-player game and the virtual environment described above—because the arrangement of the data on the grid is tied naturally to the way humans perceive information in the physical world. We propose to implement specific data-mining, data-aggregation, and context adaptation processes with specific instances of a multi-player game. For example, we might have a game in which teams of human and/or autonomous players search through large low-level data sets, looking for context relevant information for a decision maker at a higher level of data aggregation. A multi-player game framework can support centralized and/or de-centralized coordination, human and/or autonomous processing, multi-processing in space and time, advanced algorithm development, dynamic environment reconfiguration, and other interesting pattern recognition and information discovery strategies. The effectiveness of the proposed multi-player framework for information processing will be demonstrated in this SBIR project.

In summary, a multi-player virtual environment framework offers potential advantages for the processes of data-mining and information discovery, in addition to the implementation of a layered data management strategy. In this SBIR project we will evaluate the merit of the proposed game-based strategy as a component of information processing systems. Ultimately, we seek to commercialize a new technology based on the research and development results generated in the Phase I and Phase II projects.

1.2.3.3 Human Cognitive Data

The multi-player game approach proposed in this paper offers potential advantages to the process of human cognitive data input. Humans are an integral feedback component of any predictive information system because we can easily correct for many kinds of errors, inconsistencies, contradictions, etc. generated by autonomous processes [Antony and Karakowski, 2007]. We claim that a 4-D based virtual representation of location-time dependent data facilitates human cognitive data inputs due to its intuitive advantages for data organization, visualization and user interface. We suppose, for example, that a human operated player uncovers a link between data encountered at different points during a game. Such a discovery might be informed by human experience and human cognitive reasoning, and therefore constitutes a non-autonomous system process. Human-cognition based data is again readily mapped to a geo-spatial organizational structure with its corresponding location-time attribute. An advantage of using a 4-D virtual environment multi-player game framework for semi-autonomous information processing is that it offers an intuitive approach for the inclusion and exploitation of human-based cognitive reasoning data.

A key component of this SBIR proposal is to demonstrate the effectiveness of a game simulation environment as a design approach for a predictive information processing. The personnel participating on this proposal will include a software programming expert. This position is scheduled to be hired pending acceptance of this proposal. In summary, a game-simulation environment offers the potential for a successful predictive-information and data-mining framework in which players may base their decisions on an intuitive representation of the time-location dependent data.

1.2.4 CONTEXT ADAPTATION

Context adaptation is a prevalent theme throughout this work. Results in the reference papers [McGirr and Keenan, 2011] and [Johnson, *et al.*, 2011] show that predictive gains are obtainable from context specific system adaptation (*e.g.* time-location dependent model formulation). Finding the best model and estimator for the predicted values in a given context, environment, and/or application is clearly advantageous to the performance of the system. We have therefore focused on an adaptive approach to system function based on factor-graphs and the sum-product algorithm [Kschischang, *et al.*, 2011], described in Section 1.1.2. We propose the application of a dynamic factor graph model for the development of a predictive system. A key component of the context adaptive approach that we adopt in this paper is the development of a re-configurable graphical structure (finding an efficient factorization) for a given instance of the application. Due to complexity advantages of implementing algorithms on factor-graphs, we argue that the adaptive factor-graph approach is consistent with the goal of finding an efficient, practical implementation.

In addition to “context adaptive re-factorization” described above, this proposal includes the evaluation of an adaptive training scheme. Training is a requisite component of any predictive information system. We propose to adjust parameters of the predictive algorithms (*e.g.* coefficients of a tapped delay line filter) for a given location-time context based on selective emphasis of local and global training data sets. We claim that adaptive training may be beneficial to system performance. We first assume that time-location proximate data is the most relevant data for a time-location based query. In this case, we might weigh locally proximate data more heavily than non-proximate data when training the algorithms. For example, temperature forecasts in Chicago and Atlanta may both benefit from a global training data set (*e.g.* the global warming trend). We claim, however, that the predictive algorithm for Chicago will benefit by emphasizing Illinois state archive data over Georgia state archive data (*e.g.* record temperatures data).

