Using Fuzzy Logic in Evaluating User Tabled Correlation Rules for COMINT

May 17, 2000

John M. Palmer
mailto:jpalmer@ausinfo.com
Austin Info Systems
301 Camp Craft Road
Austin, TX 78746

ABSTRACT

This paper presents a methodology for performing level-1 fusion involving organizational structures and for determining which structures, if any, can be recognized. Available data consists of contact reports received via various communications circuits. Information for correlation arises from three main sources: parametric signature data such as radio frequency, nominative categorical information such as unit identifier, and location information such as intercept position. Both individual and aggregations of such data separately give rise to correlation sub-scores. Each sub-score provides evidence about whether a new report correlates with an existing entity. The paper identifies methods and reports some preliminary results correlation, which use the operators of fuzzy logic, and adaptive weighting to enforce analysts specified correlation rules.

1. Introduction

COMINT is a valuable source for assessing threat intentions and plays an important role with respect to the identification of units, intelligence systems, and C2 organizations. While improved surveillance provides plentiful reports, correlation techniques remain relatively time-intensive and human intensive. One of the problems with understanding reported data is that only for a fraction of the intercepted entities are the associated signals, features, networks and functions known in any detail. Often, the parent organizations themselves are not well understood.

These needs serve as a primary motivation for developing automated correlation tools for the COMINT analyst-the standard goal being to reduce the COMINT analyst's workload by using his/her own rules to perform correlations. Automating, where possible, allows the analyst to concentrate on the more important tasks he/she can do best, such as the recognition of emerging patterns and the comparative evaluation of complex alternative interpretations or the battlefield.

This paper presents a methodology for performing level-1 fusion involving organizational structures and for determining which structures, if any, can be recognized. Available data consists of contact reports received via various communications circuits. Each report often contains information about more than one type of entity that is needed in constructing a global picture. Report data may be provide information that describe nodes (units), links (communication paths), and/or networks (collaborating collection of communicating nodes). Taken separately, theses elements provide a sometimes-confusing web of interrelationships. It is the function of the correlator to make the most of this data.

Report Documentation Page

Report Date	**Report Type**	**Dates Covered (from... to)**
17052000	N/A	-

Title and Subtitle Using Fuzzy Logic in Evaluating User Tabled Correlation Rules for COMINT	**Contract Number**
	Grant Number
	Program Element Number
Author(s) Palmer, John M.	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Austin Info Systems 301 Camp Craft Road Austin, TX 78746	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) Director, CECOM RDEC Night Vision and Electronic Sensors Directorate, Security Team 10221 Burbeck Road Ft. Belvoir, VA 22060-5806	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UNLIMITED
Number of Pages 14	

2. Overview

The COMINT correlator assesses whether newly reported information about a communication node or net will be paired (correlated) with existing information or whether the new information arises from a new and heretofore unseen entity. Because the reported information can be incomplete, erroneous, or have irregular specificity, differing methods of analysis can produce contradictory results.

One approach in resolving inconsistencies is to declare problematic reports to be ambiguous and await additional information for the proper resolution of the report. Other correlators form multiple hypotheses about the disposition of the new data.[1] The current COMINT correlator takes the first approach – decisions are deferred through ambiguity formation. When the ambiguity rate[2] is too high, the clarity of the tactical picture is degraded so it is clearly of interest to minimize the ambiguity rate while maintaining accurate correlation.

Information for correlation arises from three main sources: parametric signature data such as radio frequency, nominative categorical information such as unit identifier, and location information such as intercept position. Both individual and aggregations of such data separately give rise to correlation sub-scores (S_i ; $i = 1, 2, \dots , n$).

