A Combinatorial Analysis of Link Discovery

John Cheng
Global InfoTek, Inc.
1920 Association Dr., Suite 200
Reston, VA 20191 USA
jcheng@globalinfotek.com

Ted Senator
DARPA/IPTO
3701 N. Fairfax Dr.
Arlington, VA 22203 USA
tsenator@darpa.mil

Keywords: Link Analysis, Multi-INT/Fusion, Information Sharing and Collaboration, Knowledge Discovery

Abstract

Link discovery and analysis, or connecting the dots, is a key component of intelligence analysis. Little principled analysis of this process has occurred. This paper presents a quantitative analysis of this process using a model based on the metaphor of identifying and assembling pieces of jigsaw puzzles. Specifically, it evaluates the probability that a particular puzzle can be recognized and classified based on various parameters describing data volumes, number of analysts, number of pieces required for recognition, and fraction of interesting puzzles. Combinatorial techniques are used to provide a closed-form solution for both single-analyst and multi-analyst collaborative situations. Computational experiments that demonstrate the effects of different parameters and structures are described. The key result is that factors that affect the likelihood of related pieces' being presented to a single analyst – such as the collection of more data – dominate the solution probability.

1. Introduction: The Puzzle Model

Intelligence analysis is sometimes described metaphorically as putting together the pieces in a jigsaw puzzle to enable recognition of the picture. Link discovery and analysis is a formal name for this process of connecting the dots [1]. It consists of inferring interesting, relevant information from masses of relational data by considering the patterns of connections between individual data elements. The key difficulty is that no piece of information is significant in isolation; rather, it is the combination in context of many related pieces of data that provide indications of significance. Much data are ultimately irrelevant, but this can be determined only after they are connected together. However, it is impossible to consider all possible connections because of the combinatorial complexity. Hence, an iterative process of connecting elements of information and evaluating significance is needed. Analysts perform this process to the best of human ability, matching available data against implicit or explicit patterns, using the partial matches to guide the acquisition of additional data, and repeating until uncertainty is reduced and a conclusion can be obtained.

The likelihood of detecting patterns of interest is affected by many factors, including:

- Number of entities and relationships
- Graph-structure of relationships
- Data completeness and correctness
- Number, size, and structure of interesting (and uninteresting) patterns
- Similarity between interesting and non-interesting patterns

Perhaps, however, the most important factor is none of the above; rather, it is the number of analysts, the amount of information they can analyze, and the organization of the analytical processes.

To analyze the effects of these factors on the likelihood of detection we create an abstract model that captures the significant aspects of link discovery and analysis while obscuring other details. We imagine that every element of available data is an individual jigsaw puzzle piece with the picture obscured and that pieces from multiple puzzles arrive all mixed together. Recognition of a puzzle (i.e., determination of its significance, modeled as the emergence of the picture) depends on obtaining a minimum number of pieces of the puzzle. Puzzle pieces are assigned randomly and possibly repeatedly to analysts. The model is depicted in Figure 1. An analysis of this model can answer the following questions:

- What is the probability that a person can solve a puzzle of interest (i.e., obtain enough pieces to recognize a particular picture)?
- How does the solution probability depend on various parameters such as the number and workload of analysts, the number of puzzles and of pieces per puzzle, the number of pieces required to recognize a puzzle, and the number of interesting puzzles?
- If analysts collaborate in teams, how does the solution probability change?

More formally, we assume that during a specified unit of time, there are **N** puzzles of size **P**, for a total of **NP** pieces. Of these **N** puzzles, **I** are of interest, and **N-I** are

not. **S** pieces are examined independently by each of **A** analysts. Recognizing a puzzle requires a minimum of **M** pieces of that particular puzzle. (There are obvious constraints between these parameters required for a sensible and useful interpretation, e.g., **I<N**, **M<S**, etc.) Model parameters are summarized in Table 1.