We plan to use current and archived crime statistics data published by city, state, and federal agencies as a key data source for this project. We propose to use homicide statistics data, which is well reported across many agencies, as an approximation to the rate gun violence per reporting agency per reporting interval (it is possible to use a more sophisticated estimate for the rate of gun violence based on all reported crime data rather than sole reliance on homicide statistics). The goal of this study is develop an adaptive predictive information processing framework for identifying and predicting patterns of gun violence in the United States. An important technical contribution of this project is the adaptation to a diverse nature of available data. For example, some agencies publish statistics per month per locality (granular or “rich” data set), however, some agencies report aggregate data at the granularity of per locality per year (aggregate or “coarse” data set). In addition to predicting the future state of system variables, the system developed in this proposal can be used to mitigate the inequity of available data by making predictive estimates of historical data points. In particular new discoveries of historical entities or relationships (variables or checks on the bipartite graph) can be used to inform predictions of future variables.

1.2.5 LAYERED DATA AGGREGATION APPROACH

Performance gains are anticipated by incorporating a plurality of conventional and modern data sources, *e.g.* Human Intelligence (HUMINT), Signals Intelligence (SIGINT), and Twitter-sourced data. A key issue that arises in the multi-source large-data set problem is too much data and therefore source-data compression, pre-filtering, and aggregation are key functions of future intelligence systems. We propose to manage large data set issues with a layered data aggregation framework in which different layers of a “data stack” represent different levels of data aggregation, where the lowest layer is the closest to the raw source data and the highest layer represents the highest level of data aggregation.

Actually, the question of how best to aggregate data is a question of what data is most relevant for a given query and therefore, in order to obtain an efficient system implementation, the process of data-aggregation needs to be driven by the context. In the Phase I feasibility study we propose to evaluate the effectiveness of a quality-of-information (QOI) metric for determining the relevance of data sampled from a large set for a given application. The QOI metric is supposed to identify the most pertinent data and/or data-aggregate for passing up to the next level of the data stack. The algorithm for determining the QOI will leverage the same underlying framework as the system model developed Section 1.1.2 (*i.e.* factor-graph model), or, alternatively, the 4-D virtual environment framework developed Section 1.1.3 (*i.e.* multi-player 4-D virtual environment framework). For example, we could envision a school of autonomous players navigating the virtual environment while iteratively re-evaluating the QOI of any encountered data based on the current needs of the system. A “QOI” player might have access to more data—in order to determine what information is most relevant—than other players performing analysis at higher levels of data aggregation. A key aspect of the proposed system is that the function of the component programs is reconfigurable based on dynamic system objectives. In summary, we develop an adaptable layered data management strategy that shifts emphasis to the most relevant data for a given objective.

2 PHASE I TECHNICAL OBJECTIVES AND BEYOND

The Phase I-III technical objectives are given by the OSD13-LD3 topic description from the solicitation. We include some comments addressing the Phase I technical objectives.

2.1 Select and automate the population of a feature layer;

The feature layer chosen for the Phase I study is a 4-D virtual space-time environment based on open software packages. Specifically we propose to base the feature layer on open license software from www.opensimulator.org.

2.2 Pre-process data as required to approximate a uniform grid;

We propose to adopt a hierarchical space-time grid framework for mapping time-location dependent data to the feature layer. Data pre-processing functions will include a data compression function that condenses raw source data into relevant data points for the system application.

2.3 Uncover linear/ nonlinear correlations between environmental, social and cultural variables and feature importance;

In this proposal we develop a concept of context adaptive graph re-factorization. We plan to use this concept for uncovering statistical relationships between variables. We further plan to use hypothesis evaluation to uncover previously unknown variables or features of interest.

2.4 Mature models for why/when a location/area would be of interest;

Locations/areas of interest are identified based on a proposed dynamic factor-graph based model and a multi-player virtual data visualization environment.

2.5 Mature pattern recognition algorithms that approximate human recognition and classification skills utilizing cognitive insight; provide a means to visualize current and predicted states (*e.g.* heat maps showing locations/areas of interest given a context). Define and apply metrics to measure modeling and prediction accuracy and potential for success of products produced.

In this project, we develop “message passing” algorithms based on a dynamic factor-graph framework to obtain an efficient implementation of an advanced semi-autonomous predictive information system.

3 STATEMENT OF PHASE I WORK AND BEYOND

We define six primary Phase I tasks, a reporting schedule, and the outlines for Phase II and Phase III.