Each sub-score provides evidence about whether the new report correlates with an existing entity. One technique for combining the scores (S_i) is to use weights (w_i) to form a final score (F) that is a linear combination of the sub-scores:

(1)
$$F = (w_1 S_1 + w_2 S_2 + \dots + w_n S_n) = \sum_{i=1}^{n} w_i S_i$$

or, if the final score is to be normalized:

$$F = \frac{(w_1 S_1 + w_2 S_2 + \dots + w_n S_n)}{(w_1 + w_2 + \dots + w_n)} = \sum_{i=1}^{n} w_i S_i \Big/ \sum_{i=1}^{n} w_i$$

Although it seems counter-intuitive that "truth" can be attained so cheaply as through simple addition, Fisher, Hotelling, and Mahalanobis independently discovered a statistical technique called *discriminant analysis* which utilizes the functional form in (1) and which may have been the progenitor of this commonly encountered correlation calculation [1]. The linear discriminant function strives to minimize the frequency of misclassification[3] by calculating the weights based upon a large, representative sample of observations with known classifications.

There are several reasons this approach is questionable in the current application. First, the development of the discriminant function is based upon an assumption that the underlying sub-scores follow a joint normal distribution. This assumption is a bit of a stretch for nominative data such as echelon that has only a very limited number of discrete values. At the same time, it must be recognized that it is customary to

[1] A multiple hypotheses correlator maintains a number of views of how all the contact reports could plausibly be arranged into entities (tracks). Each view of how all the reports can be arranged is a hypothesis. Probabilities are assigned to each hypothesis. This approach provides multiple global views of how otherwise ambiguous reports may be integrated into tracks. This is accomplished at the expense of making ambiguous which view, if any, is correct.

[2] The Ambiguity Rate is the fraction of ambiguous report dispositions generated per report received.

[3] The classifications are "the new report is an update to an existing entity" and "the new report is a new entity".

make a normality assumption in non-normal settings because of the robustness of many of the techniques that are based upon the normal distribution.

A more important criticism involves the method of determining the weights (w_i). The computation of the weights used in discriminant analysis is achieved by maximizing the size of the discriminant's linear contrast over a representative collection of reports with known dispositions.[4] While ground truth labeled scenarios are available for ELINT correlation[5], no such labeled scenarios were available for COMINT. Lacking substantive weights derived from representative data, some have been led to the use of subjective values for the weights in (1).[6] Such weights are difficult to devise or to justify – either for one person or for a committee because, lacking ground truth tagged scenarios and a powerful measure(s) of correlator effectiveness, it is difficult to produce demonstrable evidence that one set of weights is superior to another.[7]

A third criticism of this approach is that the weights in (1) are static – reflecting the assumption that the information embodied in each sub-scores, S_i, has an unchanging value in the correlation process (i.e. for $t > 0$, $w_i(t)$ is constant). In fact, this may not be the case. For example, in a setting where several units are massing their forces, the value of location information will initially be a good discriminator among the units when they are geographically separated but will decline in value as elements of the forces intermingle. Likewise, the value of knowing that a report and a candidate's branch match and is "Armor" is probably less valuable in a dessert setting than in a jungle environment because the appearance of that type of force will be more common in the former setting than in the latter.

3. Methods

Recognizing these shortcomings prompted the development of an approach whose objectives were to produce a correlation procedure which:

1. Used adaptive weights appropriate to the data evaluated
2. Accurately reflected analyst methods of association
3. Allowed methods to be customized to the target environment

Because the difficulties encountered in establishing and interpreting the coefficients in equation (1) are exacerbated by a lack of calibration data, the use of such linear combinations is relegated to a secondary role within the prototype. Instead, the primary correlation approach relies upon an appraisal of the suitability of one or more user specified correlation rules. These rules initiate correlation actions when specified antecedent conditions are fulfilled as adduced by fuzzy logic operators. Currently, these rules are obtained at run-time from an editable file provided by the user

Only when the evaluation of these rules is ambiguous does the correlation algorithm rely upon the calculation of a linear combination of sub-scores. In such cases, the magnitude of the weights is adjusted

[4] That is, each data element is known to be either an "update" or a "new track".

[5] During 1990, the author participated with APL, NSWC, and SPAWAR in tests of multiple correlators at NOSC (now NRAD). Ground truth was provided from both simulated and real data – the later being provided by NRL.

[6] The OverTheHorizon (OTH) correlator, circa 1985, utilized such weights.