In this puzzle model, each piece represents an element of linked data. The context-free nature of linked data is captured by the fact that there is no way of distinguishing the puzzle pieces *a priori* and that single puzzle pieces are not meaningful – only a "critical mass" of **M** related pieces enables the analyst to recognize the puzzle. The ratio **M/P** captures the difficulty of detecting a pattern; the inverse may be thought of as the "pattern quality.") The low signal-noise ratio – resulting from much captured data's arising from uninteresting activities – is represented by the model's characterization of interesting vs. non-interesting puzzles. The signal (i.e., data of interest) corresponds to the **IP** interesting pieces; the noise to the **NP-IP** pieces of the uninteresting puzzles. Analyst productivity is modeled by the fact that the analyst only sees **S** puzzle pieces.

Simplifying assumptions of the model are:

- Puzzle sizes are identical
- Information content of all pieces is identical – ignores data quality issues
- No additional structure besides pieces and puzzles
- Probability of assignment of a piece to an analyst is random and uniformly distributed
- Each piece is relevant to only a single puzzle
- No redundant pieces received by an individual analyst – we use sampling without replacement
- No "contradictory" pieces – corresponds to ignoring data inconsistencies
- Pieces are analyzed as a group; i.e., a batch process. Hence, issues such as data distribution over time and decay of memory are ignored.
- Explicit modeling of the time or effort required to recognize puzzles is ignored.

Despite these limitations the model yields useful insights.

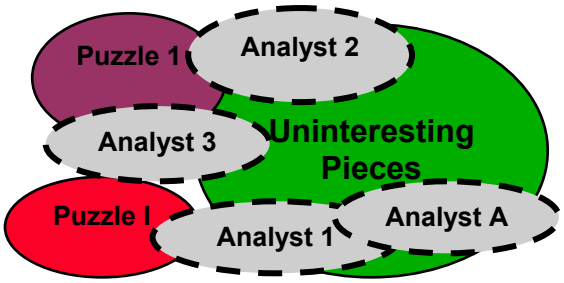


Figure 1

Table 1: Model Parameters

N:	number of puzzles
P:	pieces per puzzle
I:	number of interesting puzzles
A:	number of analysts
S:	pieces per analyst
M:	puzzle recognition threshold

2.0 Single-Analyst Link Discovery

This section develops the mathematics behind the model for a single analyst and presents and discusses results of various experiments.

2.1 Analysis

The model is analyzed using counting arguments. We first consider the simplest situation in which there is only one puzzle-solver (**A = 1**), and only one puzzle of interest (**I = 1**). What is the probability that the analyst finds a solution?

First, all possible ways in which **S** unique pieces can be sampled from the total number of pieces is:

$$\binom{NP}{S} \quad \text{where} \quad \binom{x}{y} = \frac{x!}{y!(x-y)!} \qquad \text{(Eq. 1)}$$

Using the fundamental counting principle [2], the number of distinct ways that **M** pieces from the single interesting puzzle can be chosen from all pieces is:

$$\binom{P}{M}\binom{NP-P}{S-M} \qquad \text{(Eq. 2)}$$

The first term is the number of ways **M** pieces from the interesting puzzle can be chosen from its set of **P** pieces. The second gives the number of ways the remainder of puzzle pieces – the non-interesting pieces – can be chosen, since the number of pieces in that set is **NP-P**, and the number of pieces picked from that set is **S-M**.

The puzzle can be solved, however, whenever *at least* **M** pieces from the puzzle of interest are drawn. The number of ways this can happen is

$$\binom{P}{M}\binom{NP-P}{S-M} + \binom{P}{M+1}\binom{NP-P}{S-M-1} + \ldots + \binom{P}{S} \qquad \text{(Eq. 3)}$$

assuming that **M < S < P**. The first term in the sum is identical to Equation 3. Each succeeding i^{th} term gives the number of ways that **M + *i*** pieces from the interesting puzzle can be chosen from all pieces.

