Using Set Theoretic Estimation to Implement Shannon Secrecy Theory

Albert H. Carlson[1], Richard B. Wells[2], and Robert E. Hiromoto[1]

1) Computer Science Department, University of Idaho, Moscow ID 83844-1010
2) MRC Institute, University of Idaho, Moscow, ID 83844

Abstract -- *In Claude Shannon's paper, Communication Theory of Secrecy, a concept of decrypt is developed and defined in terms of the number of keys of the cipher used, the number of symbols in an alphabet, and the redundancy of a language. In this paper, we show that Shannon's theory can be formalized as a Set Theoretic Estimation (STE) that promises to be applicable to more general decryption techniques. We discuss the unification of Shannon's decryption techniques and STE. A summary of experimental results using the STE implementation is presented, as well as, relevant insights and advances in decryption performance.*

Keywords -- *Set Theoretic Estimation, Shannon's Theory, Decryption, Cryptography*

I. BACKGROUND

Set Theoretic Estimation (STE) is an emerging technique that has seen successful applications in a number of diverse fields that range from speech, signal processing, medical imaging, and image projection problems [1,2,3,4]. One important field is that of class identification, which has application to not only system analysis but also to analysis of regular languages, connectedness problems, discernibility, and communications [5,6]. An excellent summary of the basics of STE is found in an article entitled, ``The Foundations of Set Theoretic Estimation", written by P. Combettes in [7]. Set Theoretic Estimation deals primarily with logic expressions and set theory, many of the symbols normally associated with both fields are used. The notation of both First Order Predicate Logic (FOPL) [8] and standard set operations apply. In this paper we explore the potential applicability of set theoretic space selection and abstraction to the problem of decryption. In this endeavor, we show that STE embeds Shannon's theory of secrecy and provides a practical framework that can be applied to various classes of decryption algorithms.

II. SHANNON THEORY

The paper, Communication Theory of Secrecy [9], describes Shannon's use of language statistics in decryption. In his approach, Shannon defined several measures of interest in the Secrecy paper. One of the more important measures is ``unicity distance." Unicity distance [10] is the average number of encrypted symbols needed to break a cipher. The unicity distance is given as:

$$n \approx \frac{\log_2 |K|}{R_L \log_2 |A|}$$

where R_l is the redundancy of the language, $|K|$ is the number of keys in the cipher, and $|A|$ is the number of symbols in the encrypted alphabet. R_l has been calculated for English, for instance, to have a value of approximately 0.7. In general, languages follow statistical patterns that vary slightly from user to user and message to message. Implied in the unicity distance is the need to find ways to increase unicity in order to keep encrypted data secret. This follows from Information Theory and the effect of large message lengths result in fewer possible decryptions for that message.

Entropy is a measure of how many bits are needed to transfer information. Shannon relates entropy, H(X), and unicity distance using the formula:

$$n = \frac{H(X)}{R_L}$$

where

$$H(X) = \sum_{i=1}^{n} P_{x_i} \log_2 \frac{1}{P_{x_i}}$$

with P_{x_i} the non-uniform probability of symbol x_i appearing.

The redundancy in a particular language, R_L depends on the probability density function for the language. Redundancy provides a clue to the possible role of the duplicated symbol, or collection of symbols. With respect to letter probability, non-uniform distribution of letters in the language can be exploited to correlate their occurences to symbols in the encrypted alphabet. Cryptographers have used this method for many years in guessing keys to encrypted messages. Overall, the probabilities have proven very useful for decryption. Some of these measures include letter and word frequency, word size, and combinations of letters. Shannon empirically determined the redundancy [11], and specified it as a lower limit. Further refinement of the redundancy has not included restrictions placed on letter order by considering word and sentence constructs.

Among the techniques that Shannon explored is the use of language regularity that appear as word repetition and patterns. The premise is that these regularities can be statistically characterized. Since the elemental analysis is done at the character level, identifying word patterns results in letter patterns, which can then be measured and described statistically. The chance of encountering a particular combination of letters is based on their frequency in the language.