3.1 TASK A.1 (Code Development Task A.1): This task is comprised of developing a 4-D virtual environment based on open source code from www.opensimulator.org. Multi-source data mined from internet web sites and downloaded from open code projects will be mapped to the 4-D interface. The source data for the Phase I test system will include: (i) crime statistics data published by local, state, and federal agencies, (ii) open maps data downloaded from www.openstreetmap.org, (iii) text data mined from searches of online news archives, *e.g.* search of historical L.A. Times articles, and (iv) human operator based input data. A key component of Task A.1 is to implement a method for navigating across the 4-D virtual data environment (*i.e.* a continuous path in space-time, or jump discontinuities, etc). During the first three months of Phase I we plan to develop the first iteration of the virtual environment interface, including the mapping of multi-source data to a grid and a basic navigation function. We plan to demonstrate the potential of the proposed framework with a reduced-complexity example, such as forecasting patterns in a single U.S. city (*e.g.* Los Angeles).

3.2 TASK A.2 (Code Development Task A.2): During the second three months of Phase I, we will focus on further development of the virtual environment and the development of human/autonomous multi-player strategies for the test application of gun control in the U.S. By the completion of Phase I, we plan to demonstrate that diverse data types across a wide geographic region can be effectively utilized for information discovery, pattern recognition, and prediction using a 4-D virtual environment framework. We anticipate that the proposed strategy will prove to be advantageous for certain pattern recognition and information discovery applications.

3.3 TASK B.1 (Code Development Task B.1): This task is comprised of developing a modified web-crawler for automated processing (*e.g.* data mining, quality of data analysis, and/or data aggregation) of both World Wide Web data and system internal graph-structured data. The code we develop will be used to evaluate the design concept of using a modified web-crawler for autonomous information discovery with graph structured data. We plan to base our implementation on open source code, from www.openwebspider.org. The first iteration of the code development will be completed in the first three months of Phase I project.

3.4 TASK B.2 (Code Development task B.2): During the second three months of the Phase I project, we plan to further develop a modified web-crawler program for autonomous information processing on

graphs using results obtained during first three months. The program developed in Task B.1 and B.2 will serve as a main component of the autonomous processing sub-system, in addition to a serving as an essential module for data mining operations. We plan to evaluate the utility of a modified web-crawler as a platform for implementing the information processing algorithms developed in Tasks C.1 and C.2.

3.5 TASK C.1 (Algorithm development and documentation task C.1): By the completion of the first half of Phase I we will have an outline of the basic system algorithms and the overall algorithm implementation strategy. The specific algorithms will include: (i) a data pre-processing algorithm for filtering raw source data and translating to a 4-D grid, (ii) a data aggregation algorithm (tentatively based on a QOI metric) for determining context relevant data, and (iii) a message passing algorithm for estimating unknown system variables. The preliminary specifications developed in the first half of Phase I will be evaluated and re-iterated during the second half of Phase I.

3.6 TASK C.2 (Algorithm development and documentation task C.2): We anticipate some progress in specifying a collection of basic system algorithms during the 6 month Phase I project. By the completion of Phase I, we will have a technical memo that includes a description of the system components, processes and algorithms, that were developed, and the major technical accomplishments that were achieved in the Phase I research and development effort. In addition to the technical memo, we include the possibility of seeking to publish the results obtained in the Phase I algorithm development studies (as mentioned in the topic description in the solicitation). The algorithms developed in Phase I effort will serve as a starting point for the development of a Phase II prototype system.

3.7 PHASE I REPORTING TASK

We propose a monthly progress reporting schedule. The P.I. is the correspondent for all reporting responsibilities. The six month milestone report will include specific progress updates addressing the Phase I tasks described above and a comprehensive summary of the overall Phase I technical accomplishments. We are committed to accommodating any reporting schedule at the sole discretion of the sponsor.

3.8 PHASE II WORK PLAN OUTLINE

The Phase II work plan outline consists of producing a prototype system that is capable of rapidly identifying locations/areas of interest for a given context based on adaptive information processing and visualization with a 4-D virtual environment framework. In developing the Phase II system, we will build upon the research results of the Phase I effort, including initial algorithm specification, performance metrics of interest, and lessons learned from the Phase I project. The system will be developed to automatically process multi-source data such as sequential batch files or streaming data and accept all standard raw source data formats for images, maps, tracks, text and graphs. It is necessary that the context and pedigree of information be maintained for operator review. We have proposed a QOI metric for this purpose. By the completion of Phase II we will seek to have a proof-of-concept system capable of transition to a commercial technology.

