Arbitrary Bit Streams Classification as Encrypted, Scrambled or Compressed, as Voice, Text or Image Technical Abstract

In modern digital communication systems the ability to intercept and categorize or even to 'fingerprint' data is desirable for many disciplines. This effort refines a powerful multivariate statistical algorithm which categorizes arbitrary bit streams according to various data formats and data security/transmission operations. The algorithm calculates 13 statistical parameters from a small sample of a demodulated, de-encoded bit stream, and uses this vector of parameters to frame and format the data and determine if the bit stream is open data, encrypted data, scrambled data or compressed data. This effort is intended to demonstrate the feasibility of using this algorithm to perform multiple real time bit stream categorizations to enable the sorting and selection of digital communication intercepts for intelligence purposes. The multivariate analysis includes a powerful 'fingerprinting' capability that is of keen interest to digital collection platforms. This allows high interest signals to be separated from others for later analysis or decryption. The multivariate statistical analysis done by this algorithm allows calculation of the probabilities of the sampled signal belonging to various categories. These probabilities make this algorithm a powerful objective classifier rather than a subjective analysis tool. The high fidelity of this algorithm opens multiple applications to its versatility.

Anticipated Benefits/Potential Commercial Applications of the Research or Development.

The use of digital communications and the necessity to sort through a sea of bit streams is proliferating in our world. The multivariate statistical analysis applied in this development has multiple applications for analysis of not only bit streams but of any statistically diverse pattern matching problem. The algorithm developed in this effort will be implemented in Java and applied to interception of key digital intelligence. The necessity to automatically frame and un-format intercepted digital data is essential to any military intelligence interceptor. Further, the ability to 'fingerprint' data of keen interest and record it for further analysis will be paramount for intelligence purposes. The development and testing of this algorithm will lead to the prototyping of digital classification code which will be implemented on an intelligence collection platform. RBI will take the lead in the implementation of this beta capability where the algorithm will perform against actual digital data collections. Upon successful demonstration of this beta capability, various other collection platforms will be upgraded with this powerful sorting capability. The first beta tests are expected to take 18 months because of code implementations. The successive upgrades to various collection platforms is expected to take less than 6 months because of the open architecture of our developments. Additionally RBI is exploring alternate uses of this powerful Multivariate Statistical Analysis algorithm which can have applications in the commercial search engines which abound in our interconnected society. The ability to search an entire hard drive and detect any file that was encrypted with a triple DES 112 bit key would be of keen importance to the military or law enforcement and this multivariate statistical algorithm can currently perform this task. This effort is a key step in demonstrating the feasibility of this algorithm and to move it to an implementation in actual code in a beta release.

List of 8 Key Words that describe the Project. Digital Communication, Intelligence, Bit Stream, Digital Categorizing, Multivariate Statistics, Encryption, Data Scrambling, Data Compression

Arbitrary Bit Streams Classification as

Encrypted, Scrambled or Compressed, as

Voice, Text or Image

1. COVER SHEET (see Electronic Submittal)

RBI
Principal Investigator: Edward G. Rice
9511 W.Waneta Lake Rd. Hammondsport NY 14840
Phone: (607) 292-6639 Email: edrice4@linkny.com
DUNS# 102385056 Web: www.linkny.com/edrice4/engnr

For any purpose other than to evaluate the proposal, the data referenced below shall not be disclosed outside the Government and shall not be duplicated, used or disclosed in whole or in part, provided that if a contract is awarded to this proposer as a result of or in connection with the submission of this data, the Government shall have the right to duplicate, use or disclose the data to the extent provided in the funding agreement. This restriction does not limit the Government's right to use information contained in the data if it is obtained from another source without restriction.

This proposal submitted to:

SBIR Topic Num: AF03-094
SBIR Title: Innovative Information System Technologies
SBIR Research & Technical Areas: Information Systems
SBIR Topic Author: Janis Norelli,
Phone: (315) 330-3311, Fax: (315) 330-2784,
Email: Janis.Norelli@afrl.af.mil

2. IDENTIFICATION & SIGNIFICANCE OF THE OPPORTUNITY

The objective of this proposal is to demonstrate the feasibility of classifying arbitrary bit stream types in near real time. The real time categorization of an intercepted bit stream as encrypted data, scrambled data, or compressed data along with a determination of the type of data as voice, text, image or other; is viable and an extremely attractive ability in the intelligence environment[1]. This ability has previously been marginally demonstrated[2] but this effort will demonstrate the feasibility of making this capability robust and of very high fidelity.

In the modern digital communication systems the ability to intercept a myriad of data streams has proliferated. In the vast sea of information being broadcast the intercepted signals containing useful intelligence can be lost. This vast amount of data available taxes even the ability to record data for later analysis. A real time ability to categorize a data stream as encrypted or non encrypted, as scrambled or non-scrambled or as voice, text, image or other data; would be a tremendous asset for the retrieval of any data stream that contains useful intelligence.

