Encrypted Bit Stream Classification

as Encrypted, Scrambled or Compressed,

as Encrypted DES, PGP, Blowfish or RSA-MDx

Technical Proposal

Phase I Small Business Inovative Research (SBIR) Program
Topic Number N03-150 Agency Navy

By

R B I

9511 W.Waneta Lake Rd
Hammondsport NY 14840

Project Manager/Principal Engineer
Edward G. Rice
(607) 292-6639 edrice4@linkny.com

Encrypted Bit Stream Classification as Encrypted, Scrambled or Compressed, as Encrypted DES, PGP, Blowfish or RSA-MDX

Technical Abstract

The effort proposed herein refines a multivariate statistical pattern algorithm which can separate intercepted bit streams which have been encrypted and determine what encryption algorithm was used on the data. This multivariate algorithm was developed to classify portions of intercepted arbitrary bit streams as scrambled, compressed, or encrypted. The fidelity of the algorithm was so advanced that it could easily distinguish the scrambling polynomials used for scrambled data. Additional testing on encrypted data determined a refined ability to measure the entropy and randomness features of a bit stream to the extent that it could distinguish what encryption algorithm had operated on the data. Initial tests have even indicate a likelihood of determining the type of data that was originally encrypted, text, image or audio. This proposal outlines the development and feasibility tests of this multivariate statistical pattern algorithm which would lead to the fielding of such a encryption classifier. This type of classifier would be a tremendous asset for the collection of pertinent communication intelligence for homeland defense and for military intelligence.

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, Encryption, Decryption, Intelligence, Bit Stream, Digital Categorizing, Multivariate Statistics, Scrambling

Encryped Bit Stream Classification
as Encrypted, Scrambled or Compressed,
as Encrypted DES, PGP, Blowfish or RSA-MDX

1. COVER SHEET (see attached) 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)

2. IDENTIFICATION & SIGNIFICANCE OF THE OPPORTUNITY

The objective of this proposal is to demonstrate the feasibility of classifying intercepted encrypted bit streams or retrieved files 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 encryption algorithm used on the data, is viable and an extremely attractive ability in the intelligence environment[1]. This ability has previously been only marginally demonstrated[2] but this effort will demonstrate the feasibility of making this capability robust and of very high fidelity, even able to determine the encryption algorithms used on data.

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. The ability to separate the data encrypted with a DES or 3xDES algorithm, for example, is even more vital to intelligence collection. RBI's multivariate statistical classifier can do such classifying.

In this effort RBI will refine and demonstrate a statistical pattern recognition algorithm that can perform just such a discrimination on an intercepted bit stream [3]. The algorithm has demonstrated superb performance separating scrambled data based on the scrambling algorithm used. The results of this effort will demonstrate the feasibility of using this specialized statistical pattern recognition algorithm for completely discriminating encryption algorithms, with very minimal amounts of data 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. The need to automatically detect, categorize and prioritize such an intercepted digital signal would be greatly enhanced by the algorithm that separates the encrypted data and classifies the encryption algorithm that was used on the data. Our preliminary tests on our multivariate statistical analysis algorithm demonstrate not only this high fidelity discrimination, but even an ability to separated encrypted text from encrypted image or audio data.

2.1.2 Data Encryption Algorithms.

Encryption is intended to disguise the data in such randomness that no cipher pattern or decryption can be devised without apriori access to the encryption key. The ingredient here that makes encrypted data vulnerable to the detection and classifying is its measured lack of a discernible pattern. The more completely random the data appears the better the encryption technique. Encryption algorithms are not perfect randomizers however, and we found the measures of randomness in an encrypted file leave a fingerprint pointing to the algorithm which encrypted the data, and perhaps even the type of data encrypted. RBI's discriminator is not attempting to break encryption, or determine patterns. It is only attempting to classify an arbitrary bit stream as encrypted and to what randomness level it was encrypted. It can do this because of the multivariate statistical pattern recognition method we use.

Several encryption algorithms have already been tested with RBI's discriminator. Below is a short excerpt giving some background information on the DES and early PGP encryption algorithms.