Shannon uses the adjacency of letters, called m-grams [9]. M-grams are a run of m letters in a row that appear in a text of length L. Because there are L - m -1, m-grams in a section of code being processed, and m-grams with 2 ≤ m ≤ L are possible, a great deal of information is available about a ciphertext. Shannon uses the m-grams as the basis for his decryption methodology for the shift cipher, a subset of the substitution cipher. Starting at an arbitrary position in a ciphertext, Shannon begins by analyzing the resulting text from decryptions using each of the keys for the cipher. Shift ciphers are an easy case, because there are only 26 possible keys, and easy to illustrate for the same reason. The resulting plaintext from each key is checked for the probability that it is a (the) key that produces a readable message. Because some combinations of m-grams do not result in an understandable message, the probability of one of the keys producing the correct message converges to 1. Other keys will produce plaintext whose probability of being correct converges to 0. Additional symbols are added and analyzed until the probability of one of the resulting plaintext streams becomes 1. The probability of an m-gram occurring is measured empirically by actually counting the number of occurances in a representative corpus. Prior to Shannon's work, cryptographers used compilations of empirical statistics on the subject [12]. It would be easier to calculate a simple probability density function using Baysian statistics [13]; however, the variability of style and language usage [14] make the calculation very difficult. Knowledge about the probability of letters and m-gram frequency is important.

III. SET THEORETIC ESTIMATION

Set Theoretic Estimation (STE) [7,15] is a set theoretic methodology that attempts to find an answer to a problem governed by set membership rules, where solutions are described by intersections of non-empty sets. Each possible solution, called an estimate [7], is listed as a member of the overall solution set. Different properties of the solution, given an input, are specified as groups called property sets. Taking intersections of all of the property sets yields a reduced set of possible answers to the problem. If the correct property sets are applied, the resulting set contains only the correct estimate. The requirements for an STE application are:

1. The problem in question must have a deterministic $f(x)$ and $f^{-1}(x)$
2. The problem must be bounded for an input
3. Each possible answer to the function must be a discrete point, not a range of values
4. Different properties can be stated as sets that differentiate between groups of inputs, inferring set membership

Any information known about the problem is encoded in sets in the solution space. Each rule or constraint is represented by its own unique set. Data may be a member of more than one set, but must be in at least one set to be considered. Information known about both the inputs and about the rules is treated similarly. Rules are expressed as assertions in STE. An assertion *A* takes the set of possible inputs and gives a set of resulting outputs, or solutions, for the operation, *O*, as specified in the rule. For a particular input *i* this output is expressed as

$$O_i = A(i)$$

The set of all possible solutions for all possible inputs, called a ``property set,"

$$O = \bigcup_{j=1}^{m} O_j$$

is found in an m-dimensional solution ``space" known as $\Xi^m$. The space is composed of elements called ``points", each representing a member of a solution set, $\Phi_n = O$. Points in Ξ^m are formally known as the set theoretic estimates. Multiple rules can be asserted, resulting in multiple "property" sets in the solution space. For each rule asserted there is a Φ_n representing the solutions, and Φ^{-1} consisting of all other points in the solution space. Values of Φ are determined by the nature of the application. For instance, a communications application may be expressed in terms of voltages, so Φ would be a range of voltages. There may be many distinct subsets of Φ that are defined by the input to a rule. The total space is defined by

$$\Xi^m = \bigcup_{i=1}^{n} \Phi_i$$

where *i* represents the solution set of each of *n* rules.

Starting in Ξ^m each set Φ_n is considered in turn. Each Φ_n contains only those possible solutions that follow the rule that the assertion describes. Since Ξ^m contains all possible solutions, if a solution exists then the solution *P* must be in the solution space and must be a member of the property set for each assertion. That is, if $\exists P$ then:

$$P \in \Xi^m$$

and

$$\forall \Phi_n \rightarrow P \in \Phi_n$$

Further, if *P* is in each of the solution sets, then it must also be true that

$$P \in \bigcap_{i=1}^{n} \Phi_i$$

If a solution is found in the intersection of the sets, the intersection is said to be ``consistent." If it is not, then the intersection is said to be ``inconsistent." If the resulting solution has exactly one answer, ∃!, the solution is said to be ``ideal."

In order to apply STE to Shannon Theory, one must show that the requirements of STE are met. Unlike Shannon's m-gram approach, which involves finding the most likely m-gram for the input string, our approach eliminates m-grams that are not valid. Forbidden m-grams encountered during decryption indicate that the key used in the decryption is incorrect. Since forbidden combinations of letters cannot occur in a valid string of a message in the target language, impossible m-grams can only come from an incorrect decryption cause by using the wrong key for the message being decrypted (we deal separately with the case of deliberately-inserted ``pads'' of nonsense plaintext in another paper). When forbidden m-grams are encountered the key used to create the decryption can then be eliminated from further consideration. In effect, we look for strings that cannot occur in normal speech, m-grams with a 0 probability. Shannon stopped his consideration of a key when a forbidden m-gram is encountered, but he continued seeking the key whose probability is 1. Eliminating all impossible keys does the same thing, leaving only the one key that is possible. We term this to be the *last-man-standing* technique, where the elimination of all forbidden possibilities leaves the only correct solution.