In this effort RBI will refine and demonstrate a statistical pattern recognition algorithm that can perform just such a discrimination on arbitrary bit streams [3]. The algorithm has demonstrated superb performance on the varying word lengths and framing formats found in modern communication links. The results of this effort will demonstrate the feasibility of using this specialized statistical pattern recognition algorithm for completely discriminating arbitrary bit streams in real time, with very minimal amounts of data demodulated and collected. Such a demonstration will lead the way to greatly enhanced and selective digital SIGINT collection

systems.

2.1 Background

2.1.1 Intelligence Data Escalation: There is an acute need for enhancements to intelligence gathering systems that enable them to gather, sort, record and report pertinent intelligence in near real time. This necessity is aptly demonstrated by the unchecked terrorist activities of 2001. Further *"Technology advances are enabling new satellite communications systems that combine broad band data rates with small terminals. These novel systems are being designed to provide affordable 'last-mile' network access to home and small business users world wide*. [4]" The new availability of the Internet's Transmission Control Protocol (TCP) over the advanced broad band satellite systems distinct from narrow band voice services greatly advances this proliferation of data available for intelligence interception. In the sea of information now available for interception careful selection of data that should undergo more careful scrutinization can greatly increase the probability of a useful intelligence interception. Figure 1 shows the normal block diagram of digital data transmission and the interception and birth of an arbitrary digital bit stream which may be unframed, of various word length and format, and only partially intercepted. The need to automatically detect, categorize and prioritize such an intercepted digital signal would be a great enhancement to intelligence collection.

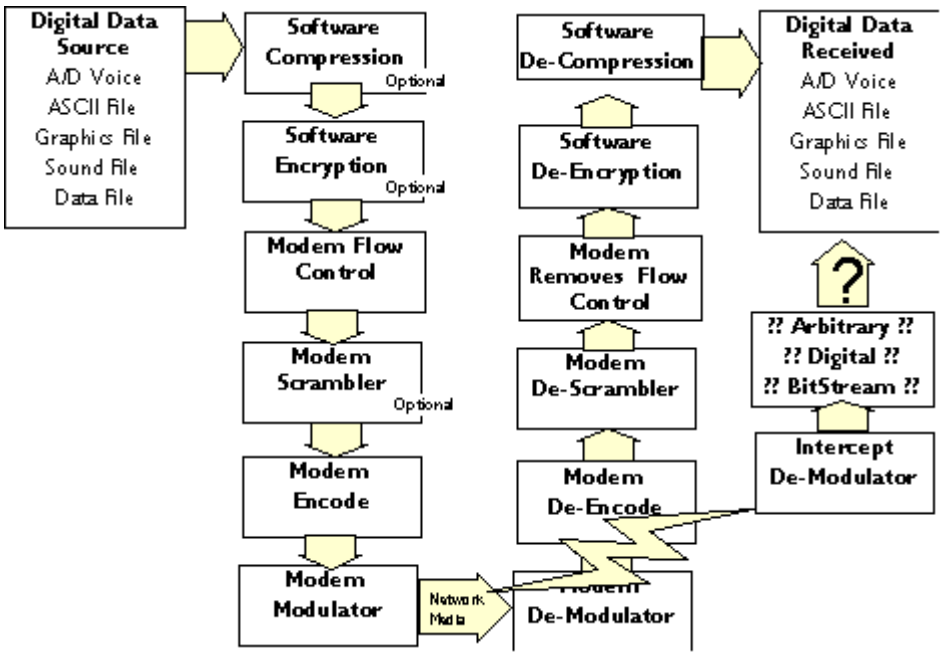

Figure 1 Digital Data Broadcast, The Intercept and The Arbitrary Bit Stream

2.1.2 Arbitrary Bit Stream Categorization.

A primary consideration in filtering through the sea of information available from digital communication channels is the categorization of an intercepted and reconstructed bit stream. Intercepted digital communication may be properly demodulated, properly decoded, and properly unscrambled into several hundred bit streams. These bit streams are arbitrary because of uncertainty in word length, asynchronous timing data, parity data or additional scrambling. Additionally it is uncertain if a bit stream represents encrypted or compressed data and the type of data further contained may be voice, text, image or other data. Current software used for analysis of these bit streams uses a trial and error approach to determine word length and timing/parity aspects followed by manually looking at the results to determine if the data may be usable. These trial and error methods cannot obtain real-time results. An algorithm that makes a determination of the category of bit stream will enable the real time processing and recording of only pertinent, high interest, 'fingerprinted' data.

In Figure 2, one can see obvious differences between encrypted and scrambled text in a histogram of each formatted bit stream, but training a computer to characterize these differences has been a challenge. Again this recognition has been centered on analyzing the measure of randomness in the signals (largely through Shannon Entropy measures) but there are other parameters which provide more pertinent insight to these differences. Even the difference introduced into encrypted text bit streams by formatting the bit stream can be visually detected in a histogram as seen in Figure 3. Adding this type of structure to an encrypted or scrambled (i.e. Randomized) signal drastically changes the Shannon Entropy. The blue histogram in that figure has a 28 bit word length with 1 odd parity bit 2 stop bits and 1 start bit. The violet histogram in that figure has no format data in the bit stream and a simple 8 bit word length.