3.9 PHASE III WORK PLAN OUTLINE

The Phase III work plan outline consists of producing a system capable of deployment and operational evaluation. The system shall consume available operational data sets and focus on areas that are of interest to specific transition programs and/or commercial applications. Autonomous processing steps and system meta-data will be accessible by operator and presented in human readable form. The software and hardware will be modified to operate in accordance with guidelines provided by transition sponsor.

4 RELATED WORK

Predictive information science, artificial intelligence, database management, and information theory are well researched fields of technology. An important contribution of this project will be an application of coding theory (error control codes) to the problem of predictive information processing with a large number of system variables. In particular, we propose to develop algorithms based on research results in the field of inference processing on graphs, factor graphs, and the sum-product algorithm. The P.I. has research experience on the topic of error control codes [Jacobsen, *et al.*, 2007—2011]. Specifically, as a research and development engineer with Alcatel-Lucent, the P.I performed research on the construction of near capacity performance error control codes for multi-rate and multi-terminal wireless communication systems. This research effort resulted in a U.S. Patent [Jacobsen and Soni, 2011].

Reference contact #1: Dr. Thomas L. Marzetta Bell Labs Research, Alcatel-Lucent Phone #: (908) 582-3090 E-mail: tlm@research.bell-labs.com	Reference contact #2: Dr. Alexei Ashikhmin Bell Labs Research, Alcatel-Lucent Phone #: (908) 582-4107 E-mail: aea@alcatel-lucent.com

Probability theory and linear/nonlinear systems theory are fundamental building blocks for this work—the P.I. has adjunct teaching experience in the subjects of Probability Theory (NYU-Poly) and Linear

Systems Theory (Columbia University).

Reference contact #3: Dr. Keren Bergman	Reference contact #4: Dr. Xiaodong Wang
Dept. of Electrical Engineering	Dept. of Electrical Engineering
Columbia University	Columbia University
Phone #: (212) 854-1744	Phone #: (212) 854-6592
E-mail: bergman@ee.columbia.edu	E-mail: wangx@ee.columbia.edu

5 RELATIONSHIP WITH FUTURE RESEARCH AND DEVELOPMENT

Aquerre Technologies LLC is seeking to advance the state of the art of predictive algorithms research through both industry-government and industry-academia partnerships. We envision ourselves as an innovator and catalyst of future research and development in the field.

The Phase I portion of the proposed research and development effort shall provide a foundation for the Phase II effort by evaluating the effectiveness of context adaptive message passing algorithms and a multi-player virtual environment for predictive information processing. The test application chosen for the Phase I fasibility study is gun control in the United States. The developed system will be broadly re-configurable pending target application anwe hope to attract U.S. Government customers. Predictive information systems have widespread applications and we claim that our approach has potential advantages for pattern recognition, prediction, inclusion of multi-source and HUMINT data, and intelligence systems data management.

Note that the P.I. does not currently possess an active security clearance and we are open to applying for a clearance(s) if desired.

6 COMMERCIALIZATION STRATEGY

Aquerre Technologies LLC is seeking to be an industry leader in the field of predictive information systems. We are actively pursuing the research and development of a “Gun Control Network.” Specifically, we are developing new gun control technologies for the purpose of advanced threat detection, prediction, and deterrence, including application to security systems of schools, colleges and universities across the United States, and beyond. This represents a huge market opportunity. The primary funding sources that we have sought since our inception (May, 2013) are U.S. Government SBIR awards from the Department of Homeland Security and the National Science Foundation.

Predictive information systems have matured across various sectors of the economy, including the financial and military sectors. We believe that the currently available complex and expensive predictive systems are still inadequate for broad commercial markets. A key challenge towards accomplishing our goal of developing a viable Gun Control Network is the challenge of developing a system that is both effective and practical in light of cost and complexity constraints. Complexity costs arises from the sheer volume of data that is available to the system and from approximating the performance of an ideal system. Other cost constraints include the challenge of developing a practical system for wide deployment, competition from other vendors, and funding availability.

In addition to the school/college/university security market opportunity that we have identified, the system developed in this SBIR program has broader application to the strategic interests of the Department of Defense and the Office of Naval Research. Therefore a key customer of the proposed technology is the United States Government, its agencies and subcontractors.

