1. Introduction

The problem treated in this paper is described as follows. We are to analyze a massive quantity of information which is naturally partitioned into discrete pieces. We have *a priori* only a nebulous description of *interesting information*, in particular, generic characterizations of how interesting information might differ from *mundane information*. By expending large sums of money and/or time we can pursue a seemingly interesting information piece and determine if it is truly interesting; we benefit greatly if indeed the information pursued was truly interesting. Since we can afford to pursue only a tiny fraction of the information with which we are presented, our goal is to prioritize (i.e., rank) the information pieces, allowing us to pursue the most appropriate collection of α pieces, for any α.

The information prioritization problem has a wide range of applications, including:

1. *Computer security*: The information is a computer system's audit trail and our task is to prioritize the recorded activities in order to facilitate the exploration of those pieces indicative of malicious intent (e.g., an intruder or a virus).

2. *Astronomy*: The information is radio signals collected from deep space and our task is to prioritize bursts of signals in order to facilitate the exploration of those pieces indicative of an intelligent life form.

3. *Biomedical testing*: The information is batteries of medical test results and our task is to prioritize the subjects in order to facilitate the exploration of those pieces indicative of the early stages of serious disease.

Our problem differs in several fundamental ways from traditional classification [1,2,3,5,6,9,10,13,16,17,18]. Firstly, we cannot obtain an optimal ranking from an optimal classification, since if two information pieces optimaly are placed in the same category, we have no guidance regarding their optimal relative prioritization. A surprising result is that, in the context of the problem version of most interest to us, classification and prioritization are *not necessarily consistent*. In particular, it may be optimal to rank x as less likely to be interesting than y, yet optimal also to classify y as mundane and x as interesting. Hence, it need not be the case that an optimal ranking can be turned into an optimal classification with appropriate choice of cutoff point.

Another key difference from the more traditional problem is that we are willing to consider contexts in which our training data is exclusively from the mundane process. Our approach assumes neither that an exact characterization nor a sample of the interesting process is available, and relies on various statistical abstractions of what is likely to distinguish mundane and interesting.

Two classes of **phenomena detectors** are considered here, and a third class of detectors is under study. A Class 1 detector must assign a priority value to each information piece as it is encountered, and this value may not depend on the information pieces that have been evaluated previously. A Class 2 detector waits until it has seen all the information pieces and then ranks the entire collection of information pieces; each of its evaluations may depend on the identity of all the information pieces to be evaluated. A Class 3 detector is intermediate; the value it assigns to information piece x may depend on the information pieces evaluated prior to encountering x.

Class 1 defines the most restrive class of detectors, but many applications require them. The motivations for a Class 1, rather than Class 2 or Class 3 detectors, include:

1. We need to provide a real-time evaluation of each information piece x as x enters the system and act upon this evaluation immediately. For example, if x and y are the first information pieces encountered and x is evaluated as being more likely to be interesting than y, we may begin exploring x while awaiting the arrival and evaluation of additional information pieces. We continue to explore x until a higher ranked information piece z arrives; we may resume exploration of x after all higher ranked information pieces have been explored fully. In this way, even if x later falls out of the top α ranked information pieces, the dovetailing exploration strategy will not delay the time to completion of the exploration of the top ranked α pieces, since no time is expended on an information piece so long as an already encountered, higher ranked information piece is not explored completely.

2. When the amount of information is massive (e.g., as in radio astronomy), we may be able to store only a small fraction of the information we collect. If we know in advance that we will be able to investigate at most α information pieces, a Class 1 detector requires only that we maintain a priority queue of the top α information pieces (along with statistics extracted from the training data), allowing us to discard any information piece that does not rank among the current top α. In contrast, both Class 2 and Class 3 detectors require that all information pieces encountered, and statistics associated with the information pieces (which can be substantially more information than that extracted from the training data) be kept until the entire collection of information pieces has been seen. For example, when using a Class 3 detector, it is possible that at some point of the processing information piece x ranks below the top α information pieces encountered thus far, but later in the processing it joins the top α information pieces as the statistics associated with the information pieces change. Hence, no information piece, nor statistics associated with the information pieces, can be discarded by either a Class 2 or Class 3 detector until all pieces have been seen.
3. Practically speaking, a Class 1 detector operates faster on-line (i.e., once constructed and put into operation) than do the other two classes of detectors, since it's evaluation of an information piece depends on a static body of statistics, statistics computed only in the course of the off-line training phase and not updated in the course of on-line evaluation. On-line efficiency often is critical for real-time applications.
4. In some applications, it is desirable not to allow information encountered in the course of monitoring to influence the evaluation. In computer security applications, for example, a classic scenario sees the sophisticated intruder defeating a detection system by manipulating the usage patterns. Under this scenario, the intruder gradually introduces into the system abnormal behavior, in a sense priming the system to consider as mundane an action that would otherwise be treated as interesting. Both Class 2 and Class 3 detectors are prone to this type of manipulation.