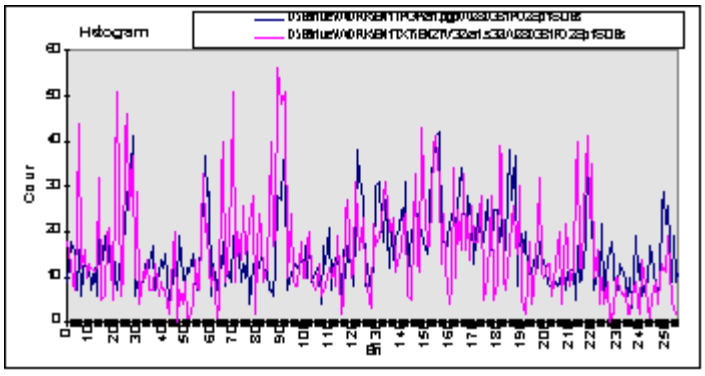

Figure 2 Obvious Histogram Differences Between Encrypted and Scrambled Text

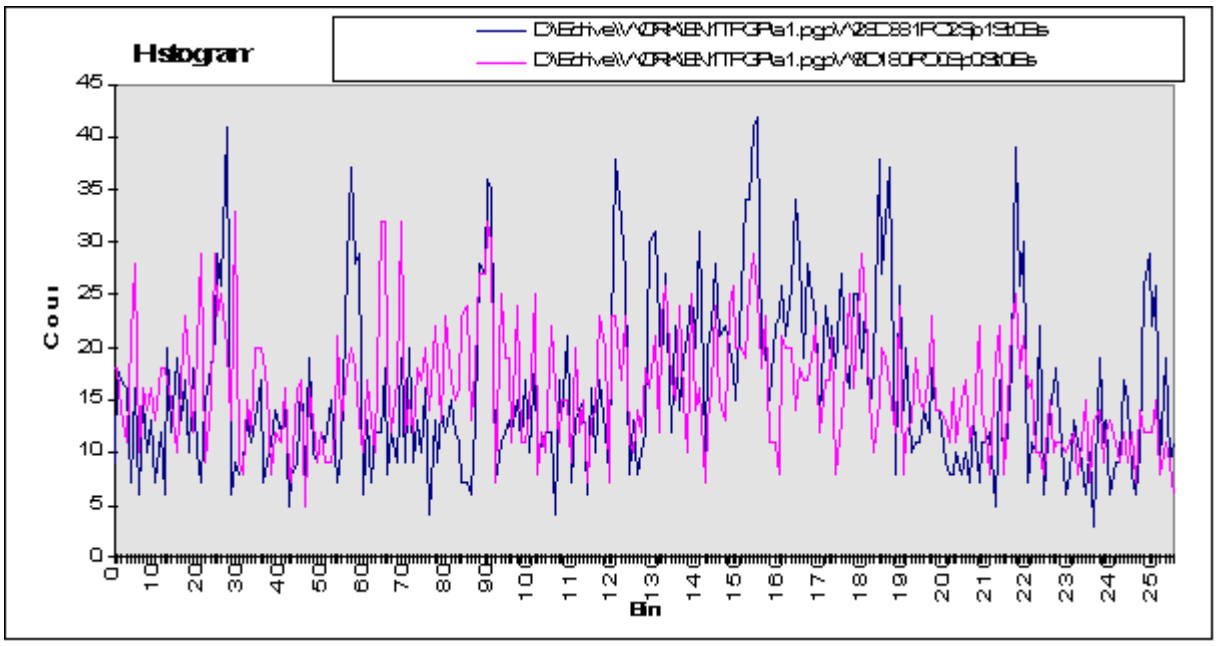

Figure 3 Histogrammed Differences Between Formatted and Unformatted Encrypted Text

2.1.3 Bit Stream Categorizing Mechanics

Significant research has gone into the problem of categorizing arbitrary bit streams using bit entropy[5]. The usefulness of this method has been limited by the fact that the Shannon entropy calculation[6] is word length and bit stream shift dependent. Investigation of this dependence is ongoing but fruitless. Determining a bit streams word length or framing information the statistical categorization of the unknown bit stream is most effective. This statistical categorization is accomplished by introducing additional statistical parameters used in this effort. In initial investigation of pertinent parameters the following 13 parameters, calculated on 4 Kbytes of an arbitrary bit stream sample, were of paramount importance.

1	Shannon Entropy	The Shannon Entropy calculation based on 8 bit word length
2	Mean	The μ value of all the 8 bit words in the sample. Normalized to $0 < \mu < 1$
3	Stdv	The σ value of all the 8 bit words in the sample. Normalized to $0 < \sigma < 1$
4	Maximum Value	The maximum value of all the 8 bit words in the sample. Normalized to $0 < max < 1$
5	Histogram Stdv	The σ of the histogrammed sample Normalized to $0 < \sigma < 1$
6	Histo Max Value	The maximum occurrences in the histogram. Normalized to $0 < max < 1$
7	# of Word Reps	The number of 8 bit word repetitions. Normalized to a percentage of words in the sample
8	# 8 Bit Seq	The number of 8 bit sequences of all ones or zeros.
9	# 7 Bit Seq	The sequence counters include bit sequences that cross the 8 bit word frames.