> In essence, DES processes plain text by passing each 64-bit input through 16 iteration, producing an intermediate 64-bit value at the end of each iteration. Each iteration is essentially the same complex function that involves a permutation of the same complex function that involves a permutation of the bits and substituting one bit pattern for another. The input at each stage consists of the output of the previous stage plus a permutation on the key bits, where the permutation is known as a subkey.
> Triple DES uses two keys and three executions of the DES algorithm with the function following an

encrypt-decrypt-encrypt (EDE) sequence. There is no cryptographic significance to the use of decryption for the second stage. Its only advantage is that it allows users of triple DES to decrypt data encrypted by users of the older single DES; if the same key is used for all three stages, the effect is the same as a single stage. ... With three iterations of the DES function, the effective key length is 112 bits.
IDEA (International Data Encryption Algorithm) is a block-oriented conventional encryption algorithm developed in 1990 by Xuejia Lai and James Massey of the Swiss Federal Institute of Technology. It uses a 128-bit key to encrypt data in blocks of 64 bits. ... Each block is manipulated in eight rounds, of iterations, followed by a final transformation function. The algorithm breaks the input up into four 16 bit sub-blocks. Each of the iteration rounds takes four 16 bit sub-blocks as input and produces four 16 bit output blocks. The final transformation also produce four 16 bit blocks, which are concatenated to form the 64-bit cipher text. Each of the iterations uses six 16-bit sub-keys, and the final transformation uses four sub-keys, for a total of 52 sub-keys. ... These 52 sub-keys are all generated from the original 128-bit key.

Each iteration of IDEA makes use of three different mathematical operators, each one performed on two 16-bit inputs to produce a single 16-bit output. The operations are:

* Bit-by-bit exclusive or
* Addition of integers modulo 216 (modulo 65536)
* Multiplication of integers modulo 216 + 1 (modulo 65537)

One of the first public-key schemes was developed in 1977 by Ton Ticest, Adi Shamir, and Len Adleman at MIT. The RSA scheme has since reigned supreme as the only widely accepted and implemented approach to public-key encryption. RSA is a cipher in which the plain text and cipher text are integers between 0 and (n-1) for some n. For a long message, the message is broken up into blocks, with each block of size log2n bits.

Encryption and decryption make use of modular arithmetic for some plain text block M and cipher text block C with keys e and d: Where

$C = M^e \bmod n$

$M = C^d \bmod n = (M^e)^d \bmod n = M^{ed} \bmod n$

Both sender and receiver must know the value of n and e; only the receiver knows the value of d. Thus, this is a public-key encryption algorithm with a public key of (e,n) and a private key of (d,n).[5]

RBI's initial tests were done with the public domain PGP encryption algorithm. When pattern recognition tests were done with the DES algorithm and then the triple DES algorithm we discovered much higher Shannon entropy values emerged. The Blowfish algorithm and later EEE algorithms gave the same high entropy values but other parameters differed and it was discovered that our multivariate algorithm could distinguish between various encryption algorithms.

2.1.3 Bit Stream Categorizing Mechanics

Significant research has gone into the problem of categorizing arbitrary bit streams using bit entropy[6]. The usefulness of this method has been limited by the fact that the Shannon entropy calculation[7] is word length and bit stream shift dependent, and because it is only one measure of a bit streams randomness. Determining a bit streams randomness fingerprints the encryption algorithm and is best done with the multivariate statistical categorization algorithms we developed. 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 created. 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. That performance enabled accurate discrimination of encryption algorithms.

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 1.

In Figure 1 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 latter and the full feasibility of this approach begs to be tested.