[7] Typical measures of effectiveness include *new track ratio*, *track purity*, and *miscorrelation rate* as well as the more comprehensive measures *clarity*, developed by the author, and *fidelity*, developed by APL. All of these measures are routinely used to describe afloat correlator performance but do require a ground truth labeled scenario.

to enhance the discrim inatory ability of the correlation variables within the available candidate set.[8] In this paper, the specific linear combination used is referred to as the *Preponderance of Evidence*. The characteristics and method of calculation of this score are discussed first.

3.1. Assumptions:

It is assumed that an initial candidate selection and specific correlation have been performed[9]. Assume that there are m (> 0) remaining candidates. Further, assume that there are n_{Var} (> 1) correlation variables, each serviced by their own comparative function.[10]

It is required that each sub-process, when applied to a report and a candidate, produce a result that can be cast into one of three mutually exclusive results. These outcomes are that:

1. The variable provides evidence that the report is an update to this candidate;
2. The information element provides evidence that the report is not an update to this candidate;
3. The information element provides insufficient evidence for either of the previous assertions.

3.2. Transformation of Scores:

Because of this assumption, the results of each correlation sub-process (S_{ij}) can be equivalently categorized in terms of thresholds c_i and n_i, which may vary from one correlation variable to another. Such an "equivalent scoring" categorization is defined in the subsequent table.

Comparative Results for Information Element i and Candidate j	**Equivalent Threshold Condition**
Update candidate j	$c_i \le S_{ij} \le 1$
Do not update candidate j	$S_{ij} \le n_i$
Inconclusive	$0 \le n_i < S_{ij} < c_i$

[8] Other approaches for assessing agreement/disagreement or fot for measuring the concordance of information can be found in Anderson [2].

[9] Candidate Selection is that process which eliminates those historical "tracks" which are incompatible with the new observation. Specific Correlation are associations formed straightaway because of the match of a unique identifier.

[10] The comparative function produces correlation scores on individual elements by methods appropriate to the data. For example, for elements whose value is represented in relatively short strings, the sub-score computed is a variant of the Hamming metric that reports the maximum concordance of the strings without permutation. For continuous univariate data, the Student's t CDF is used.

The modular nature of the procedure accommodates the use of other distributions or measures (e.g. a beta distribution in place of a t).

Of future interest are comparative methods, which measure the association between larger bodies of text that may be reported (e.g. analysts comments ~ 500-2000 words). One method that is under investigation is based on the approach used in SCAM, a software package developed (and in use) to detect plagiarism.

Although each variable may have a unique set of thresholds, each is transformed to a constant scale by the application of a simple linear mapping:

Correlation Sub-Process Results for Variable i and Candidate j	**Numerical Representation**	***Transformed Value (* a$_{ij}$ *)***
Update candidate j	$S_{ij} \geq c_i$	$a_{ij} = 100 (S_{ij} - c_i) / (1 - c_i)$
Do not update candidate j	$S_{ij} \leq n_i$	$a_{ij} = 100 (S_{ij} / n_i - 1)$
Inconclusive (Ambiguity)	$n_i < S_{ij} < c_i$	$a_{ij} =0$

These mappings transform each sub-score to the interval [–100, 100]. The more negative the value of a_{ij} , the more strongly the sub-score argues against updating the candidate. Similarly, large positive values of a_{ij} reflect support for an update decision. The transformation of *S* to *a* retains whatever strength and discriminatory power embodied within the original scoring.

3.3. Construction of the Agreement Table.

For each variable (*j*) and each candidate (*i*), the transformed sub-scores are tabled as

$$(a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_{\text{var}}} \\ a_{21} & a_{22} & \cdots & a_{2n_{\text{var}}} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m\,2} & \cdots. & a_{mn_{\text{var}}} \end{bmatrix}$$

The transformed sub-scores for each candidate that passed the initial selection criteria comprise one row of this table. These entries are the basis for subsequent inference. Each tabled variable is weighted to reflect both the user's perceived value of the variable and its discriminatory capability within the context of the current environment.