The current paper studies properties of Class 1 and Class 2 detectors, leaving Class 3 for future work. Our results demonstrate many theoretical differences between Class 1 and Class 2 detectors, including: (A) Optimal Class 2 prioritization is completely consistent with optimal Class 2 classification (that is, when the classifier may inspect all the information before classifying any of the information pieces), while optimal Class 1 prioritization is not always consistent with optimal Class 1 classification (that is, when the classifier must classify each information piece as it is encountered); and (B) Given a polynomial time oracle for evaluating the posterior probability of phenomena as defined in Section 2.4, Optimal Class 2 prioritization is solvable in time polynomial in the problem size (as measured in the number of information pieces encountered and the length of the encoding of the auxiliary distribution information to be specified later), while Class 1 prioritization is not, unless P=NP. As is discussed in Sections 3 and 4, these results are of both theoretical and practical interest. The final section the paper specifies efficient Class 1 detectors that are optimal for restricted problem instances, including symmetric **anomaly detection**.

2. Mathematical Formulation

In [12] we present a mathematical formulation of the detection problem in the context of computer security, in particular, misuse detection. Reference [12] defines a penalty function, gives optimal performance bounds, specifies the form of an optimal detector under perfect information (processes known exactly), and considers methods for heuristic detection, including data reduction (attribute selection and value aggregation). The formulation developed in the current paper extends the results to the more interesting and complex problem in which processes are not known, but described by prior distributions. The results of the current work subsumes [12] in that the earlier work is an instance of the current when exactly one normal-phenomenon process pair has nonzero density, *a priori*. The theory developed here also gives new mathematical justification to methods based on attribute selection, including those methods described in [2,4,11,12,13].

2.1 Basic Assumptions and Notation

The information to be analyzed is represented as a collection of **transactions**. A **transaction template** defines the structure of the transactions, that is, specifies the set of attributes to which each transaction assigns a value. For example, in a computer security application, a transaction template might include the attributes *user_id, command, port, time,* and *elapsed_cpu*. A computer's audit trail then collects **instances** of this transaction template, such as

user_id=Fred, command=edit file, port=xterm20, time=1994.06.07.08.01.0002, elapsed_cpu=2.2.

We assume the collection of transactions is homogeneous, each being an instance of a single, fixed and predetermined, transaction template. We assume also that the domain of each attribute (i.e., the set of values an attribute can assume) is finite, and hence the **transaction space** (i.e., the set of possible transaction instances)

$$S = \{x_1, x_2, \ldots, x_i, \ldots, x_{|S|}\}$$

is finite.

We model the transactions as generated by two, independent, stationary stochastic processes, the **normal** (mundane) process N and the **phenomena** (interesting) process M, each with its image in S. This simple, binary partitioning simplifies the presentation, but our results easily generalize to applications for which we wish to refine the partition.

A problem instance consists of a finite **training sample** *DB* and a finite **rank stream** *RS*, the actual transactions which must be ranked. Each of these transaction sequences is written as $< y_1, \ldots, y_j, \ldots, y_m >$, where $y_j \in S$ is the j^{th} transaction in the sequence, and the y_j are not necessarily distinct. *RS* is assumed to be generated by a mix of N and M, with the proportion determined by $\lambda = Pr\{\mathrm{N}\}$, the *a priori* probability of process N. The stochastic process T that selects between N and M for the generation of each transaction in *RS* is itself stationary and independent of each of N and M.

In the applications under study, *DB* is assumed generated by N only (though there are important applications in which the training data is of the same mix as *RS*). Hence, we lack any "tagged" occurrences from process M. That we lack tagged representatives of the process M is one of the characteristics that distinguishes our problem from the traditional classification problem; this work considers methods for coping with this severe problem constraint.

The formulation developed below is simplified by the assumption that each of *DB* and *RS* is of a fixed size, denoted $|DB| = n$ and $|RS| = k$ (where multiple occurrences of a transaction are counted multiple times toward the count). This assumption can be relaxed, replacing the fixed sizes by distributions over the possible sizes of DB and RS; all results apply in this case, provided that the resulting quantities are convergent.

The processes N and M are characterized by their probability distributions on S, that is, by the vectors

$\theta_N = <\theta_N(x_1), \cdots, \theta_N(x_i), \cdots, \theta_N(x_{|S|})>$ and $\theta_M = <\theta_M(x_1), \cdots, \theta_M(x_i), \cdots, \theta_M(x_{|S|})>$,

such that the coordinates of each of θ_N and θ_M are nonnegative, real valued, and sum to 1. The entries $\theta_N(x_i)$ and $\theta_M(x_i)$ represent the probabilities $Pr\{x_i|\theta_N\}$ and $Pr\{x_i|\theta_M\}$, respectively.