10	# 6 Bit Seq	The sequence counts are normalized to a percentage of 8 bit words in the sample
11	# 5 Bit Seq	A sequence of 5 would also count as 2 ea 4 bit sequences, and 3 ea 3bits.
12	# 4 Bit Seq	Theoretically the #of 3 bit sequences, normalized to the percentage of 8 bit words
13	# 3 Bit Seq	in the sample could exceed 100%, but not in a meaningful sample.

A simple Euclidean distance vector comparison method was initially used to separate files into categories. The parameters were calculated for several known bit streams. For example, 10 each encrypted text files were formatted into a 28 bit word length, with 1 odd parity bit, 2 stop bits and 1 start bit. The 13 parameters were calculated on these 10 bit streams and the mean values were calculated and stored in a 13 dimension reference vector. Fifty such reference vectors were calculated. The 13 parameters were now calculated for unknown bit streams and their vector was compared to the reference vectors via a Euclidean distance vector comparison. The reference vector which was closest to the unknown bit stream's vector was chosen as the classification for this vector. This method was able to discriminate how a bit stream had been modified, i.e. compressed, encrypted or scrambled, and remarkable was often able to determine the data type, text, image, voice, etc., even through the modification. Analysis of these parameters further demonstrated an intriguing ability to determine the type of scrambling algorithm that had operated on the original data. It was not until the statistical distributions of each of these parameters was calculated and displayed that the superb performance of this method was realized.

Replacing the Euclidean vector with a multi-variant statistical analysis in this comparison has remarkably enhanced fidelity and can easily recognize various word lengths and framing (bit slip) errors. The multi-variant statistical analysis comparison requires both a mean reference vector and a standard deviation reference vector. The Shannon entropy distributions were found to follow a Beta Distribution and careful normalization of all the other parameters, (keeping their values between 0 and 1) causes each to fit the beta statistical distribution very well. These reference vectors are calculated only for the signals of keen interest. Of course the unknown bit stream has no standard deviation vector since it is a one time sample, but the probability that each parameter of it's parameters is within the distribution of the reference vector can be calculated. The multivariant statistical analysis tallies all these probabilities into a single probability that this unknown bit stream belongs to the particular reference category. This probability has a subjective cut off value (we initially used 60 %) that can be evaluated or varied for various applications. Such a multivariant statistical analysis gives you 2 capabilities, first to determine the higher representative likelihood of a sample belonging to reference A or to reference B; and secondly to determine an objective likelihood that the sample belongs to reference A at all. Divergent statistics for similar bit streams are shown in Figure 4.

In Figure 4 notice that an unknown file with a Shannon entropy calculation of 0.6 could as likely be in the first distribution or the second (the two Shannon entropy distributions are indicated with blue diamonds, and Shannon Entropy typically does not have distributions that allow distinguishing between many key categories.) However, a standard deviation measure of 0.2 (the distribution with green triangles) and a maximum value of 0.7 (the distribution with the violet x's) would verify with very high certainty that the unknown signal falls in the second distribution pattern and is thus more likely an HTML document than a text document. Thus in order for the an unknown bit stream to be classified in a particular category, it must make it through 13 separate parameter distribution "hoops", several which are often unique to only that type of signal. These 'hoops' can be of various widths (larger sigma's) depending on the fidelity of the categorization to be achieved. (i.e. Differentiation between scrambled data of all types and scrambling algorithms and encrypted data of all types and encryption algorithms, vs. differentiation between encrypted text and encrypted image bit streams.) So this classifier can work around transmission formats and framing errors and still distinguish data types. Or this classifier can wade into the transmission formats and framing errors and classify word lengths in the data transmission. It has already demonstrated that kind of fidelity and is ready for this feasibility demonstration.

This high fidelity 'fingerprinting' of data can first separate an arbitrary bits stream into its encrypted, scrambled or compressed category, then further determine the proper framing and word length format of the bit stream. It is extremely powerful. The multivariant method also enables comparisons against selected categories and formats without calculation of thousands of reference vectors for all the possible datas, scrambling methods and formats found in the digital communications world. To use this great classifying power one must know a particularly application and direction to pursue prior to demonstrating the full feasibility. We have currently honed our testing and development tools in 2 directions, first in the framing and formatting analysis of a completely arbitrary intercepted bit stream, and secondly at evaluating encrypted data and determining what algorithm was used to encrypt it. Both areas have great promise. This proposal focuses on the former and the full feasibility of this approach begs to be tested.