User weights are obtained from an initialization file. If u'_j denotes the user supplied weight for correlation variable *j,* then these weights are normalized by dividing by their sum so that

$$u_j = \frac{u'_j}{u'_1 + u'_2 + ... + u'_{n_{\text{var}}}} = \frac{u'_j}{\sum_{i=1}^{n_{Var}} u'_i}$$

An additional weight (v_j) factor represent the uniqueness of correlation variable j. Uniqueness is measured through a proxy variable that is merely the running ratio

$$v_j = n_d / n_r$$

where there ar   e n_d distinct values among n_r instances wherein the information element was reported.[11]

(2) $$a'_{ij} = v_j \; u_j \; a_{ij}$$

Once normalized, each column of the table (a_{ij}) is then examined to assess the discriminatory power of the corresponding correlation variable within the available collection of candidates. This is accomplished by tallying each score based upon its sign – a sum of the positive scores and a sum of the negative scores. Specifically, these positive and negative column totals are obtained for each correlation variable j as:

$$T_j^+ = \sum_{i,\, a_{ij} > 0} a_{ij}$$

$$T_j^- = \sum_{i,\, a_{ij} < 0} a_{ij}$$

These totals are then factored back into the correlation sub-scores along with the normalized user weights:

(2) $$a'_{ij} = \frac{m \, v_j u_j \, a_{ij}}{T_j^+} \quad if \;\; a_{ij} > 0$$

$$a'_{ij} = -\frac{m \; v_j \; u_j \; a_{ij}}{T_j^-} \quad if \;\; a_{ij} < 0$$

Besides factoring in the user-supplied weights, the revised scores (a'_{ij}) re-scale each score based upon the average propensity of the variable to score "for correlation" (i.e. a candidate's score is positive) or "against correlation" (i.e. a candidate's score is negative).

The sum over all candidates of each transformed variable is zero when scores carrying both signs are present in a column. Because the sum of the transformed scores is zero, the magnitude of any negative score will be increased in the presence of a large number of positive scores in order to achieve the zero sum. Likewise, the magnitude of positive scores will be increased in the presence of a large number of negative scores in order to achieve the zero sum. Consequently, the net effect of the transformation for a particular variable is to magnify the weight of atypical scores.[12]

3.4. Preponderance of Evidence Score:

[11] Initial values of these weights are provided with the default rule set. These weights differ with the type of data being addressed (e.g. simulator vice real).

[12] A cautionary note: The issue of candidate based weighting requires additional attention. More recent investigations indicate that the net effect of using the last factor (Tj / m) is negligible (except in POE calculations) and, in some instances, may promote the formation of ambiguities-exactly counter to its intent.

After the transformation (2) is applied to obtain:

$$(a'_{ij}) = \begin{bmatrix} a'_{11} & a'_{12} & \ldots & a'_{1n_{\mathrm{var}}} \\ a'_{21} & a'_{22} & \ldots & a'_{2n_{\mathrm{var}}} \\ \ldots & \ldots & \ldots & \ldots \\ a'_{m1} & \dot{a}_{m\,2} & \ldots. & a'_{m\,n_{\mathrm{var}}} \end{bmatrix}$$

then the score for each candidate (i) is obtained by summing over the columns within its associated row. In this paper, this sum is referred to as a candidate's *Preponderance of Evidence* score. Algebraically, it is

$$P_i = \sum_{j=1}^{n_{Var}} a'_{ij}$$

The Preponderance of Evidence (POE) represents a tally "for" or "against" correlation over the "voting" correlation variables. Apart from the adaptive weighting, the POE is the usual measure of correlation. If the collection of information elements generally favor a particular candidate, then its Preponderance of Evidence score will be exceed those of the other candidates. Because the transformation ($a_{ij} \rightarrow a'_{ij}$) magnifies the weight of atypical scores assigned to a candidate, atypical scores contribute more to a candidate's sum than the "routine" scores assigned to "typical" candidates. In this way, the performance of the correlation variables on each particular candidate set is used to select weights that enhance the discriminatory power of P_i.