7 KEY PERSONNEL

7.1 Principal Investigator

Noah Jacobsen, Ph.D. Curriculum Vita	
Citizenship	United States Citizen
Date of Birth	02/15/1978
Contact Information	• Home address: 1445 Colby Ave #3, Los Angeles, CA 90025 • Cell phone: (310) 857-8049 • Fax: (424) 270-0709 • E-mail: noah@aquerre-technologies.com • Web page: http://aquerre-technologies.com
Objective	I represent my company, Aquerre Technologies LLC. We are seeking industry, academic, and government partnerships for a new project to research and develop a new gun control technology.
Experience	
Research and Development	• Aquerre Technologies (present) • LGS Innovations, Alcatel-Lucent, Lucent Technologies (Jul. 2006-Oct. 2011)

Scientist	◦ Signal processing and error control coding for wireless communication systems
Adjunct Professor	• Columbia University, Dept. of Electrical Engineering ◦ Linear Systems Theory, Spr. 2012 • Polytechnic Institute of New York University, Dept. of Electrical and Computer Engineering ◦ Probability Theory, Fall 2010
Education	
	• University of California, Santa Barbara ◦ Ph.D. Electrical and Computer Engineering, Sept. 2005 ◦ M.S. Electrical and Computer Engineering, Jun. 2003 • Cornell University ◦ B.S. Electrical Engineering, Jun. 2000
Publications	

- N. Jacobsen and R. Soni, “Method and system for encoding data using rate-compatible irregular LDPC codes based on edge growth and parity splitting,” U.S. Patent No. 7966548, Jun. 2011.
- N. Jacobsen, “Practical cooperative coding for half-duplex relay channels,” In Proc. Conf. on Information Sciences and Systems, Baltimore, MD, Mar. 2009.
- N. Jacobsen and U. Madhow, “Coded noncoherent communication with amplitude/phase modulation: from Shannon theory to practical turbo architectures,” IEEE Trans. Communications, 56(12): 2040-2049, Dec. 2008.
- N. Jacobsen, G. Barriac, and U. Madhow, “Noncoherent eigenbeamforming and interference suppression for outdoor OFDM systems,” IEEE Trans. Communications, 56(6): 915-924, June 2008.
- N. Jacobsen and R. Soni, “Design of rate-compatible irregular LDPC codes based on edge growth and parity splitting,” In Proc. IEEE Vehicular Technology Conf. (VTC), Baltimore, MD, September 2007.
- K. Solanki, N. Jacobsen, U. Madhow, B.S. Manjunath, and S. Chandrasekaran, “Robust image-adaptive data hiding using erasure and error correction,” IEEE Trans. Image Processing, 13(12):1627-1639, December 2004.
- N. Jacobsen, K. Solanki, U. Madhow, B.S. Manjunath, and S. Chandrasekaran, “Image adaptive high-volume data hiding based on scalar quantization,” In IEEE Military Communications Conf., Anaheim, CA, October 2002.

- G. Barriac, N. Jacobsen, and U. Madhow, “Beyond BAD: A parallel arbitration framework for low-complexity equalization,” In Proc. Allerton Conf. on Communications, Control, and Computing, Monticello, IL, October 2001.

Honors/Recognitions

- 3GPP2 Ultra Mobile Broadband (UMB) Air Interface Specification: Recognition of Contribution to LDPC Ad Hoc Group, 2007.
- National Science Foundation (NSF) and Japan Society for the Promotion of Science (JSPS) East Asia Summer Institutes Fellowship, Yokohama National University, Japan, 2003.
- Microelectronics Innovation and Computer Research Opportunities Scholarship, University of California, Santa Barbara, 2000–2001.
- Theodore C. Ohart Scholarship in Engineering, Cornell University, 1999–2000.
- Cornell University School of Engineering Cooperative Education Program Participant, Floyd R. Newman Laboratory of Nuclear Studies, Cornell University, 1998–1999.
- Cornell University Dean of Students Service Award: Cornell Concert Commission, 1998 and 1999.

7.2 Other Key Personnel

Pending outcome of this proposal, Aquerre Technologies LLC will hire at least one employee. We will seek to hire a part-time computer software programming expert, to perform 500 hours of work on the Simulation Environment and the Modified-Crawler tasks described in Section 3. We include the possibility of hiring an additional employee (computer scientist) pending project needs.

8 FOREIGN CITIZENS

All personnel performing work on this SBIR project are expected to be U.S. Citizens. The inclusion of any foreign citizens is subject to sponsor approval.