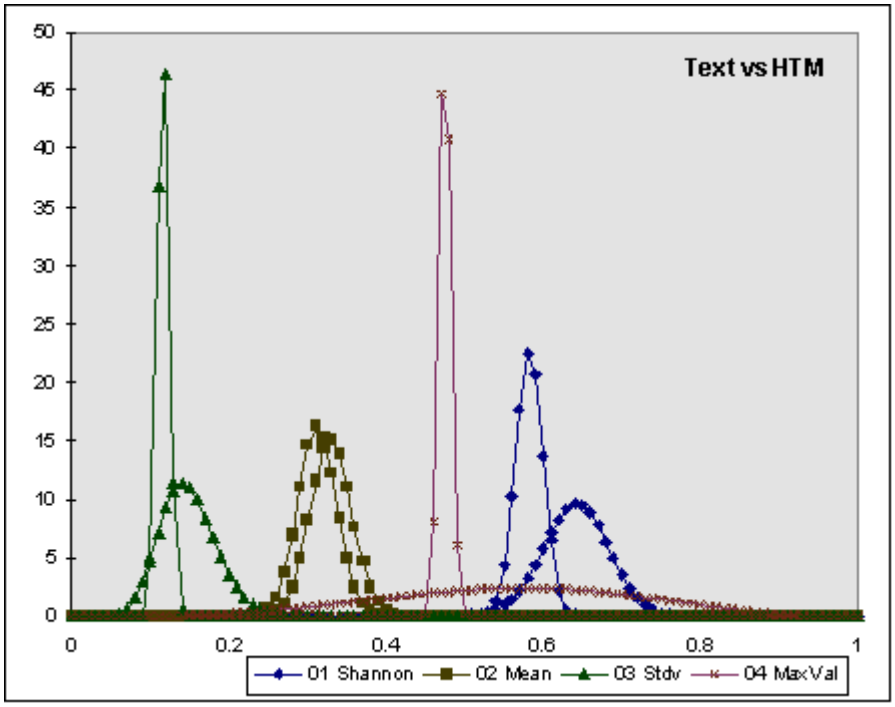

Figure 4 Similar Bit Streams with divergent statistics.

3. PHASE I TECHNICAL OBJECTIVES

The overall objective of the proposal is to develop the multi-variate statistical analysis for categorizing arbitrary bit streams and investigate it's robustness against various word lengths and framing errors that are encountered in arbitrary bit streams and against various scrambling and encryption techniques. The specific objectives of the program can be enumerated as follows:

1. Investigate the various word lengths normally encountered in modern digital communication systems and develop sample sets of these word lengths with various data types and implementations.

2. Refine the calculations of the 13 vector parameters to provide robustness to investigate the dependence on word length and framing errors which may effect overall classification of an arbitrary bit stream.

3. Refine and optimize the multivariate statistical comparison algorithm to capture and categorize word length and framing error variations in various data types.

4. Review the mathematics of the Multivariate Statistical Vector fitting along side the Euclidean Distance Vector Fitting Algorithm and explore the possible additional benefits of an Eigen value evaluation of the vectors. Evaluation of multi dimensional vectors and pattern matching methods

is still wide open for high performance payoffs in this research area.

5. Evaluate the performance of the refined algorithm to discriminate between various bit stream formats, framing errors, data types and data encoding (encryption, compression or scrambling).

This multivariate statistical approach has already demonstrated the ability to differentiate data found with various word lengths and framing errors. It is ready for these feasibility tests, for prototype and for deployment into operational intelligence platforms.

4. PHASE I - WORK PLAN

Phase I research will be restricted to refining the algorithm and showing feasibility of using this algorithm to categorize arbitrary bit streams with various format and framing uncertainties. The Phase I work plan will include the following tasks for achieving the stated objectives:

4.1 Task I - Investigate and Refine the Algorithms versatility

A. Investigate Applicable Data Formats: During this task the various formats of keen interest will be determined with contact with applicable AF offices. These formats with various others from prior tests will be constructed and examined for accuracy. An adequate test base of files to include encrypted scrambled and compressed, text, image, audio, data and other files will be formatted with these formats. Anomalies of these formats and files will be examined with existing bit stream classification tools. A sample set of these files will be selected as reference bit streams and another set for “unknown” samples.

B. Refine Parameter Calculations: During this task the 13 parameters of interest will be scrutinized for applicability and their calculations subjected to accuracy tests. Additional statistical parameters will be examined for inclusion since some were previously eliminated because of redundancy. (i.e. The histogram mean was found to be completely redundant but an applicable replacement characterization of the histograms distribution was not included.) The normalization of each variable will be revisited to ensure that it’s distribution can best be characterized as a Beta distribution between 0 and 1. The sequence counters as implemented in visual basic will be particularly scrutinized for more efficient implementation. This counter is the ‘long pole in the tent’ for processing efficiency and more versatile code will be pursued.

C. Refine Multivariate Statistics: During this task the Beta distribution of each parameter will be subjected to applicability and accuracy. Further refining of each parameters statistical distribution will be examined to ensure that a representative probability can be produced, and then that these probabilities can be synthesized into one representative probability of a classification match. Much work has been done in the field of multivariate statistics, however this effort is a new application in this field. We are matching a known set of distributions to a sample set of one. The exploration of this new found and powerful analysis technique, its comparison to Euclidean Vector and Eigen Vector methods will be ongoing throughout this phase

but centered in this task.