3.5. Correlation Rules

When numerous variables are available to provide information regarding correlation, it may be difficult to grasp what, if any, scalar thresholds for a measure like POE are meaningful in determining correlation. A reason for this is that the POE inherits the variability of all its components whether or not the component is a definitive summand in the POE.

Moreover, measures like POE have little or no explanatory power when a large number of variables are included. During an audit, it is meaningless to ask why a POE driven correlator made a decision in any particular instance because the only explanation is that the POE exceeded or failed to exceed a (sometimes farcical) threshold. This inability to meaningfully partition the observation outcome space hinders domain experts in identifying misapprehensions in the automated correlation process.

In discussions with domain experts, the developer may be confronted with contradictory statements such as "location is critical" or "location is meaningless." To overcome these issues, the approach adopted was to allow the user to encapsulate "correlation concepts" in an editable table. Each concept constitutes a sufficient reason to correlate a new report to an existing entity. The use of correlation concepts allows the functioning of the correlator to be customized to the target and environment.[13]

[13] As with the weights assigned to informational elements, best performance appears to be attained by using somewhat different rule sets in differing environments.

These correlation concepts can be illustrated by assuming we are "intercepting" local TV stations, trying to identify the source of the intercepted broadcast. The data elements available might be *Channel Number* (6, 24, etc.); *Spiffy Name* (e.g. Keye42, Fox7, News36) and *Network* (CBS, NBC)

Without intending to convey any correlative merit to these propositions, the following examples illustrate the structure of the correlation rules a user might provide.

Simple assertions:

Correlate the report to the node **if** *Channel Number* matches.

Conjunctions:

Correlate the report to the node **if** *Channel Number* matches **and** *Spiffy Name* match.

Disjunction's:

Correlate the report to the node **if** Channel Number matches **or** Network matches.

Compound:

Correlate the report to the node **if** Channel Number matches **and** either Spiffy Name matches **or** Network matches

Because of the idiosyncrasies of reporting, certain data is often only sporadically present. As a result, some comparisons can't always be made. To allow the user flexibility in addressing situations of this kind, another rule format is offered. This structure is illustrated with the following example:

Enthymemes:

Correlate the report to the node **if** Channel Number matches **and** there is **no contradictory information** from either Spiffy Name **or** Network.

It is the unstated premise of such a rule that if the missing data were present (and non-ambiguous), that data would usually support the correlation decision. In order for a rule that does not make such a provision to be construed as applicable, all referenced data must be reported and unambiguous.

Allowing the user to encapsulate and record those conditions and only those conditions that justify correlation activities allows questions to be resolved regarding anecdotal acts of correlation or non-correlation.

Where there is a question as to why a correlation occurred, it is possible to point to a specific correlation rule that was applied. When a question arises in a particular instance, it will be found that correlation was or was not performed because of the presence or absence of a rule whose applicability was (or was not) supported by the available data. User identified errors can then lead either (a) to a refinement, addition, or deletion in the user-supplied rules or (b) to an identification of data that is unreliable (and whose weight should therefore be reduced), (c) the determination that a comparative process is inadequate or (d) an individual decision threshold is erroneously set.[14]

[14] In some cases, the domain expert may feel that no automated rule can be trusted to perform adequately. Such a situation might arise, for example, if the analyst is privy to side-information not accessible to the automated system. At least a partial solution to this circumstance is to provide an additional rule set that embodies the "ambiguity concept". Observations satisfying any of these rules would be set aside for analyst disposition. (This feature is not currently implement.)

Couplin g correlator actions with specific rules and data and individual thresholds furnishes a more understandable correlation rationale whose premise the user or domain expert can accept or reject.

3.6. Rule Specification

If there are n_{Var} information elements that the user may incorporate into one or more correlation rules, then theoretical limit to the number of rules is $4^{n_{Var}}$ - 1. In practice, the user will probably be able to identify all the conditions that warrant correlation processing with a limited number of concept statements.