The solution probability for **A = I = 1** is then:

$$\frac{\sum_{i=M}^{\min(S,P)} \binom{P}{i}\binom{NP-P}{S-i}}{\binom{NP}{S}} \qquad \text{(Eq. 4)}$$

Next we solve the more general case, allowing the puzzles of interest, **I,** to range inclusively from 1 to the total number of puzzles, **N**. We continue to assume, however, that the number of analysts is 1 (**A** = 1). To simplify the mathematics, we compute the complement of the solution probability, i.e., the probability that no puzzles can be solved, and then subtract this from 1 to get the actual solution probability.

The ways that the number pieces per analyst (**S**) can be chosen such that no puzzle of interest can be solved is:

$$\sum_{X_1=0}^{M-1} \cdots \sum_{X_I=0}^{M-1} \binom{P}{X_1} \cdots \binom{P}{X_I} \binom{NP - IP}{S - \sum_{j=1}^{I} X_j} \qquad \text{(Eq. 5)}$$

where $S \geq \sum_{j=1}^{I} X_j$

Here, each term in the sum gives the number of ways that no puzzles of interest can be solved for a unique combination of pieces per puzzle of interest – i.e., fewer than **M** pieces of any puzzle of interest are present. The **I** nested sums give all possible combinations for which this can be true, resulting in the total number of ways that no puzzle of interest can be solved.

The probability that the analyst cannot solve any puzzle of interest is then:

$$P_F = \frac{1}{\binom{NP}{S}} \bullet \sum_{X_1=0}^{M-1} \cdots \sum_{X_I=0}^{M-1} \binom{P}{X_1} \cdots \binom{P}{X_I} \binom{NP - IP}{S - \sum_{j=1}^{I} X_j}$$

$$where \quad S \geq \sum_{j=1}^{I} X_j \qquad \text{(Eq. 6)}$$

Hence, the solution probability – i.e. the probability that the analyst solves at least one puzzle of interest – is:

$$P_S = 1 - P_F \qquad \text{(Eq. 7)}$$

Although Equation 7 gives a closed-form solution, its form is such that the solution behavior is difficult to intuit. Hence, we next present a set of experiments to show the solution probability's characteristics.

2.2 Experiments and Discussion

In the following experiments, a range of parameter values is considered (Fig. 2). They were chosen to roughly match real-world intelligence data characteristics. Analysts can examine about 200 messages/day. That number is used as a basis for **S**, which ranges from 200 to 5000, reflecting the number of "puzzle pieces" an analyst might see in a day, a week, and a month. Message traffic volume is approximately 10,000 messages/day. Assuming that a reasonable number of puzzles is 20 for small problems (and increasing that by a factor of 5 and 25 for medium and large problems, continuing the day/week/month idea), the number of pieces per puzzle, **P**, must be fixed at 10,000/20 = 500. Finally, the puzzle recognition threshold is chosen at 25, which is 5% of the pieces of each puzzle, a fairly conservative estimate.

Problem Size ->	Small	Medium	Large
N (total puzzles)	20	100	500
P (pieces per puzzle)	500	500	500
S (pieces per analyst)	200	1000	5000
M (recognition threshold)	25	25	25

Fig. 2. Parameter ranges considered in experiments

2.2.1 Methodology

The following methodology was used:

- Java code was written to compute the probabilities. BigInteger/BigDecimal datatypes were used to ensure sufficient accuracy in the computation.
- Windows XP (1.4 GHz Intel Pentium 4 Mobile CPU, 256MB RAM) was used to execute the code.
- Wolfram Research's Mathematica 4.0 were used to graph the results. Some of the axes are labeled oddly – e.g., "**Number of Puzzles – 75**" indicates that the axis' scale range from 0 to 50 represents an actual **Number of Puzzles** from 75 to 125.

Only a subset of the experimental results is presented here due to space limitations.