Let Θ_N and Θ_M denote the spaces of all normal and phenomena distributions, and let $\Theta = \Theta_N \times \Theta_M$ denote the set of all distribution pairs. For any $\theta = <\theta_N, \theta_M> \in \Theta$ and any $x_i \in S$:

$\theta_r(x_i)$ denotes the ratio $\theta_M(x_i)/\theta_N(x_i)$

$\theta_\lambda(x_i)$ denotes the linear combination $\lambda * \theta_N(x_i) + (1 - \lambda) * \theta_M(x_i)$.

Note that $\theta_\lambda(x_i)$ is the probability that, under θ, transaction x_i is generated as the j^{th} transaction, for any j, that is, $\theta_\lambda(x_i) = Pr\{x_i|\theta\}$.

2.2 Three Classes of Phenomena Detectors

The current paper considers two classes of phenomena detectors and demonstrates that the classes have rather different theoretical properties. A third class of detectors is introduced and is the focus of current research.

A phenomena detector pd from each of the classes trains on DB. A Class 1 detector, after training on DB, is instantiated to a map $pd: S \rightarrow \mathbf{R}$. Hence, as the j^{th} transaction x of an RS is generated, a pd of Class 1 provides an immediate evaluation $pd(x)$ that is independent of the RS that contains x. In contrast, a Class 2 detector does not provide an evaluation of x until it has examined both the training data DB and the entire rank stream RS that contains x. We shall demonstrate that Class 2 detectors are in a sense both potentially of greater power and "easier" to construct. Unfortunately, as was discussed in Section 1, the requirements of many applications mandate the use of a Class 1 detector. The Class 3 detector examines the rank stream up to the current point in time. Since it uses all of the currently available information, but does not yet have access to all the information, it is intermediary between the Class 1 and Class 2.

More formally, we have the following definitions. Let G denote the set $G = \{g | g$ is a (prior) density or probability function over $\Theta\}$.

A **Class 1 detector** is a function

$$pd\colon G \times \{samples\ DB \ni |DB| = n\} \times S \to \mathbf{R}.$$

Hence, such a pd's evaluation of each $x \in S$ does not depend on the RS which is encountered. Effectively, pd defines a single permutation $pd_{g,DB}(S) = \Pi(S)$ of S and pd's ranking of any RS is the restriction of $\Pi(S)$ to the distinct transactions appearing in RS. Practically, such a pd is instantiated to a map from $S \to \mathbf{R}$ once DB is presented during the (off-line) training phase.

A **Class 2 detector** is a function

$$pd\colon G \times \{samples\ DB \ni |DB| = n\} \times \{rank\ streams\ RS \ni |RS| = k\} \times S \to \mathbf{R}.$$

Hence, such a pd's evaluation of $x \in S$ may depend on the RS which is encountered, as well as on the training set DB. Effectively, pd defines a permutation $pd_{g,DB}(RS) = \Pi(RS)$ for each RS encountered† and these permutations may be independent, i.e., $\Pi(RS)$ and $\Pi(RS')$ need not be consistent. Practically, such a pd is not fully instantiated to a map from $S \to \mathbf{R}$ during the (off-line) training phase; analysis of RS must be performed in the course of on-line operation.

A **Class 3 detector** is a function

$$pd\colon G \times \{samples\ DB \ni |DB| = n\} \times \{rank\ streams\ RS_m \ni |RS_m| = m, m = 1, \ldots, k\} \times S \to \mathbf{R}.$$

Here, such a pd's evaluation of $x \in S$ may depend on the members of RS encountered up to the point x is encountered, as well as on the training set DB. This pd is also not fully instantiated to a map from $S \to \mathbf{R}$ during the (off-line) training phase; analysis of RS_m must also be performed in the course of on-line operation. Class 3 detectors are sequential estimators or empirical Bayes

† pd's prioritization of transactions in S but not RS is inconsequential and hence $pd(g, DB, RS, x)$ may be defined arbitrarily for any $x \in S$ not appearing in RS.

estimators; investigation of their properties is the subject of ongoing research. New methods for approximating such estimators have been reported in the literature [8,15], and we hope to apply these methods in our setting.

Observe that the second class of detectors properly contains the first -- a Class 1 detector is a Class 2 detector that is constant in the argument *RS*. Also note that Class 3 detectors are, in a sense, a bridge between Class 1 and Class 2 detectors, contained in Class 2, but containing Class 1.

Our prior distribution g on Θ typically will be fixed, in which case we write $pd(DB, x)$ (or $pd(DB, RS, x)$ for a Class 2 detector). This notation emphasizes the perspective that specialized algorithms are used to construct pd depending on the characteristics of g appropriate for a particular application.

2.3 Bayesian Risk Formulation

Loss Function

The following notational conventions are used.