Assume, for definiteness, that the user develops *R* rules. These are encoded in a table or matrix:

$$(r_{k\,j})^* = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1\,n_{\text{var}}} \\ r_{21} & r_{22} & \cdots & r_{2\,n_{\text{var}}} \\ \cdots & \cdots & \cdots & \cdots \\ r_{R1} & r_{R2} & \cdots\cdot & r_{R\,n_{\text{var}}} \end{bmatrix}$$

where the following conventions are employed:

The element r_{kj} describes, in rule k, the treatment of the comparative score available for variable j. The interpretations currently used are codified below.

Coding of r_{kj}	Meaning
0	Correlation variable *j* is not referenced nor used in rule *k.*
1	Correlation variable *j* must be perceived to match the same information element in the candidate.[15]
2	Correlation variable *j* is considered only when the variable has been reported in both the new observation and the candidate; otherwise, the variable is not used in this rule.
3	If both data elements are present and the variables are perceived to match satisfactorily, the data is used; otherwise, the variable is not considered.

Rules are interpreted in the matrix as being related through the logical operators “and” across columns and “or” across rows; i.e., the matrix $(r_{ij})^*$ is interpreted as:

Rule 1: ($r_{1\,1}$ and $r_{1\,2}$ and $r_{1\,3}$ and ... and $r_{1\,Nvar}$)
or
Rule 2: ($r_{2\,1}$ and $r_{2\,2}$ and $r_{2\,3}$ and ... and $r_{2\,Nvar}$)
or
...
or

[15] As noted earlier, this “perception” may be guided by any separate algorithm – the most common ones being either some variant of string matching or some form of statistical distance assessment.

$$\textit{Rule R: } (\, r_{R\,1} \textit{ and } r_{R\,2} \textit{ and } r_{R\,3} \textit{ and } \ldots \textit{ and } r_{R\,Nvar} \,)$$

For example, the codifications of the initial illustrative assertions are:

Sample Codifications			
Channel Number	Unit ID	*Network*	Rule
1	0	0	Correlate the report to the existing entity if *Channel Number* matches.
1	1	0	Correlate the report to the existing entity if *Channel Number* matches and *Spiffy Name* matches
1	0	0	*Correlate the report to the existing entity if Channel Number matches* or *Network* matches
0	0	1	
1	1	0	Correlate the report to the existing entity if *Channel Number* matches and either *Spiffy Name* matches or *Network* matches
1	0	1	
1	2	2	Correlate the report to the existing entity if *Channel Number* matches and there is no contradictory information from either *Spiffy Name* or *Network.*

3.7. Rule Weights:

Each rule is assigned a weight – either by the user or by the system. Weights are of value in adjudicating conflicts wherein different rules argue for different decisions. The rules, adjoined with their weights, are represented as:

$$(r_{k\,j}) = \left\langle \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1\,n_{\text{var}}} \\ r_{21} & r_{22} & \cdots & r_{2\,n_{\text{var}}} \\ \cdots & \cdots & \cdots & \cdots \\ r_{R1} & r_{R2} & \cdots. & r_{R\,n_{\text{var}}} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \cdots \\ \omega_R \end{bmatrix} \right\rangle$$

3.7.1. If the user does not wish to supply weights, weights are provided for them via

$$\omega_k = \max{}_j \{u_j \mid r_{kj} \neq 0\}$$

where the u_j are the user supplied variable weights discussed earlier. With this weighting assignment, each rule inherits the importance of the most important correlation variable it references.[16]

3.8. Rule Evaluation:

For each user-supplied correlation rule, a "truth value" is estimated based upon the correlation sub-scores calculated earlier $(a'_{i\,j})$ and the combinatoric methodology of fuzzy logic [3]. Employing this approach, each of the scores $a_{ij} \in [-100,100]$ can be viewed as a membership function that portrays the degree of belonging evidenced in variable *j* between the report and candidate *i*. Since each rule is a conjunction of specifications about the state of these variables, the function or "truth value" of rule *k* for candidate *i* is

(3) $$t_{i\,k} = \omega_k \min{}_j \{a^*_{ij}\}$$

where for candidate *i* and variable *j* in rule *k*, the value of a^*_{ij} is determined according to the table:

$a^*_{ij} =$

a'_{ij} is not considered if $r_{kj} = 0$
a'_{ij} is used regardless of value if $r_{kj} = 1$
a'_{ij} is used if $r_{kj} = 2$ and variable j is unambiguously reported and is not considered otherwise
a'_{ij} if $r_{kj} = 3$ and variable $a'_{ij} > 0$ and is not considered otherwise

In (3), t_{ik} also includes the scaling factor (ω_k) that reflects the importance of the rule.[17]

[16] In the default installation, rule weights are always automatically computed. A rule hierarchy may emerge as analyst feedback is obtained.