2.2.2 Puzzle Solution Probability vs. Noise

Figure 3 depicts how the solution probability $\mathbf{P_S}$ varies as the number of puzzles **N** increases while the other parameters are held constant, essentially displaying $\mathbf{P_S}$ as noise increases. Note that in this graph, $\mathbf{P_S}$ is shown on a logarithmic scale.

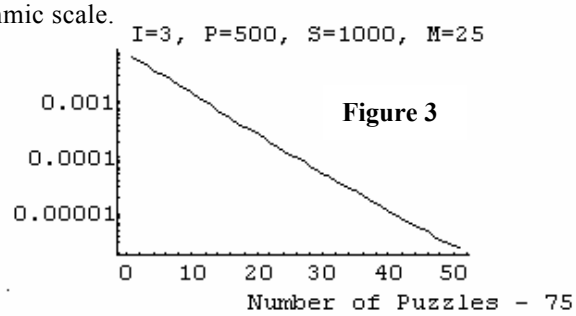


Figure 3

This experiment demonstrates how quickly $\mathbf{P_S}$ falls as noise increases. It suggests that collecting more data that do not contain the phenomena of interest will do more than simply obscure interesting patterns; it will break them apart into pieces too small to enable recognition.

2.2.3 Signal vs. Noise

In Figure 4, the total number of puzzles is varied along one dimension while the number of interesting puzzles is varied along another, showing how the solution probability $\mathbf{P_S}$ is affected by both these parameters.

The graph shows that increasing the number of interesting puzzles **I** tends to increase $\mathbf{P_S}$; however, this gain seems fairly slow. In the other dimension, we once again see how quickly $\mathbf{P_S}$ falls with increasing noise.

Since increasing signal and noise clearly have opposite effects on $\mathbf{P_S}$, it is interesting to consider their relative effects on $\mathbf{P_S}$. Due to the limited range of **I** in Figure 4, this tradeoff is difficult to see. Figure 5 addresses this issue by plotting $\mathbf{P_S}$ against the number of puzzles when the signal-to-noise ratio is constant.

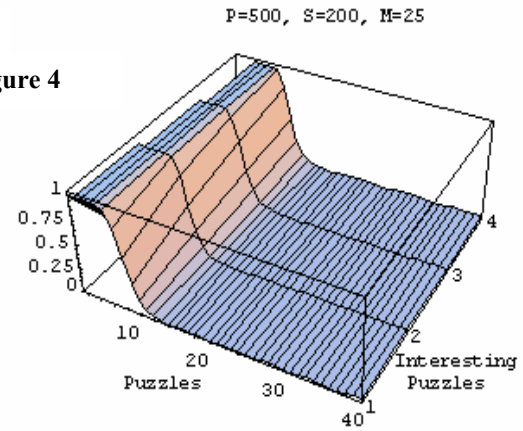


Figure 4

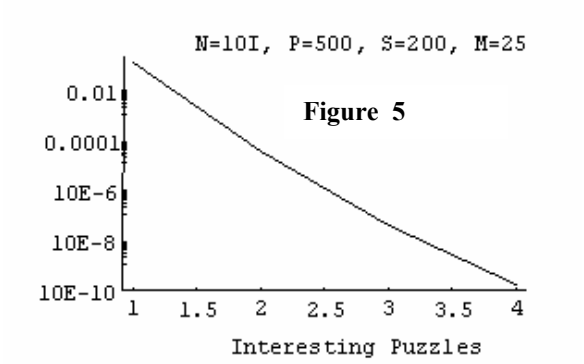


Figure 5

$\mathbf{P_S}$ is plotted on a logarithmic scale; the graphs show that even if **I** varies linearly with **N** (i.e., constant signal/noise), $\mathbf{P_S}$ falls extremely quickly. The prior set of experiments (Figures 3-4) showed that the solution probability falls exponentially whenever extraneous data is introduced, given a *fixed* set of interesting data. This experiment demonstrates the stronger result that this exponential decrease in link-discovery efficacy results as the data volume increases *even when the proportion of interesting data remains constant*.