4.2 Task II - Optimize the Algorithm and Evaluate it's Effectiveness

A. Optimize Algorithms: Performance evaluation begins during this task. This includes the overall evaluation of the whole process of classification of unknown signals. The parameter calculations and the multivariate analysis is revisited and refined to work together as a whole classification algorithm. The particular classifications are evaluated to optimize the algorithm for peek performance.

B. Evaluate Algorithm Performance: During this task selected unknown bit streams are classified by the algorithm. The classification results and probabilities of error are evaluated and reported.

4.3 Reporting.

A. Bimonthly Progress reports: In addition to day-to-day informal contacts with the program monitor, technical progress reports will be submitted every two months.

B. A Final Report will be generated. A final report detailing all of the development and evaluation will be submitted at the end of Phase I.

The Phase I work would follow the tentative schedule below:

Task	Days	Hrs	Start	Stop
SBIR Phase I	273	720	03/03/03	12/01/03
Kick Off	1	3	03/03/03	03/04/03
T1A Investigate Applicable Data Formats	56	100	03/04/03	04/29/03
T1B Refine Parameter Calculations	56	100	04/15/03	06/10/03
T1C Refine Multivariate Statistics	56	100	05/27/03	07/22/03
T2A Optimize Algorithms	56	150	07/08/03	09/02/03
T2B Evaluate Algorithm Performance	56	150	08/19/03	10/14/03
Final Report	42	117	09/30/03	11/11/03

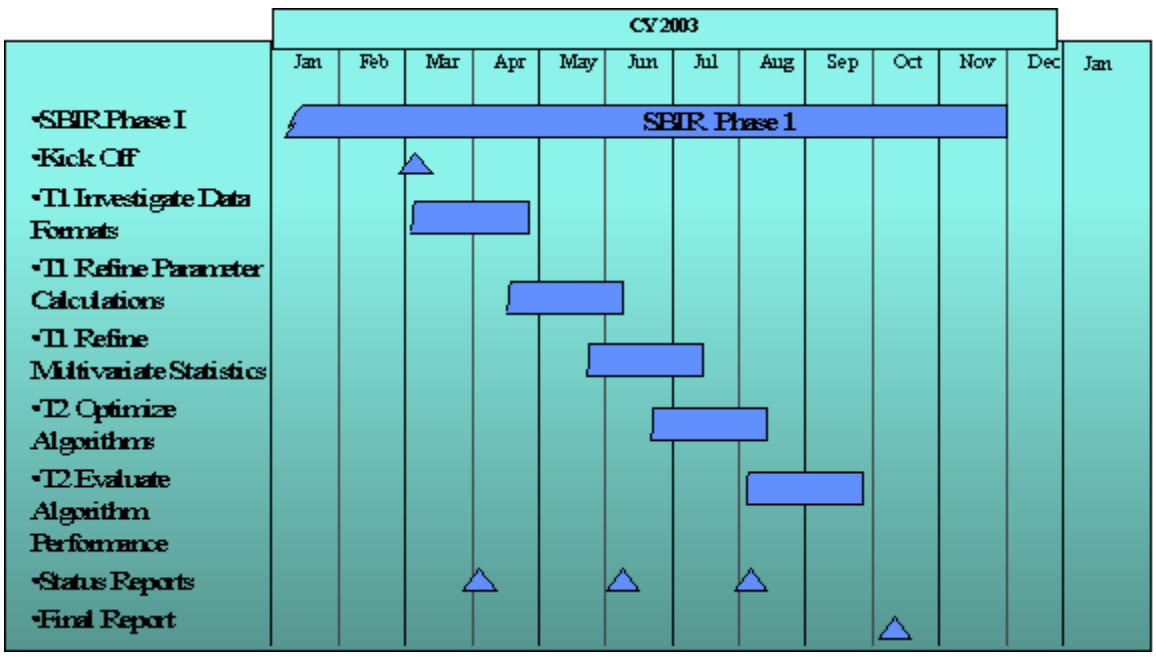

Figure 5 Phase I Schedule

5. RELATED WORK

The Principal Investigator has been involved in pursuing categorizing arbitrary bit streams for the past seven years. At his previous place of employment he was responsible for the programs that enhance the performance of numerous intelligence collection platforms that would be enhanced by this capability. He is in close contact with other researchers exploring intelligence gathering techniques and digital communications.

6. Relationship With Future Research Or R&D

Anticipated improvements in sorting and categorizing digital bit streams will be of immediate use where conventional intelligence collection techniques currently have limitations due to data overloading. The proposed Phase I work will determine this fundamental algorithm as the best to categorizes arbitrary bit streams. The prototyping and deployment of this algorithm as a software package will be accomplished in the Phase II effort.