[17] The *min* function used in (5) mirrors the performance of the Aristotelian "*and*" when the membership function is Boolean {0,1}. The current implementation uses the standard fuzzy logic operators. However, code has been developed (but not yet implemented) that optionally utilizes operators from either the Yager or Weber family of t-norms and t-conorms. An initial investigation indicates that by taking the parameter ω (see [?]) as ω = 1+ 1 / m in the Yager family and ω = 1-1 / m in the Weber class of functions, correlation can be improved. Work with significant test sets is needed here.

After each rule is scored for candidate i, the aggregate value of all rules is computed for the candidate by taking the maximum:

$$t_i = \max_k \{t_{ik}\}$$

From a fuzzy logic viewpoint [3], t_i is the measure of the "truth" of (rule r_1 or rule r_2 or … or rule r_r). Positive values indicate the rule is applicable to one degree or another while negative values denote the obverse. Zero values indicate the rule stands mute.

3.9. Candidate Scaling and Sorting:

Once the t_i values are obtained for each candidate, they are ranked and sorted for use in the decision logic. Ranking is performed in terms of percentiles. This preserves the original ordering of the t_i but provides additional information about the magnitude of variation in the t_i that is useful in arriving at an update decision.

Denote the percentile rank of t_i by p_i. The p_i are calculated using a Student's t distribution and then sorted in terms of descending percentiles:

3.9.1. $t_{(1)} \geq t_{(2)} \ldots \geq t_{(m)}$

3.10. Decision Logic:

After sorting, the candidate with the most important rule that provides the most evidence for correlation is at the top of the list. At this point, a decision is rendered according to the following hierarchical logic (the first applicable decision is accepted):

1. If the top score $t_{(1)}$ is negative, then the new observation is considered incompatible with all candidates and consequently considered the first report initiating a new track.

2. If the top score, $t_{(1)}$, is zero, then the available rule set is indeterminate. In this case, the *Preponderance of Evidence* score (discussed earlier) is examined, per options the user has selected. If the POE scores of all candidates with truth value zero are negative, the report is declared a new track – provided the user has enable this feature. If a POE score among those candidates with truth-value 0 is positive (and significantly greater than any other positive POE score[18]), that candidate is updated (per user option). Otherwise, the report is declared an ambiguity requiring operator resolution.[19]

[18] This feature is somewhat in a state of flux. The intention is to assess "significance" by assuming that under a null hypothesis of no differences, the POE score is asymptotically normal. The corresponding percentiles would then be uniformly distributed and the corresponding order statistics would then follow a beta. Unfortunately, the system does not currently retain global statistics on the POE distributions.

[19] Apart from the difficulties noted, if the *Preponderance of Evidence* score is positive, that is evidence that the

In all of the following cases, the t value of the highest ranked candidate is assumed positive.

3. If there is exactly one candidate, that candidate is updated.[20]

4. If the value of *t* is negative for the second ranked candidate ($t_{(2)}$), then the top-ranked candidate ($t_{(1)}$) is updated.

5. If the *Preponderance of Evidence* score is negative for the second ranked ($t_{(2)}$) candidate, then the top-ranked candidate ($t_{(1)}$) is updated.

6. If the difference in the percentile rank of the top ranked candidate and the second ranked candidate (($t_{(1)}$) - ($t_{(2)}$)) is positive, the top ranked candidate ($t_{(1)}$) is updated.