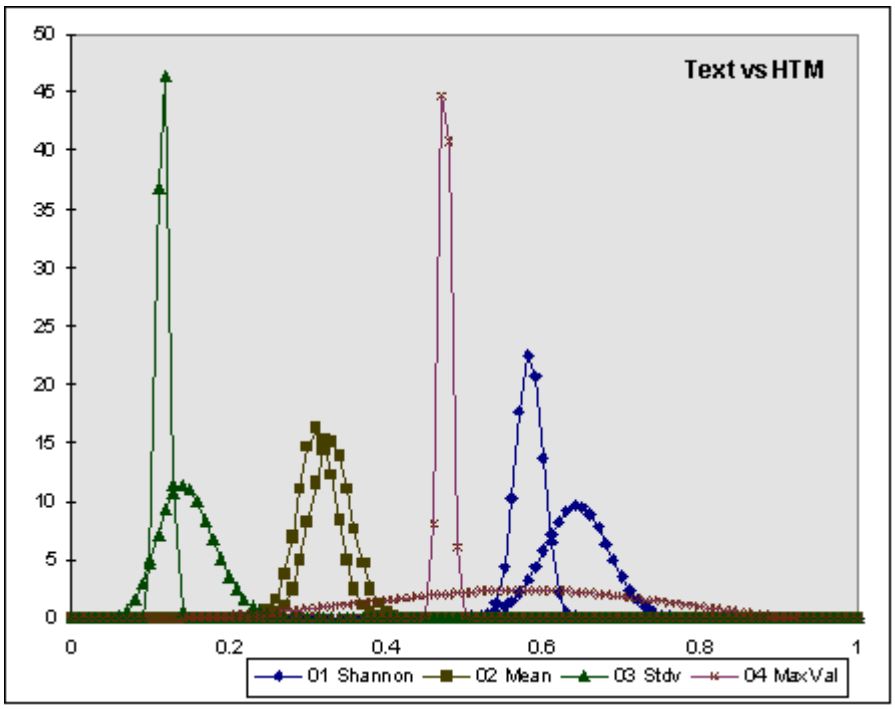

Figure 1 Similar Bit Streams with divergent statistics.

The discrimination between two encryption algorithms is shown in Figure 2. In this figure the Shannon entropy and mean value distributions are obviously adequate to differentiate between a DES and a PGP encryption algorithm.

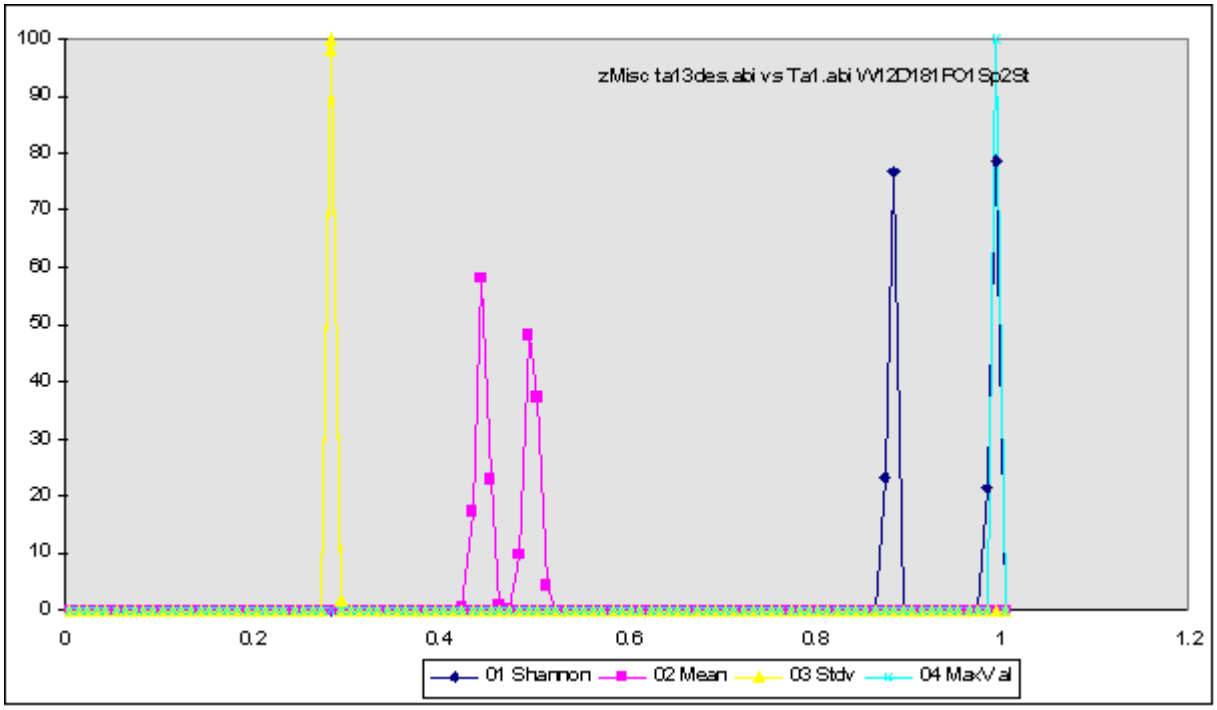

Figure 2 DES and PGP Divergent Statistics of Entropy and Mean

In Figure 3, however, when comparing a DES with a Blowfish algorithm, the distinction between the entropy and mean distributions are obscured. This differentiation must be made by other parameters. Figure 4 shows some of the parameters that can provide the distinction between the DES and Blowfish algorithms.