2.2.4 Signal vs. Number of Pieces per Puzzle

To what extent does the number of pieces per puzzle **P** influence $\mathbf{P_S}$? If the recognition threshold **M** remains fixed, increasing **P** effectively reduces the proportion of the puzzle required to solve it, which should increase $\mathbf{P_S}$. Figure 6 graphs $\mathbf{P_S}$ vs. **P** and **I**.

We see that increasing **P** has an almost negligible effect, especially compared to **I**. This result suggests that increasing the amount of data available to analysts, *even if such an increase assumes that the number of interesting and non-interesting puzzles is fixed*, will not result in significant gains in link discovery performance. More data about the same phenomena does not help much.

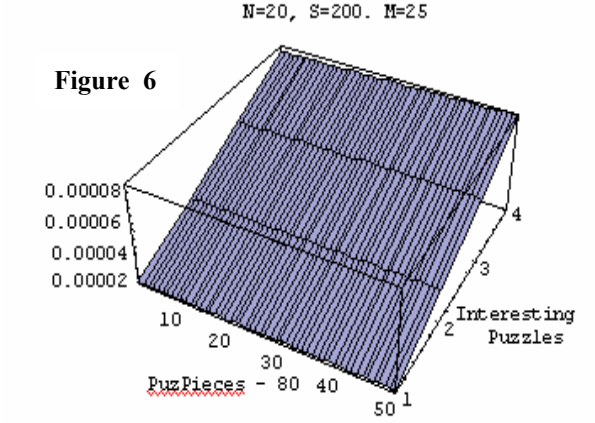
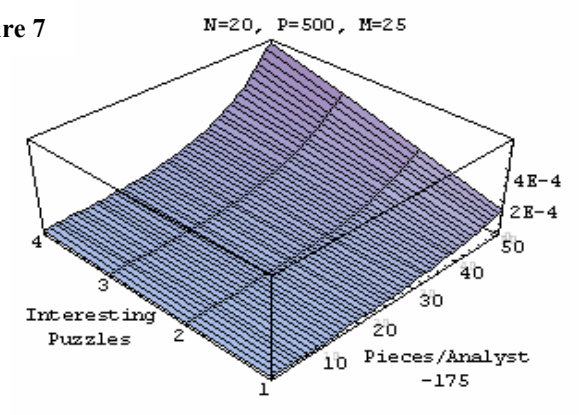


Figure 6

2.2.5 Signal vs. Pieces/Analyst

As we have seen, $\mathbf{P_S}$ increases with **I**. We would expect the number of pieces per analyst **S** to have a similar effect, since increasing **S** means the analyst is exploiting a larger proportion of the total amount of information, but how do **I** and **S** compare against each other? Figure 7 shows that **S** dominates, causing faster growth in $\mathbf{P_S}$. Increasing **I** while keeping the total number of puzzles constant increases the signal to noise ratio, and essentially amounts to decreasing the amount of irrelevant information available to the analyst – i.e., an increase in data relevance. This experiment shows that in order to increase link discovery effectiveness, increasing the amount of data an analyst can process is more important than increasing data relevance.

Figure 7

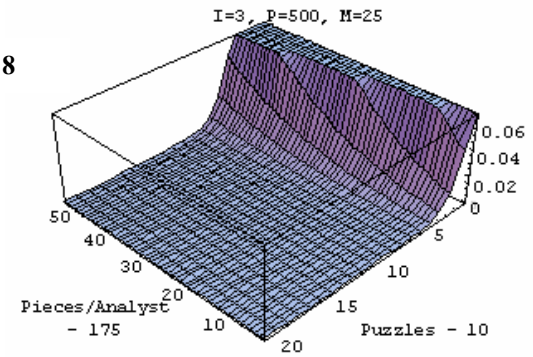


2.2.6 Noise vs. Pieces/Analyst

Since **S** dominates signal, we now consider the question of **S** compared to noise. Figure 8 shows that although **S** has a considerable affect on $\mathbf{P_S}$, $\mathbf{P_S}$ is much more sensitive to noise. That is, the negative effects of extraneous data on link discovery will always overwhelm any gains that result from improved analyst data cognizance.