9 FACILITIES/EQUIPMENT

9.1 CURRENT OPERATIONAL OFFICE

The current sole business office of Aquerre Technologies LLC is located at 1445 Colby Ave #3, Los Angeles, CA 90025. Any facilities, equipment, and data management supporting this SBIR proposal will be provided from the primary business office address.

9.2 OFFICE SPACE CONTINGENCY

Pending the outcome of SBIR proposals applicable to the year 2014 and beyond, we foresee the possibility of requiring additional office space. In this event, a small business office location will be leased in Los Angeles. Any possible change in facilities circumstance is subject to sponsor approval.

10 SUBCONTRACTORS/CONSULTANTS

We do not foresee the need to include subcontractors or consultants on the Phase I project.

11 PRIOR, CURRENT OR PENDING SUPPORT OF SIMILAR PROPOSALS OR AWARDS

Aquerre Technologies LLC has no prior, current, or pending support for the work proposed in this SBIR project.

12 REFERENCE PAPERS

12.1 U.S. Department of Defense, “Sustaining U.S. Global Leadership: Priorities for the 21st Century Defense,” Jan. 2012.

12.2 U.S. Fleet Cyber Command, “U.S. Navy Information Dominance Roadmap 2013-2028,” Mar. 2013.

12.3 R. Antony and J. A. Karakowski, “Fusion of HUMINT and Conventional Multi-Source Data”, National Symposium on Sensor and Data Fusion (NSSDF), Jul. 2007.

12.4 S. McGirr and L. Keenan, "Environmental Influence on Insurgent Activity in Afghanistan", National Symposium on Sensor and Data Fusion (NSSDF), Oct. 2011.

12.5 N. Johnson, *et al.*, “Pattern in Escalation in Insurgent and Terrorist Activity,” Science, vol. 333 (6038): 81-84, Jul. 2011.

12.6 R. Sagarin, *et al.* “Decentralize, adapt, and cooperate.” Nature. 465: 292-293, 2010.

12.7 F. R. Kschischang, B. J. Frey, H.-A. Loeliger, "Factor Graphs and the Sum-Product Algorithm", IEEE Transactions on Information Theory, vol. 47 (2): 498-519, Feb. 2001.

12.8 D. J. C. Mackay and L. Peto, “A Hierarchical Dirichlet Language Model,” Natural Language Engineering, vol. 1 (3): 1-19, 1995.

12.9 N. Jacobsen and R. Soni, “Design of rate-compatible irregular LDPC codes based on edge growth and parity splitting," In Proc. IEEE Vehicular Technology Conf. (VTC), Baltimore, MD, September 2007.

12.10 N. Jacobsen, "Practical cooperative coding for half-duplex relay channels," In Proc. Conf. on Information Sciences and Systems, Baltimore, MD, Mar. 2009.

12.11 N. Jacobsen and R. Soni, "Method and system for encoding data using rate compatible irregular LDPC codes based on edge growth and parity splitting", U.S. Patent No. 7966548, Jun. 2011.

APPENDIX I: LIST OF OPEN SOFTWARE AND STANDARDS REFERENCED IN THIS PROPOSAL

- Open Office: http://www.openoffice.org
- Open Simulator: http://www.opensimulator.org
- Open Street Map: http://www.openstreetmap.org
- Open Web Spider: http://www.openwebspider.org
- MongoDB (geo-spatial indexing): http://www.mongodb.org

APPENDIX II: A NON-COMPREHENSIVE LIST OF DATA SOURCES

- Online Database: Justice Dept. Crime Statistics Data, Federal Database of Lost or Stolen Guns.
- Geographical and Maps Data: Open Maps Data, Global Positioning System (GPS) Data, Geographic Information Systems (GIS) Data.
- News Service Data: Online News Archive Data, Global Financial Markets Data.
- Electronic Communications Data: Telephony Data, SMS, E-Mail, Twitter, Online Message Board Data
- Web Content Data: text, images, audio, video and meta-data.
- Social Networking Data: Twitter, Facebook, Open and Closed Social Networks Data.
- Human Intelligence Data (HUMINT): eye-witness reports, human cognitive analysis data.
- Imagery Intelligence Data (IMINT): aerial and satellite imagery data.
- Signals Intelligence Data (SIGINT): electro-magnetic (EM) signals data, radio frequency (RF) signals, visible, acoustic, infrared (IR), X-ray, and millimeter wave signals, radar systems and sensor networks data, broadcast signals, radio, television signals.