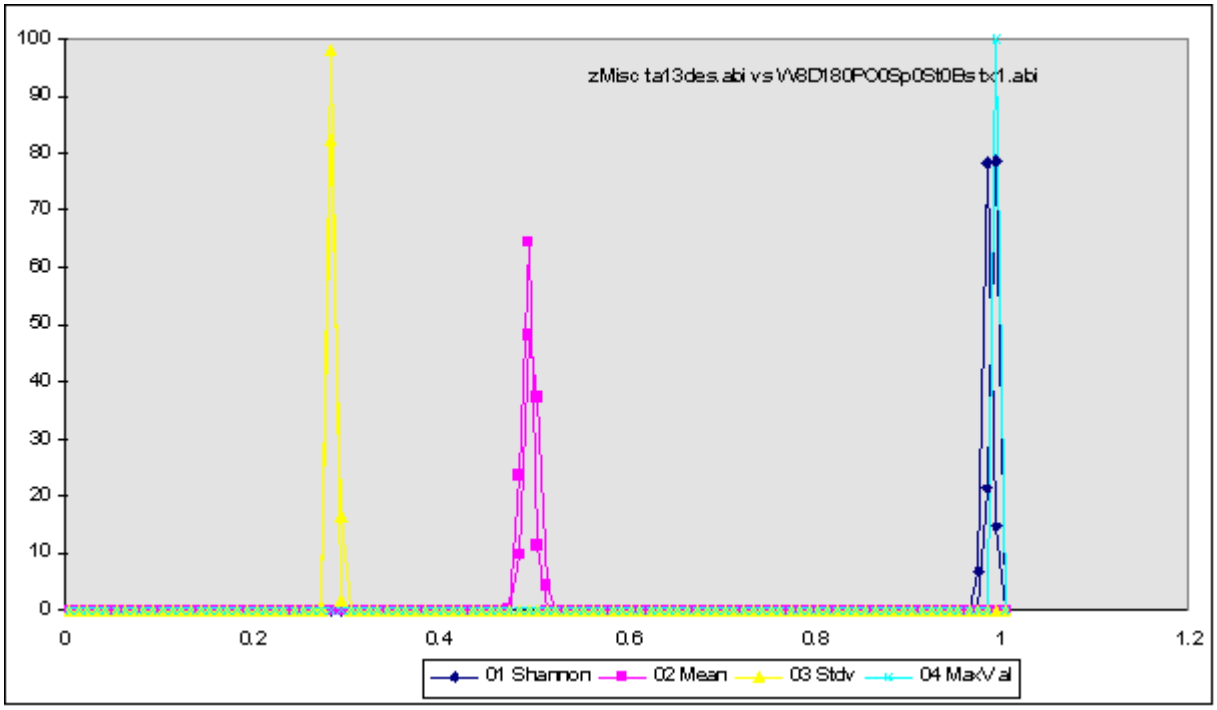

Figure 3 DES and Blowfish Entropy and Mean Distributions

The power of the multivariate statistical pattern algorithm is this grouping of several parameters into one discriminator, and the comparison of the distributions such that only one selection will line up with all the parameters of an unknown sample. The statistics of that match may be calculated to determine the maximum likelihood of any match and the absolute likelihood of a specific match. Notice in the distributions of Figure 4, that all the distributions overlap. However the parameters of the sampled bit stream will fit one set of distributions better than the other set and a determination can be made.