!

Figure 8



2.2.7 Data Volume vs. Pieces/Analyst

From previous experiments we know that increasing data volume, even with a constant signal-to-noise ratio, decreases $\mathbf{P_S}$. What if the pieces/analyst is increased? Does that offset the poor data-scaling performance of $\mathbf{P_S}$? Figure 9 shows that $\mathbf{P_S}$ is much more sensitive to data volume than **S**. As in earlier experiments, even when the proportion of relevant and irrelevant data remain fixed, any increase in the amount of available data has a powerful effect on $\mathbf{P_S}$ – significantly more powerful than that of increasing **S**.

2.2.8 Recognition Threshold vs. Pieces/Analyst

One would expect $\mathbf{P_S}$ to grow rapidly as the recognition threshold M decreases. How does this decrease compare

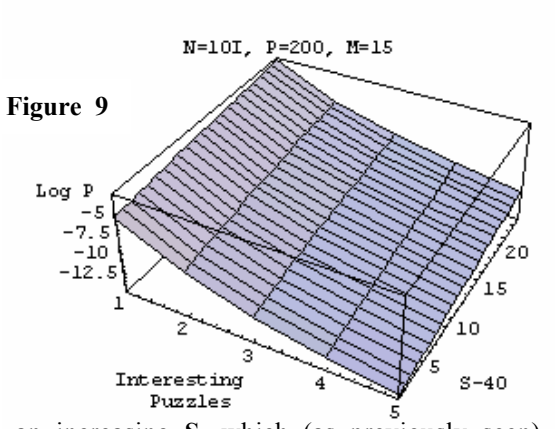


Figure 9

with an increasing **S**, which (as previously seen) also causes $\mathbf{P_S}$ to grow?

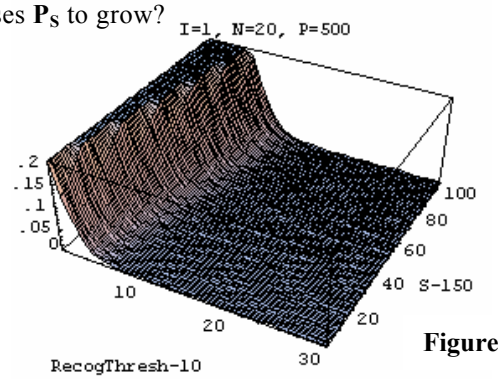


Figure 10

Figure 10 shows that **M** is significantly more important than **S**. This result suggests that efforts to reduce **M** are more valuable than efforts to increase **S** (or efforts to reduce noise, as previously shown). Reducing **M** is somewhat analogous to increasing **S** and the pieces per puzzle, **P**. Hence the dominance of **M** over **S** seems intuitively clear.

Taken together the experiments show that the dominant effect is the critical need to ensure that enough pieces sufficient for recognition will reach an individual analyst. The breaking apart of a puzzle into groups too small for an individual analyst to recognize is what makes link discovery such a difficult analytical task. Conversely, technologies that can reassemble the puzzles, or at least group pieces likely to have come from the same puzzle in a way that enables them to be assigned to the same analyst, should be most valuable.

3.0 Multiple-Analyst Link Discovery

Link discovery is difficult for analysts working in isolation because they cannot process enough data. Hence, we now consider the case where multiple analysts attempt to solve the puzzle problem by working in teams. An analysis of independent analysts is conducted, followed by that of collaborating analysts.