1. pd is shown to include *RS* as an argument; if pd is from Class 1, it is understood that pd is constant in this argument.

2. The prior distribution on Θ is understood to be g.

3. The summation $\sum_{\{x,y\} \subset RS}$, where *RS* is a rank stream (i.e., a sequence of not necessarily distinct transactions), is taken over all unordered pairs of distinct transactions from the sequence *RS*, such that if x appears multiple times in *RS*, it will appear in multiple terms of the summation paired with the same y. In particular, if x appears $c(x)$ times and y appears $c(y)$ times, the pair $\{x, y\}$ appears $c(x) * c(y)$ times in the summation.

Loss is based on prioritization error as defined in [12]. Loss based on ranking is also discussed in [14], and that modeling appears to be quite consistent with ours. For any phenomena detector pd, training set

DB, and rank stream RS, $L(pd, DB, RS; \theta)$ is the prioritization error incurred by pd over all transactions pairs from RS (see (3) above) when the sample is DB and $\theta \in \Theta$ is the true distribution vector characterizing the processes N and M.

Formally, we define loss as follows. For any pair $\{x, y\}$ of transactions, define
$l(< x, y > ; \theta) = Pr\{M|x; \theta\} * Pr\{N|y; \theta\}$.

The loss $L_{x,y}$ incurred by pd on transaction pair $\{x, y\} \subset RS$ when the training set is DB, rank stream is RS, and the processes are characterized by θ is:

$$L_{x,y}(pd, DB, RS; \theta) = \begin{array}{l} l(< x, y > ; \theta) \text{ if } pd(DB, RS, x) < pd(DB, RS, y) \\ 1/2(l(< x, y > ; \theta) + l(< y, x > ; \theta) \text{ if } pd(DB, RS, x) = pd(DB, RS, y) \\ l(< y, x > ; \theta) \text{ if } pd(DB, RS, x) > pd(DB, RS, y) \end{array}$$

Loss function L is defined as:

$$L(pd, DB, RS; \theta) = [1/(|RS| * (|RS| - 1))] * \sum_{\{x,y\} \subset RS} L_{x,y}(pd, DB, RS; \theta) \qquad (2.3.1)$$

Discussion:

1. $L_{x,y}(pd, DB, RS; \theta)$ specifies that the loss incurred by a pd such that $pd(DB, RS, x) < pd(DB, RS, y)$ is the probability that, when θ characterizes the operative processes, x is generated by the phenomena process while y is generated by the normal process. This is an appropriate penalty, since it is exactly the probability that, when θ is operative, pd's ordering of x and y is a misdirection. (When $pd(x) = pd(y)$, the penalty is the average of the two directional penalties, reflecting the fact that in this case pd provides no guidance and x and y must be ranked by arbitrary means [e.g., a coin flip].) Note also that loss does not include terms $L_{x,x}$, since no detector prioritizes a pair of identical transaction instances.

2. A loss function appropriate for classification is $pen_I * Pr\{\mathrm{M}|x; \theta\}$ when x is classified as normal and $pen_{II} * Pr\{\mathrm{N}|x; \theta\}$ when x is classified as phenomena. (Here, pen_I and pen_{II} are chosen to model the relative importance of Type I and Type II errors.) While classification error is not applicable to our ranking problem, we shall demonstrate that the ranking induced by an optimal Class 2 detector is

trivially converted, for any pen_I and pen_{II}, into an optimal classifier, where the classification of x may be based on g, DB, and RS. In contrast, we demonstrate that the ranking induced by an optimal Class 1 detector is not consistent with optimal classification, where the classification of x may be based on g and DB, but not RS.

3. Loss, equation 2.3.1, is normalized by $|RS| * (|RS| - 1)$, which is proportional to the maximum number of terms contributed by a rank stream of size $|RS|$, thus allowing meaningful comparisons of loss incurred on RS of different sizes.

From the definition of Bayes risk we have

$$rsk(pd) = \underset{\Theta}{integral} R_{pd}(\theta) g(\theta) d\theta, \text{ where}$$

$$R_{pd}(\theta) = \sum_{DB,RS} L(pd, DB, RS; \theta) * Pr\{DB, RS|\theta\}$$

$$= \sum_{DB}(\sum_{RS}[1/(|RS| * (|RS| - 1))] * (\sum_{\{x,y\} \subset RS} L_{x,y}(pd, DB, RS; \theta)) * Pr\{RS|\theta\}) * Pr\{DB|\theta_N\}$$

$$= \sum_{DB}[1/k * (k - 1)] * (\sum_{RS}(\sum_{\{x,y\} \subset RS} L_{x,y}(pd, DB, RS; \theta)) * Pr\{RS|\theta\}) * Pr\{DB|\theta_N\}. \tag{2.3.2}$$

(Recall that k is the size of the rank stream to be analyzed.)