Implied in the forbidden m-gram technique is the knowledge of which m-grams are permissible and which are not. No ordering is required. The only information needed is whether or not a particular m-gram is allowed. For m-grams of size n there is no relationship between m-grams to determine which are, and which are not, allowed. Noting that for an input of a forbidden 2-gram a value of 1 is returned, and for a non-forbidden 2-gram a 0 is returned easily partitions the set of 2-grams.

$$forbidden = \left\{ \begin{array}{ll} 0 & non-forbidden \\ 1 & forbidden \end{array} \right.$$

The same is true for other m-gram sets.

The set of possible keys is bounded and must be deterministic. If the cipher is NOT deterministic for a key, then it is not possible to recover encrypted information. Properties of the keys are determined by the result of applying possible decryptions and then evaluating the results for feasibility.

To demonstrate our approach, we first look at the application of STE to the shift cipher and then to the substitution cipher. We begin with texts known to be written in English. Samples of literature taken from the Project Gutenburg [16] library in .txt format serve as the corpus for m-grams, as well as the source for encryption. The entire process is automated.

For the shift cipher, we follow Shannon's *m*-gram technique but also include the effects of using randomly chosen words and sentences. Property sets of words and sentences can also be applied to decryption. Breaking up the plaintext resulting from decryption using a possible key must form a string of words in English that can also be understood as a sentence, or group of sentences. Checking the decrypted text for words requires that words be formed by the text. Partial words may be recognized, but words are definitely found in the dictionary for the language. It stands to reason that the number of symbols required for recognition should be close to the length of words, on the average. The average length of an English word is 4.83 characters. Finding the completed word and the beginning of the next word defines the first word and is necessary for unambiguous word identification. At least one additional letter beyond the first word is required to terminate the first word. The average of 4.83 symbols plus 1 symbol for bounding the first word gives an expected average of 5.83 symbols. While more than 1600 tests are run on the data, only 41 texts are used as data. Using a larger number of texts would push the number of symbols towards the average. The results are within the expected number of symbols, with deviation.

Words and sentences as property sets decrypted all texts correctly. Words have all m-grams contained within words, but do not contain m-grams formed at the boundary of words. Sentence structure restricts which words may be placed together. By restricting the word combinations, m-grams spanning the words are limited. A full set of allowable m-grams is defined by the allowable words and sentence structure. Words and sentences give more total information than m-grams alone, but they require more symbols to initially apply. While words and sentence structure are more accurate they also require more input data. For more complex encryption methods with unicity distances longer than the average word length, words and sentences add to the effectiveness of a pure letter based, or m-gram based, approach.

A summary of the result follows:

Test: *Shift cipher, m*-grams. Words and Sentences
***m*-grams used:** 2 , 5
Number of tests: 1128
Incorrect/no decryption possible: 0
Correct: 100%
$\mu = 5.0$ $\sigma = 0.5545$

where μ is the sample mean and σ is the sample variance.

Applying the property sets to a more complex code, such as the substitution cipher, showed similar success. In this application, the sets use m-grams of various lengths applied to a substitution cipher. By evaluating whether or not a key results in a forbidden string of the target language, keys are kept or eliminated. When a letter is unambiguously identified, it is possible to reduce the number of possible bindings for each key. Each m-gram is its own property set. Taking the intersection of each set is as simple as eliminating any key that does not hold for a property set. Data on allowable *m*-grams was gathered from a group of fiction and non-fiction works considered classics. Works were chosen to cross genre, authors, and the era in which each was written. The library spans Shakespeare to Asimov. Tests were conducted on the same texts to verify that decryption was possible with the property sets constructed from the texts. A point in the text, randomly selected, started a string of symbols in ciphertext presented to the STE decryption algorithm. Decryption continues until all symbols seen are decrypted. Each file was presented 12 times and the number of symbols required for the solution was recorded.