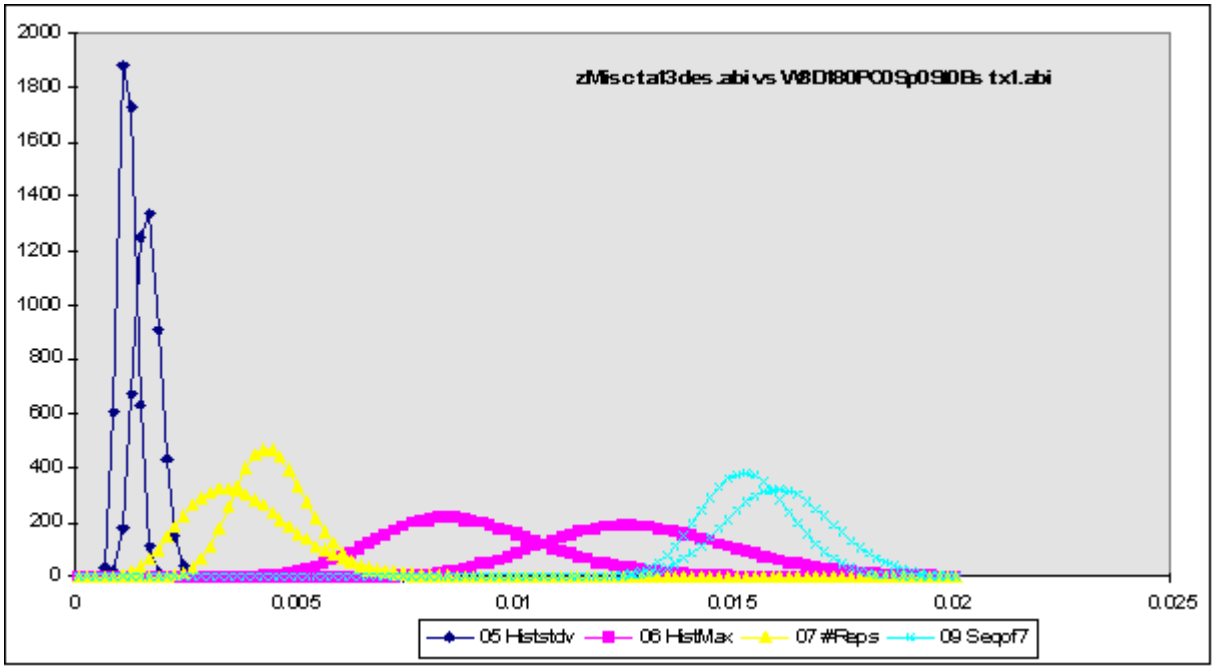

Figure 4 DES and Blowfish Divergent Statistics for Same Entropy Values

3. PHASE I TECHNICAL OBJECTIVES

The overall objective of the proposal is to develop the multi-variate statistical analysis for categorizing encrypted bit streams and investigate it's robustness against various scrambling, compression, and encryption algorithms. The specific objectives of the program can be enumerated as follows:

1. Investigate the various encryption algorithms normally encountered in modern digital communication systems and develop sample sets of these bit streams with various data types and implementations.
2. Refine the calculations of the 13 vector parameters to provide robustness to investigate the dependence on randomness measures which may effect overall classification of an arbitrary bit stream.
3. Refine and optimize the multivariate statistical comparison algorithm to capture and categorize distinguishing statistics in the 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 scrambling and encryption algorithms and between the data types that received encryption.

This multivariate statistical approach has already demonstrated the ability to differentiate data found with various scrambling algorithms. 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 encryption algorithms 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 Encryption Algorithms: During this task the various algorithms of keen interest will be determined with contact with applicable Government offices. These algorithms with various others from prior tests will be applied to a sample set of files 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:

NA 03-150 Task	Cal Days	Hrs	Start	Stop
SBIR Phase I	180	480	03/03/03	08/30/03
Kick Off	1	3	03/03/03	03/04/03
T1A Investigate Data Formats	33	69	03/04/03	04/06/03
T1B Refine Parameter Calculations	33	69	03/23/03	04/25/03
T1C Refine Multivariate Statistics	33	69	04/11/03	05/14/03
T2A Optimize Algorithms	45	100	04/30/03	06/14/03
T2B Evaluate Algorithm Performance	45	100	05/31/03	07/15/03
Final Report	45	70	07/01/03	08/15/03