Technical Note: $R_{pd}(\theta)$ is bounded and continuous; hence, we are justified in the use of the Riemann integral whenever the prior distribution on Θ has integrable density $g()$. However, in many applications (and many of our examples) $g()$ will be a discrete probability function, in which case the integral is interpreted as a sum. In fact, important families of prior distributions can be enlarged to include more general probability measures over Θ, in which case the Riemann integral is replaced by the Lebesgue integral.

The expression for Bayes risk involves the distributions $Pr\{DB|\theta_N\}$, the probability that θ_N generate sample DB, and $Pr\{RS|\theta\}$, the probability that $\theta = \langle\theta_N, \theta_M\rangle$ generates rank stream RS.

Theorem 2.1:

(a)Let DB be a sample (i.e., a sequence of transactions), with $Count(DB)$ the vector $< c_1, \ldots, c_i, \ldots, c_s >$ giving the frequency count in DB of each $x_i \in S$, including those with 0 count. Then the probability $Pr\{DB|\theta_N\}$ that process N characterized by θ_N generates this DB is given by $Pr\{DB|\theta_N\} = (\prod_{i=1}^{s}(\theta_N(x_i)^{c_i}))$.

(b)Let RS be a rank stream (i.e., a sequence of transactions), with $Count(RS)$ the vector $< c_1, \ldots, c_i, \ldots, c_s >$ giving the frequency count in RS of each $x_i \in S$, including those with 0 count. Then the probability $Pr\{RS|\theta\}$ that processes N and M characterized by θ generates this RS is given by $Pr\{RS|\theta\} = (\prod_{i=1}^{s}(\theta_\lambda(x_i)^{c_i}))$.

Proof: The transactions in each of DB and RS are generated by a sequence of i.i.d. random variables whose distributions are given by θ_N and θ_λ, respectively. □

2.4 Optimal Class 2 Detection

Since the sums are over a finite number of terms, we can write

$$rsk(pd) = [1/k * (k-1)] * \sum_{DB}(integral_{\Theta}(\sum_{RS}(\sum_{\{x,y\}\subset RS} L_{x,y}(pd, DB, RS; \theta) * Pr\{RS|\theta\})) * Pr\{DB|\theta_N\}g(\theta)d\theta). \quad (2.4.1)$$

Observe that pd minimizes risk if and only if $pd(DB, \cdot, \cdot)$ is the minimizer for each $\overline{DB}$ of

$$rsk(pd, \overline{DB}) = [1/k * (k-1)] * integral_{\Theta}(\sum_{RS}(\sum_{\{x,y\}\subset RS} L_{x,y}(pd, \overline{DB}, RS; \theta) * Pr\{RS|\theta\})) * Pr\{\overline{DB}|\theta_N\}g(\theta)d\theta. \quad (2.4.2)$$

Similarly, any optimal Class 2 detector must be optimal for each $(\overline{DB}, \overline{RS})$ pair encountered. That is, an optimal pd of Class 2 must rank $\overline{RS}$ so as to minimize

$$rsk(pd, \overline{DB}, \overline{RS}) = integral_{\Theta}(\sum_{\{x,y\}\subset \overline{RS}} L_{x,y}(pd, \overline{DB}, \overline{RS}; \theta)) * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\}g(\theta)d\theta. \quad (2.4.3)$$

We therefore have the following necessary and sufficient condition for the optimality of a Class 2 detector.

Definition 2.1: Function f_2: $T \rightarrow \mathbf{R}$ orders set T **consistently** with function f_1: $T \rightarrow \mathbf{R}$ if, for all $x, y \in T$,

$f_1(x) < f_1(y) \Rightarrow f_2(x) < f_2(y)$.

Observe that if f_2 orders consistently with f_1 then: (a) When $f_1(x) = f_1(y)$, $f_2(x)$ and $f_2(y)$ may stand in any relation, and (b) We may have $f_2(x) = f_2(y)$ only when $f_1(x) = f_1(y)$.

Theorem 2.2: *pd* is an optimal Class 2 detector if and only if, for each $(\overline{DB}, \overline{RS})$, *pd* orders $\overline{RS}$ consistently with $pd^*(\overline{DB}, \overline{RS}, x) = \underset{\Theta}{integral} Pr\{M|x;\theta\} * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\} g(\theta) d\theta$.

Proof: See Appendix

Observation 2.1: Given a fixed pair $(\overline{DB}, \overline{RS})$, the probability that transaction x is generated by the phenomena process is $Pr\{M|x\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{RS} \wedge \overline{DB}) d\theta$, where $f(\theta|\overline{RS} \wedge \overline{DB})$ is the posterior distribution on Θ, given training sample $\overline{DB}$ and rank stream $\overline{RS}$.