In this set of tests, files are encrypted from a randomly selected point in the file. Keys are randomly selected for the encryption and checked as the key is assembled. Decryption is terminated when either an incorrect key is selected or when the text that has been seen is fully decrypted. Forbidden m-gram sets for $2 \leq m \leq 6$ and $m = 13$ are used in the decryption process. Using these sets, the following result is gathered:

Test: Substitution cipher, *m*-grams only
***m*-grams used:** 2 – 6, 13
Number of tests: 492
Incorrect/no decryption possible: 0
Correct: 100%
$\mu = 17.46$ $\sigma = 5.53$

If the 5 worst results are removed from consideration, the value of μ drops to 17.05 symbols and value for σ falls to about 3.36 symbols. Similar results were found for tests on a shift cipher.

IV. CONCLUSION AND FUTURE WORK

In this paper, we show that a STE approach to decryption provides a compatible framework to formally embed Shannon's work in this area. To our knowledge, this is the first application of STE in decryption analysis. Within the framework of STE, keys are discrete and bounded, m-grams form the required property sets, and decryption using a key is quick and accurate in unambiguously determining set membership. Testing basic techniques in decryption, STE yields results that indicate its generality not only within Shannon's approach but within a broader class of decryption techniques.

Shannon's theory was developed during a time when computers were unavailable for general research. Data was gathered by groups of people manually reviewing, collecting, documenting, and analyzing data. Costs were extremely high and the time to complete a single study limited the quality of information that laid the foundations of their analysis. Given the advanced technology in today's computational systems, a reinvestigation of data related to the redundancy of languages may be useful. M-grams have proven to be of use in exploring decryption. To date only arbitrary m-grams have been employed. The use of words and sentences have not been considered. Words are variable length m-grams, and sentences are composed of varying number of words. Words limit the further combination of subsequent m-grams, or words. Sentences made up of words also limit the possible combination of m-grams. Additionally, using semantical information embedded within words and sentences, allows keys to be isolated more rapidly. Recent research by the authors into STE decryption of a substitution cipher has indicated that the unicity distance can be lower than generally believed. This research is on going.

REFERENCES

[1] Barakat, R. and Newsam, G, "Algorithms for Reconstruction of Partially Known Band-Limited Fourier Transform Pairs from Noisy Data II: The non-linear Problem of Phase Retrieval", Journal of Integral Equations, v. 9, no. 1 (supplement), pp 77 – 125, July, 1985

[2] Benidir, M. and Picinbono, B, "Nonconvexity of the Stability Domain of Digital Filters", IEEE Trans. Acoustics, Speech, Signal Process, v. 38, no. 8, pp 1459 – 1460, Aug, 1990

[3] Y. Censor, "Parallel Application Of Block-Iterative Methods In Medical Imaging And Radiation Therapy", *Math. Programming*, vol. 42, no. 2, pp 307 – 325, 1988

[4] Crombez, G, "Image Recovery By Convex Combinations Of Projections", *Journal of Math. Analysis Applications*, v. 155, no. 2, pp 413 – 419, Mar. 1991.

[5] S. R. Kulkarni, and D. N. C. Tse, ``A Paradigm for Class Identification Problems," IEEE Transactions on Information Theory, v. 40, no. 3, 1994, pp. 696 - 705

[6] Wells, R.B., ``Application of Set-Membership Techniqes to Symbol-by-Symbol Decoding for Binary Data Transmission," IEEE Transactions on Information Theory, v.42, no. 4, 1996, pp. 1285 - 1289

[7] Combettes, P.L., "The Foundations of Set Theoretic Estimation," Proceedings of the IEEE, v. 81, no. 2, pp. 182-208, February 1993

[8] Hunter, G., ``Metalogic, An introduction to the Metatheory of Standard First-Order Logic," University of California Press, Berkeley, CA, 1971

[9] Shannon, C.E., "Communication Theory of Secrecy Systems," Bell System Technical Journal v. 28, pp. 656-715, 1949

[10] Wells, R.B., "Applied Coding and Information Theory," Prentice Hall, Upper Saddle River, NJ, 1999

[11] Shannon, C.E., "A Mathematical Theory of Communication," Bell System Technical Journal v. 27, pp. 379-423 and 623-656, 1948.

[12] Pratt, F., "Secret and Urgent," Blue Ribbon Books, Garden City, NY, 1939

[13] Ross, S., "A First Course in Probability," MacMillan Publishing Company, NY, 1976

[14] Morton, A.Q., "Literary Detection," Charles Scribers, and Sons, New York, NY 1979

[15] S. G. McCarthy and R. B. Wells, "Model Order Reduction for Optimal Bounding Ellipsoid Channel Models," IEEE Transactions on Magnetics," v. 33, no. 4, pp. 2552-2568, July 1997

[16] The Gutenburg Project, Intenet, http://www.gutenburg.net