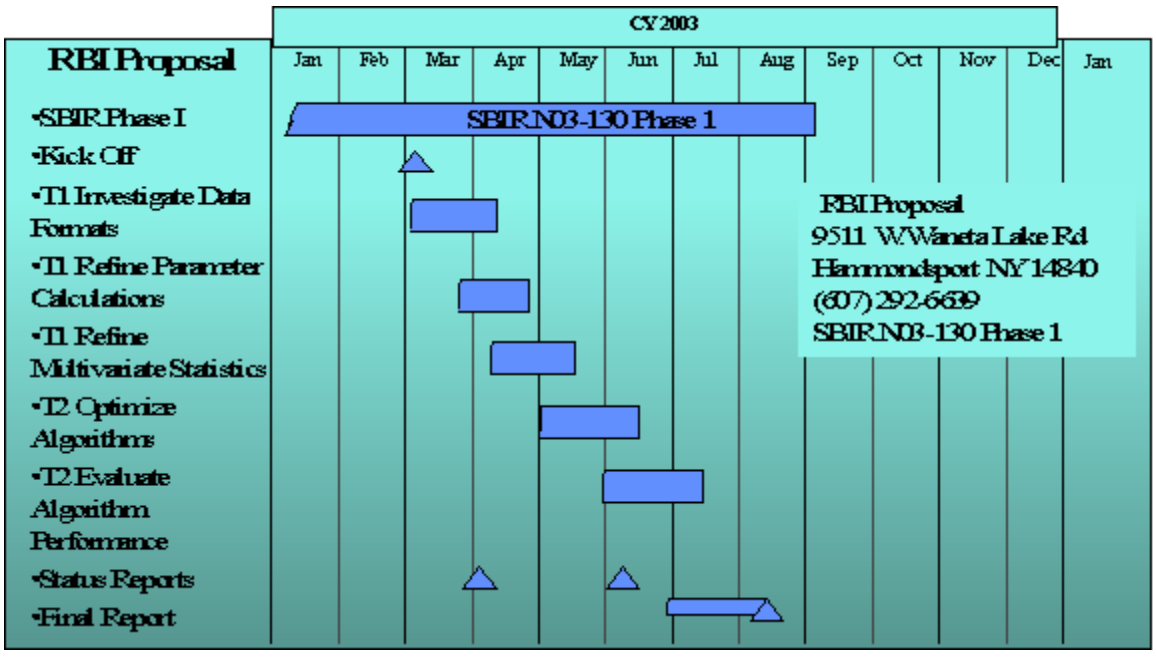

Figure 5 Phase I Schedule (N03-150)

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 encrypted 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: 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

That proposal uses the same classifier core to discriminate arbitrary bit streams with various

framing, format and bit stream slip deviations. Should both efforts be funded Task 1-B and Task 1-C above would dovetail with similar work required for AF03-094 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 intercepted bit stream may be any portion from any bit stream, including portions of files retrieved from hard disks. 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] Stallings, William "Protect Your Privacy, The PGP Users Guide" Prentice Hall PTR@1995

6 [6] "Research Results for Classifying Arbitrary Bit Streams Using Bit Entropy Features" 10 September 1997 by Ed Semplinski, QuesTech, Inc. submitted to AFRL/IFEC

[7] 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 Equation not ported to pdf]

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

9511 W.Waneta Lake Rd, Hammondsport NY 14840 Date: 26-Dec-02
Phone: (607) 292-6639 CAGE Code _____
Title: Encrypted Bit Stream Classification
Topic: N03-150 Multi-Intelligence SIGINT Sensor Processing
Total Proposal Amount $54,402.00

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

Direct Labor (1728 MH/MY)	EST MM	9MO RATE/HR 20hr/wk	EST COST	TOTAL
Principle Engineer	3.3333	45.00	$21,600	
Jr Engineer	0	24.00	$0	
Programmer	0	24.00	$0	
Publications	0	13.00	$0	
TOTAL DIRECT LABOR	480	0.37	ENGRS	$21,600

LABOR OVERHEAD	OH RATE	XBASE	EST COST		
a. IN PLANT	0.7	$21,600	$15,120		
b. ON SITE	0.55	0	$0		
TOTAL LABOR OVERHEAD	.	.	.	$15,120	
SPECIAL TESTING	.	.	.		$0
SPECIAL EQUIPMENT	.	.	.		$0

TRAVEL	EA	RATE	EST COST		
a. TRANSPORTATION	3979		$2,937		
b. PERDIEM	10	85	$850		
c. LOCAL TRANSPORTATION	10	32	$320		
TOTAL TRAVEL	.	.	.	$4,107	

CONSULTANT	Hrs	Rate	Cost		
	0	$750			
TOTAL CONSULTANT	.	.	.	$0	
OTHER DIRECT COST	.	.	.		$0
TOTAL DIRECT COST AND OVERHEAD		.	.	$40,827	

GENERAL ADMIN EXPENCE	0.25	OF COST	.	$10,207
COST OF MONEY	0	OF COST	.	$0
TOTAL ESTIMATED COST	.	.	.	$51,034
FEE OR PROFIT	0.066	OF COST	.	$3,368
*	*	*	*	*
TOTAL ESTIMATE AND FEE OR PROFIT		.	.	$54,402