Corollary 2.2: For each $(\overline{DB}, \overline{RS})$, $pd^*(x)$ is equal to $Pr\{\overline{DB} \wedge \overline{RS}\} * Pr\{M|x\}$, and hence ranking by the probability $Pr\{\mathrm{M}|x\}$ each $x \in S$ was generated by M (given $\overline{DB}$ and $\overline{RS}$) is optimal.

Proof: $pd^*(x) = \underset{\Theta}{integral} Pr\{M|x;\theta\} * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\} g(\theta) d\theta$

$= \underset{\Theta}{integral} Pr\{M|x;\theta\} * Pr\{\overline{RS} \wedge \overline{DB} \wedge \theta\} d\theta = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{RS} \wedge \overline{DB}) * Pr\{\overline{DB} \wedge \overline{RS}\} d\theta$

$= Pr\{\overline{DB} \wedge \overline{RS}\} * Pr\{M|x\}$ □

Discussion:

1. The ranking problem and the classification problem (classify each x into either N or M) are consistent when all of *RS* may be inspected before ranking or classifying any of its transactions. That is, the penalty for Type I or Type II error simply determines where in the optimal ranking to place the threshold on $Pr\{M|x\}$ for classifying x and, consequently, [x optimaly classified as N and y optimaly classified as M] $\Rightarrow$ [$pd^*(x) \leq pd^*(y)$].
2. The optimal Class 2 ranking produced by pd^* depends on $\lambda = Pr\{N\}$. (It is easy to see that $Pr\{\overline{RS}|\theta\}$ depends on λ and this quantity can be made to dominate.) Section 4 establishes that, in contrast, the optimal Class 1 ranking is independent of λ.
3. While the optimal Class 2 detector is specified by the simple expression

$Pr\{M|x\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{RS} \wedge \overline{DB}) d\theta$, the evaluation of this expression for a given prior distribution g on Θ may be infeasible. Nevertheless, the result stated in Corollary 2.2 is of some theoretical interest as it serves to highlight the differences between optimal Class 1 and Class 2 detection. It will be shown in Section 4 that the ability to evaluate the posterior $Pr\{M|x\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{DB}) d\theta$ of M conditioned on DB but not RS does not translate into the ability to perform optimal Class 1 detection. Specifically, it will be shown that (i) optimal Class 1 detection is not consistent with the ordering induced by these posterior probabilities (and hence not consistent with Class 1 classification) and, more generally, that even given a constant time procedure (i.e., an oracle) for computing these probabilities, no polynomial time evaluable Class 1 detector exists, unless $P = NP$. These results serve to delineate the computational complexities of the two classes of prioritization problems from each other, the complexities of Class 1 prioritization from classification, as well as to indicate the limits of the utility of approximations of $Pr\{M|x\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{DB}) d\theta$ as a heuristic for constructing optimal Class 1 rankings.

2.5 Class 1 vs. Class 2 Detectors

The distinction between the two classes of detectors is significant; for certain prior distributions on Θ, all optimal detectors lie in (Class 2 - Class 1). The following example demonstrates that the optimal detector indeed can depend on RS. In particular, the example demonstrates that, for a fixed DB, the optimal detector's ordering of rank stream $RS_1 =< x_1, x_2, x_3 >$ is not consistent with its ordering of rank stream $RS_2 =< x_1, x_1, x_2 >$. This demonstrates that the overall optimal detector comes from (Class 2 - Class 1), since the overall optimal detector must be optimal on each rank stream, while any detector from Class 1 induces a single ranking of S, with which the ranking of any rank stream is consistent.

Example 2.1: Suppose $S = \{x_1, x_2, x_3\}$ and that the distribution g on Θ is nonzero on only the following θ_1 and θ_2.

	θ_1	
	θ_{N_1}	θ_{M_1}
x_1	0.999	0.001
x_2	0.001-ε	0.999-ε
x_3	ε	ε

	θ_2	
	θ_{N_2}	θ_{M_2}
x_1	0.001	0.499
x_2	0.500	0.001
x_3	0.499	0.500

Suppose $\lambda = 0.5$.

The optimal detector from either class must be optimal for each individual DB. Consider $\overline{DB}$ and g such that $Pr\{\overline{DB}|\theta_1\} * g(\theta_1) = Pr\{\overline{DB}|\theta_2\} * g(\theta_2)$; hence, we can suppress this product in our analysis. Consider first $RS_1 = <x_1, x_2, x_3>$. The optimal Class 2 pd^* must be such that $pd^*(\overline{DB}, RS_1, x_2) < pd^*(\overline{DB}, RS_1, x_3) < pd^*(\overline{DB}, RS_1, x_1)$, since any optimal pd must minimize

$$\sum_{\theta_1,\theta_2} (\sum_{\{x,y\}\subset RS_1} L_{x,y}(pd, \overline{DB}, RS_1; \theta)) * Pr\{RS_1|\theta\}$$

and, for sufficiently small $\varepsilon \geq 0$, $Pr\{RS_1|\theta_1\}$ is so small that the optimal ordering is determined exclusively from the θ_2 term.