7. Commercialization Strategy

The use of digital communications and the necessity to sort through a sea of bit streams is proliferating in our world. The multivariate statistical analysis applied in this development has multiple applications for analysis of not only bit streams but of any statistically diverse pattern matching problem. The algorithm developed in this effort will be implemented in Java and applied to interception of key digital intelligence. The necessity to automatically frame and un-format intercepted digital data is essential to any military intelligence interceptor. Further, the ability to 'fingerprint' data of keen interest and record it for further analysis will be paramount for intelligence purposes. The development and testing of this algorithm will lead to the prototyping of digital classification code which will be implemented on an intelligence collection platform. RBI will take the lead in the implementation of this beta capability where the algorithm will perform against actual digital data collections. Upon successful demonstration of this beta capability, various other collection platforms will be upgraded with this powerful sorting capability. The first beta tests are expected to take 18 months because of code implementations. The successive upgrades to various collection platforms is expected to take less than 6 months because of the open architecture of our developments. Additionally RBI is exploring alternate uses of this powerful Multivariate Statistical Analysis algorithm which can have applications in the commercial search engines which abound in our interconnected society. The ability to search an entire hard drive and detect any file that was encrypted with a triple DES 112 bit key would be of keen

importance to the military or law enforcement and this multivariate statistical algorithm can currently perform this task. This effort is a key step in demonstrating the feasibility of this algorithm and to move it to an implementation in actual code in a beta release.

8. KEY PERSONNEL

Edward G. Rice, Senior Engineer

EDUCATION:

M.S., Electrical Engineering, Air Force Institute of Technology, Wright Patterson AFB Ohio, March 1992.

B.S. Electrical Engineering, Ohio State University, Columbus Ohio, March 1982.

CURRENT POSITION AND RESEARCH:

Edward Rice is a retired USAF officer and the Sole Proprietor of RBI. He has more than 18 years of experience in USAF intelligence systems. RELEVANT EXPERIENCE: Prior to starting RBI Ed Rice was the consultant that developed and tested the basic Euclidean Distance Vector Fitting Technique for Categorizing Arbitrary Bit Streams. From 1989 to 1995 he was assigned to Rome Laboratories where he managed numerous intelligence collection technology initiatives. From 1984 to 1989 as a USAF Electrical Engineer he worked with numerous data collection and weapon delivery systems furthering and rounding out his background in intelligence systems. He is currently pastoring a Baptist Church and teaching high school math and science part-time, while pursuing a M.Div. Degree. He is available to work on RBI efforts up to 40 hours per week.

9. FACILITIES/EQUIPMENT

To support its position as an analysis and algorithm development sole proprietorship, RBI currently maintains a 400 sq. ft office with networked PC computer resources. The PC's are equipped with Microsoft office and Lotus Smart Suite which are used in the construction and analysis of the data. The Internet connection in the rural area of Hammondsport NY is currently only available via dialup. It is anticipated that this capability will be adequate for the limited online work required in Phase I. The primary algorithm development is being done with Excel Spreadsheets utilizing Visual Basic to implement the algorithm particulars. These facilities and resources will be available at RBI throughout the course of this effort. Additional office space with a conference area will be remodeled and available in Mar 2003. Additional computer resources or office accommodations can be added for this effort as the need arises.

10. CONSULTANTS

No consultants are presently foreseen for the Phase I program. If a need should arise, RBI has several consultants available from previous contacts in this technical area.

11. PRIOR, CURRENT OR PENDING SUPPORT

RBI has a very similar proposal submitted under:

SBIR Topic Num: N03-150
SBIR Title: Multi-Intelligence SIGINT (COMINT/ELINT) Sensor Processing

SBIR Research & Technical Areas: Sensors, Electronics, Battlespace
SBIR Topic Author: Steve Brown, Phone: 619-524-7895, Fax: 619-524-7374,
Email: stephen.f.brown@navy.mil,
SBIR 2nd TPOC: Eric Helgeson, Phone: 619-553-1122,
Email: helgeseg@spawar.navy.mil
SBIR Acquisition Program: PMW 189 (Naval Electronic Combat Surveillance Systems)

That proposal uses the same classifier core to discriminate between encrypted data of various encryption algorithms and password lengths. Should both efforts be funded Task 1-B and Task 1-C above would dovetail with similar work required for N03-150 and the cost of these tasks would be divided evenly between the two sources, with no double billing.

A white paper entitled “White Paper on RTIBS - Real Time Identification of Bit Streams” was submitted by RBI on 03/25/02 ATTN: Chester J. Maciag, Reference BAA-96-10-IFKA, AFRL/IFGB, 525 Brooks Road, Rome NY 13441-4505 Notice Solicitation Number: BAA-96-10-IFKA Posted Date: Mar 06, 2002 Classification Code: A -- Research & Development Contracting Office Address Department of the Air Force, Air Force Materiel Command, AFRL - Rome Research Site, AFRL/Information Directorate 26 Electronic Parkway, Rome, NY, 13441-4514

Currently there has been no response to RBI’s white paper submittal.