3.1 Multiple Independent Analysts

The probability that all **A** analysts fail, assuming that they operate independently, is:

$$\prod_{i=1}^{A}(1-P_i)$$

where $\mathbf{P_i}$ equals $\mathbf{P_S}$ for the *i*th analyst. Figure 11 shows how $\mathbf{P_S}$ scales with the number of analysts, assuming $\mathbf{P_i}$ for each analyst is 1%. The function is strictly monotonically increasing, which is a desirable characteristic. Adding manpower always results in higher levels of performance, albeit with diminishing returns.

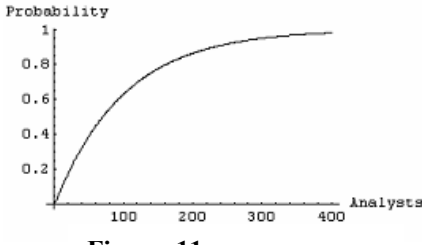


Figure 11

However, it is infeasible to use multiple independent analysts as an LD strategy. Consider the following:

- 10 total puzzles (N = 10)
- I = N/10 (1 puzzle of interest)
- 1000 pieces per puzzle (P = 1000)
- 200 pieces per analyst (S = 200)
- Recognition threshold = P/20 (M = 50)

This represents a relatively "easy" problem; puzzles of interest comprise a high percentage of the total (10%), the recognition threshold is only 5%, and analysts can see 2% of the total amount of data. These parameters give $\mathbf{P_S = 5.53x10^{-10}}$. Employing even 200 million analysts gives only a 10% probability of finding the puzzle of interest!

3.2 Multiple Collaborating Analysts

Modeling collaboration is difficult. There are many variables to consider, such as group structure, the productivity and collaboration efficiency of individuals, and even possibly group political issues. In this paper we model group collaboration at a high level of abstraction, allowing us to encapsulate overall group behavior with a relatively small set of parameters. This top-down approach simplifies the collaborative analysis considerably, while still giving insight into the collaborative process.

3.2.1 Analysis

Assume that **A** analysts collaborate, giving a single "virtual" analyst **V** with the following characteristics:

- **V** can examine $\delta\mathbf{AS}$ pieces, where $\mathbf{1/A < \delta <= 1}$
- **V** needs at least $\mathbf{M\varepsilon}$ puzzle pieces to solve it, where $\mathbf{1 <= \varepsilon < A}$

δ and ε are parameters that capture collaboration characteristics, modeling both group efficiency and information transfer issues. For instance, if the **A** analysts work together perfectly, with no information loss overhead from collaboration, the "work" that **V** can accomplish is **A** times greater than that of a single analyst; i.e., **V** can analyze **AS** puzzle pieces, and has a puzzle recognition threshold of **M**. If, on the other hand, the group is extremely unorganized and inefficient, the **A** analysts may be only slightly more productive than a single analyst.

The use of the "virtual analyst" allows us to reuse the previous analysis by simply replacing **S** by **δAS**, and **M** by **Mε** in Equations 6-7.

We now continue the analysis by determining more precise expressions for **δ** and **ε**. To do so, we introduce two additional collaboration parameters:

- **C**: represents collaborative efficiency or information transfer gain; i.e., if a person can analyze a fixed amount of data in **T** time units, he can express that analysis to others in his collaborative group in only **T/C** time units, where **C >= 1**.
- **ϕ**: represents information transfer efficiency in the collaborative process, where **0 < ϕ <= 1**. This parameter captures the observation that people don't remember everything that they are told. If, during a briefing, an analyst is given **B** bytes of information, he will recall only **ϕB** of it.