Consider now $RS_2 = <x_1, x_1, x_2>$. The optimal pd^* must be such that $pd^*(\overline{DB}, RS_2, x_1) < pd^*(\overline{DB}, RS_2, x_2)$, since any optimal pd must minimize also

$$\sum_{\theta_1,\theta_2} (\sum_{\{x,y\}\subset RS_2} L_{x,y}(pd, \overline{DB}, RS_2; \theta)) * Pr\{RS_2|\theta\},$$

and the difference in loss incurred on the θ_1 term by the orderings $x_1 < x_2$ versus $x_2 < x_1$, can be seen to dominate this difference on the θ_2 term. (Observe also that $Pr\{RS_2|\theta_1\} > Pr\{RS_2|\theta_2\}$.)

The optimal member of Class 1 must select a single ordering of S, not depending on RS, and hence the overall optimal is in (Class 2 - Class 1). The results of the sections which follow (specifically, see Example 3.1) allow us to demonstrate that, in fact, the optimal ordering of S by a Class 1 detector is $x_1 < x_2 < x_3$. It is interesting to note that this ordering differs from the way the optimal Class 2 detector must order the rank stream $RS_1 = \langle x_1, x_2, x_3 \rangle$. Intuitively, a Class 2 detector, when presented with an RS may adjust its posterior on Θ; a Class 1 detector may not depend on RS and produces a single ordering of S which minimizes an "expectation" over the distribution of possible RS. □

Hence, we have the following.

Theorem 2.3: There exist prior distributions g on Θ such that all optimal detectors lie in (Class 2 - Class 1).

Proof: Example 2.1 provides one such distribution g.

3. Theoretical Study of Class 1 Detectors

Theorem 2.2 and its corollary establish that a detector is optimal within Class 2 if and only if it orders consistently with $pd^*(x) = Pr\{M|x\} = \underset{\Theta}{integral}\, Pr\{M|x;\theta\} * f(\theta|\overline{DB})d\theta$

For many applications, however, a Class 1 detector is required. In Section 4 we demonstrate that specifying an optimal Class 1 detector is considerably more difficult than is specifying an optimal Class 2 detector; in particular, we demonstrate that, even when the distribution g is nonzero on only finitely many $\theta \in \Theta$ (or results in an easily-evaluable risk integral), the problem of determining an optimal Class 1 ranking is NP-hard.

We begin now to consider the problem of finding the best (or good) detector within Class 1.

3.1 Risk for Class 1 Detectors

Let pd be any Class 1 detector. Then the expression

$$rsk_1(pd) = \underset{\theta}{integral}\Big(\sum_{DB}\Big(\sum_{\{x,y\}\subset S} \theta_\lambda(x) * \theta_\lambda(y) * L_{x,y}(pd, DB;\theta)\Big) * Pr\{DB|\theta_N\}\Big) * g(\theta)d\theta, \qquad (3.1.1)$$

in which *RS* does not appear, is well defined. (It is not, however, well defined for a detector in (Class 2 - Class 1), since such a detector depends on the missing *RS* argument.) We now establish that rsk_1 and our original expression for risk, equation 2.4.1, yield the same value for any detector in Class 1 (where we continue to suppress *RS* as an argument to *pd* since *pd* is in Class 1).

Theorem 3.1: Let *pd* be any member of Class 1. Then for each fixed $\theta \in \Theta$ and training sample *DB*, the inner term

$$[1/k(k-1)] * (\sum_{RS} (\sum_{\{x,y\} \subset RS} L_{x,y}(pd, DB; \theta)) * Pr\{RS|\theta\})$$

of $rsk(pd)$ is equivalent to

$$\sum_{\{x,y\} \subset S} \theta_\lambda(x) * \theta_\lambda(y) * L_{x,y}(pd, DB; \theta).$$

Proof: See Appendix

In the following, when studying Class 1 detectors, we shall work with the more convenient form rsk_1 of risk.

3.2 Properties of Class 1 Detectors

The risk expression $rsk_1(pd, \overline{DB})$ (equation 3.1.1) for any fixed $\overline{DB}$ and Class 1 detector is a sum of $\{x, y\}$-terms, each of which is a function of only how *pd* orders x, y when trained on $\overline{DB}$. In particular, if *pd* orders as $pd(\overline{DB}, x) < pd(\overline{DB}, y)$, the $\{x, y\}$-contribution to $rsk_1(pd, \overline{DB})$ is

$$\underset{\Theta}{integral} Pr\{M|x; \theta\} * (1 - Pr\{M|y; \theta\}) * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta) d\theta$$

$$= \underset{\Theta}{integral} Pr\{M|x; \theta\} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta) d\theta - \underset{\Theta}{integral} Pr\{M|x; \theta\} * Pr\{M|y; \theta\} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta) d\theta.$$

Similarly, if *pd* orders as $pd(\overline{DB}, x) > pd(\overline{DB}, y)$, the $\{x, y\}$-contribution to $rsk_1(pd, \overline{DB})$ is

$$\underset{\Theta}{integral} Pr\{M|y; \theta\} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta) d\theta - \underset{\Theta}{integral} Pr\{M|x; \theta\} * Pr\{M|y; \theta\} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta) d\theta.$$

(In case $pd(\overline{DB}, x) = pd(\overline{DB}, y)$, the contribution is the average of these two quantities.)