12. COMPANY COMMERCIALIZATION REPORT (SEE ADDITIONAL ELECTRONIC SUBMITTAL)

13. COST PROPOSAL (SEE ADDITIONAL ELECTRONIC SUBMITTAL) (See Last Page)

14. REFERENCES and FOOTNOTES

1 [1] 6 December 96 SENSOR ACE Capabilities Enhancement Study, Appendix A, "Technology Survey and Recommendations" submitted to AFRL/IFEC

2 [2] “Technical Notes On Classifying Arbitrary Bit Streams Using Shannon Entropy, Statistical Distributions and Euclidean Distance Vector Fitting Technique" , by Edward G. Rice, Consultant Research Associates of Syracuse Inc. 15 September 1998, submitted to AFRL/IFEC Abstract: In modern digital communication systems the ability to intercept and categorize data is desirable for many disciplines. This report documents a 2 man-month analysis of classifying arbitrary bit streams using Shannon entropy, statistical distributions, and Euclidean distance vector fitting techniques. The effort demonstrated the feasibility of classifying and identifying arbitrary bit streams using 14 statistical parameters calculated from a bit stream sample. An Euclidean distance technique was used to match a feature vector of these parameters with a calibration matrix which identified the class of the bit stream. The demonstrated potential of this technique allowed even different scrambling polynomials to be discriminated accurately. The technique allowed the accurate discrimination between compression processes, scrambling processes, and encryption processes. Such feasibility is amply demonstrated, and the practical methods developed lend themselves to an immediate operational development of an arbitrary bit stream classifier.

3 [3] The arbitrary bit stream may be any portion from any bit stream, demodulated but of arbitrary word length and framing. Optimal performance is obtained with samples greater than 2K bytes of data.

4 [4] "TCP Performance over Satellite Channels" by Thomas R. Henderson and Randy H. Katz, University of California at Berkeley Dec 1999.

5 [5] "Research Results for Classifying Arbitrary Bit Streams Using Bit Entropy Features" 10 September 1997 by Ed Semplinski, QuesTech, Inc. submitted to AFRL/IFEC

[6] An early parameter evaluated was the Shannon Entropy. This was a leading bit-stream type indicator during the previous entropy analysis, although it was unable to differentiate between some scrambled, compressed and encrypted bit-streams. The Shannon Entropy calculation used here is accomplished as follows:
[Lotus WordPro Equations not transfered to pdf format.]

Where:
H is the uncertainty in words/symbol (for our 8 bits/word instance)
M is the number of possible symbols (*M*= 256 for our 8 bit word instance)
Pi is the probability of encountering the *i*th symbol

For a complete development of this algorithm see "Research Results for Classifying Arbitrary Bit Streams Using Bit Entropy Features", 10 Sep 1997, by Ed Semplinski, Questech, Inc. for Rome Laboratory IRAP.

RBI Cost Proposal

RBI 9511 W.Waneta Lake Rd, Hammondsport NY 14840 Date: **23-Dec-02**

Phone: **(607) 292-6639** CAGE Code: _____

Title: **Arbitrary Bit Streams Classification**

Topic: **AF03-94 Innovative Information System Technologies**

Total Proposal Amount **$78,866.68**

Direct Material	#EA	PER COST	EST COST	TOTAL
a. PURCHASED PARTS	NA		$0	
b. SUBCONTRACTED ITEMS	NA		$0	
c. OTHER	NA		$0	
TOTAL DIRECT MATERIAL			.	$0
	9	MO		
Direct Labor	EST MM	RATE/HR	EST COST	TOTAL
(1728 MH/MY)	20	hr/wk		
Principle Engineer	5	45.00	$32,400	
Jr Engineer	0	24.00	$0	
Programmer	0	24.00	$0	
Publications	0	13.00	$0	
TOTAL DIRECT LABOR	720	0.56	ENGRS	$32,400
LABOR OVERHEAD	OH RATE	XBASE	EST COST	
a. IN PLANT	0.7	$32,400	$22,680	
b. ON SITE	0.55	0	$0	
TOTAL LABOR OVERHEAD	.	.	.	$22,680
SPECIAL TESTING	.	.	.	$0
SPECIAL EQUIPMENT	.	.	.	$0
TRAVEL	EA	RATE	EST COST	
a. TRANSPORTATION	3	979	$2,937	
b. PERDIEM	10	85	$850	
c. LOCAL TRANSPORTATION	10	32	$320	
TOTAL TRAVEL	.	.	.	$4,107
CONSULTANT	Hrs	Rate	Cost	
	0	$75	0	
TOTAL CONSULTANT	.	.	.	$0
OTHER DIRECT COST	.	.	.	$0
TOTAL DIRECT COST AND OVERHEAD		.	.	$59,187
GENERAL ADMIN EXPENCE	0.25	OF COST	.	$14,797
COST OF MONEY	0	OF COST	.	$0
TOTAL ESTIMATED COST	.	.	.	$73,984
FEE OR PROFIT	0.066	OF COST	.	$4,883
*	*	*	*	*
TOTAL ESTIMATE AND FEE OR PROFIT		.	.	$78,867