We now generously assume that **ε = 1** (no collaboration overhead for the recognition threshold). Also assume that the team is fully connected (analysts communicate with all other analysts) and that communication consists of *broadcasts*, which maximize communication efficiency (vice analyst-to-analyst communication, for instance). Both assumptions are conservative. If an analyst can perform a total of **S** units of work, and he spends **G** units of work on non-collaborative analysis, collaboration requires that he spends:

- **G/C** work units broadcasting information
- **G(A-1)/C** work units receiving information

Assume the analysts are 100% efficient:

S=G+A(G/C)= G(C+A)/C (Eq. 8)

and: **G = CS/(C+A)** (Eq. 9)

The total amount of information available to the group (puzzle pieces per virtual analyst, **δAS**) is essentially identical to the information available to individual analysts since the data is shared, and can be expressed as:

δAS = G + Gϕ(A-1) (Eq. 10)

The first factor, **G**, represents the amount of "personal" analysis completed by the analyst. The second factor represents the amount of work completed by the others in the group that the analyst can actually remember.

Combining these equations gives Equation 11:

$$\delta AS = \frac{CS\,(1+\phi\,(A-1))}{C+A},\ \varepsilon = 1$$

In the following experiments, we assume that

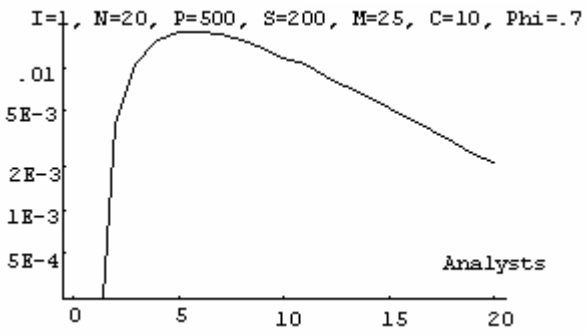
$$\phi = \frac{\phi_0}{\sqrt{A-1}} \quad \text{(Eq. 12)}$$

where ϕ_0 is a constant. The primary relevant characteristic of this function is that it decreases with respect to **A**, as does its rate of change. Although this assumption is arbitrary, it is chosen to be conservative. Hence, as **A** increases, the marginal efficiency of **ϕ** increases.

3.2.2 Experiments and Discussion

Figure 12 shows how $\mathbf{P_S}$ varies with the number of collaborating analysts. $\mathbf{P_S}$ is shown on a logarithmic scale. Our collaboration model reveals an identical characteristic shape under a variety of conditions: there is an optimal collaborative group size. As groups grow too large, the collaborative costs (even under the optimistic conditions modeled in this paper) overwhelm the collaboration gain. Collaborative effectiveness increases sharply as that optimal point is approached, and also decreases sharply after that point is passed. Optimal group sizes vary based on puzzle and group characteristics. For instance, increasing collaboration efficiency results in a larger optimal group size as well as higher values for $\mathbf{P_S}$. Decreasing data transfer efficiency does not appear to have a major effect on the optimal group size, but does reduce $\mathbf{P_S}$ significantly.

Figure 12



4.0 Conclusions

This paper has presented an abstract model of the link discovery problem, and has conducted closed-form probabilistic solution analyses for both single-person and collaborative situations. It shows that connecting the dots is an extremely difficult problem under the best of circumstances and that the difficulty arises primarily from the inability to identify related dots prior to assignment to individual analysts. Even if analysts collaborate effectively and efficiently, there is a natural limit to the ability to put together disparate clues into a coherent picture. It suggests that technologies that enable larger amounts of data per analyst can have a high payoff.

5.0 Acknowledgements

The first author's work on this paper was sponsored by the Air Force Research Laboratory, under Contract F30602-01-C-0202. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Air Force Research Laboratory, the Defense Advanced Research Projects Agency or the U.S. Government.

5.0 References

[1] Jensen, D. and Goldberg, H. Artificial Intelligence and Link Analysis: Papers from the 1998 AAAI Fall Symposium, AAAI Press, Menlo Park, CA 1998

[2] W. E. Deskins. "Abstract Algebra," pp. 45-50, 1996, Dover Publications.