7. In all other cases, the report is ambiguous and merits operator attention.

3.11. Other Considerations:

Because successful COMINT correlation was the objective of this endeavor, the approach described above was applied at two levels: node (track) correlation and link correlation. As would be expected, the system applies two different rules sets to each of these activities. Although the essential aspects of the correlative evaluations are the same, the comparative features are not.

In the first case (node correlation), the correlator is trying to put "like with like". In link correlation, the correlator attempts to detect a communications "handshake", i.e. a case where the observation represents a station communication with someone who has already talked to them. When this occurs, both a link and a net correlation have occurred. Because a "handshake" is being sought in link correlation, some comparisons are made between the dual components of elements that are naturally paired. Once a handshake is established, not only are both a link and a net correlation achieved, but also information is shared and updated between the correlated links (if that option is allowed by the user).

Because of the structuring of the data, there is little information contained in the net entity that does not also reside in some other structure. For this reason, correlations to nets are per fixed methods and do not admit varying rules. Naturally, should the need arise, this restriction can be relaxed.

To improve the effectiveness of correlation, automated merging and ambiguity re-processing are optionally available. In automated ambiguity re-processing, the list of ambiguities is periodically re-examined to determine if sufficient additional information has accumulated to allow a correlation decision to be effected. This re-processing is performed only for nodes because link ambiguities are not carried.[21]

report correlates to the indicated candidate. In testing, this processing option is usually not selected because it is assumed that if there was a valid reason to correlate, one of the user's rules would have so indicated and, by assumption in step 2, that is not the case.

[20] No candidate is retained which does not have a significant similarity.

[21] The reason for not carrying link ambiguities is that it avoids bifurcating obfuscation. The situation wherein an ambiguous link emanates from an ambiguous node is a case that, in its simplest terms, devolves to the information that "somebody is talking to somebody", hardly elucidating. In fact, in most cases, this is insufficient to assign a report to even a specific net. In terms of the prior examples illustrating rules specification, an example of this situation would be where the reported data is "somebody is watching channel 6"

Automated merging is used to reduce both node and net fragmentation. This process is invoked when ambiguities are about to be declared among apparently equivalent candidates. In this case, the surviving candidates are examined to see if they correlate to each other. If so, the entities are merged and the new report updates the merged result.

4. Results

The lack of validated test data in the development environment precludes any definitive claim of effectiveness for this or any other correlation process. For this reason, assessments are of necessity anecdotal. Testing has been performed with observations sets consisting of on the order of 5,000 observations. For observations arising from a simulator (a service approved simulator not of our creation), ambiguity rates are on the order of 2-3%. For real world traffic, ambiguity rates have varied from 7-16%.

To aid in assessment and compensate for the lack of interpreted data, the system spins off summary data that shows all update tracks and their constituent observations, all tracks, and all ambiguities. Each ambiguity is examined to determine if its declaration is plausible. Each updated track is examined to assess the degree of miscorrelation. Finally, the "all track" data is sorted on key identifying parameters to see if there are any missed correlations.

Miscorrelation appears quite low. Some track fragmentation occurs but appears justified in the environment.[22] Ambiguities are "real" in the sense that they either (a) have no significant data reported or (b) possess contradictory data in contrast to that information which has previously been reported.

The methodology described in this paper has performed well in limited testing to date. The application software is currently being integrated with effective visualization tools. Two installations have graciously agreed to evaluate and provide feedback concerning this system. Current plans call for the fielding of an "Alpha release" in June for evaluation. Feedback from these analyst evaluations will drive modifications and enhancements.

5. References

[1].Snedecor, G. W. and W. G. Cochran. *Statistical Methods (Sixth Edition)*. 1967. Iowa State University Press, Ames, Iowa. Page 414.

[2].Anderberg, M. R. *Cluster Analysis for Applications.* 1973. Academic Press, London. Page 123.

[3].Klir G. J. and B. Yuan. *Fuzzy Sets and Fuzzy Logic, Theory and Applications.* 1995. Prentice Hall, New Jersey. Page 25.

[22] An "ad-hoc" parser is in use and the source of some problems. The introduction of "Normalization", a process of standardizing reported information, has been of value but requires refinement.