Dropping the common second term of each ordering's contribution, the effect of the ordering choice $pd(\overline{DB}, x) < pd(\overline{DB}, y)$ is a contribution of

$$\underset{\Theta}{integral} Pr\{M|x;\theta\} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta)d\theta \qquad (3.2.1)$$

as opposed to a contribution of

$$\underset{\Theta}{integral} Pr\{M|y,\theta\} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta)d\theta \qquad (3.2.2)$$

for the ordering $pd(\overline{DB}, y) < pd(\overline{DB}, x)$. (In case $pd(\overline{DB}, x) = pd(\overline{DB}, y)$, the contribution is the average of these two quantities.)

Applying Bayes Theorem to the posteriors $Pr\{M|z;\theta\}$, we write the consequences of the orders as

$$\underset{\Theta}{integral} \frac{\theta_M(x)(1-\lambda)}{\theta_M(x)(1-\lambda)+\theta_N(x)\lambda} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta)d\theta \qquad (3.2.3)$$

versus

$$\underset{\Theta}{integral} \frac{\theta_M(y)(1-\lambda)}{\theta_M(y)(1-\lambda)+\theta_N(y)\lambda} * \theta_\lambda(x) * \theta_\lambda(y) * Pr\{\overline{DB}|\theta\} * g(\theta)d\theta. \qquad (3.2.4)$$

Using the identity $\theta_\lambda(z) = \theta_N(z)\lambda + \theta_M(z)(1-\lambda)$ and suppressing common terms, we write the two consequences as

$$\lambda(1-\lambda) * [\underset{\Theta}{integral} \theta_M(x)\theta_N(y)Pr\{\overline{DB}|\theta\}g(\theta)d\theta] \qquad (3.2.5)$$

versus

$$\lambda(1-\lambda) * [\underset{\Theta}{integral} \theta_M(y)\theta_N(x)Pr\{\overline{DB}|\theta\}g(\theta)d\theta]. \qquad (3.2.6)$$

When $\theta_N(x) \neq 0$ and $\theta_N(y) \neq 0$, we also can write the consequences of the orderings in terms of the ratios $\theta_r(z) = \theta_M(z)/\theta_N(z)$:

$$\lambda(1-\lambda) * [\underset{\Theta}{integral} \theta_r(x) * (\theta_N(x)\theta_N(y)Pr\{\overline{DB}|\theta\}g(\theta)\ )d\theta] \qquad (3.2.7)$$

versus

$$\lambda(1-\lambda) * [\underset{\Theta}{integral} \theta_r(y) * (\theta_N(x)\theta_N(y)Pr\{\overline{DB}|\theta\}g(\theta)\ )d\theta]. \qquad (3.2.8)$$

Consequently, a Class 1 detector can be evaluated by summing, over all pairs x, y, pd's ranking of x, y's contribution to risk, cancelling the terms and factors common to all rankings. In particular, if pd ranks x, y as $pd(\overline{DB}, x) < pd(\overline{DB}, y)$ the contribution can be written as the constant factor $\lambda * (1-\lambda)$ times

$$\underset{\Theta}{integral}\, \theta_M(x)\theta_N(y) Pr\{\overline{DB}|\theta\} g(\theta) d\theta \qquad (3.2.9)$$

plus a term common to the reverse ordering's contribution, or, equivalently, when $\theta_N(x) \neq 0$, as $\lambda * (1-\lambda)$ times

$$\underset{\Theta}{integral}\, \theta_r(x) * (\theta_N(x)\theta_N(y) Pr\{\overline{DB}|\theta\} g(\theta)\ ) d\theta. \qquad (3.2.10)$$

plus the common term. (If $pd(\overline{DB}, x) = pd(\overline{DB}, y)$, the average of $\theta_M(x) * \theta_N(y)$ and $\theta_M(y) * \theta_N(x)$ replaces $\theta_M(x) * \theta_N(y)$ in the integrand of (3.2.9), and the average of $\theta_r(x)$ and $\theta_r(y)$ replaces $\theta_r(x)$ in the integrand of (3.2.10)).

A detector is optimal within Class 1 if and only if its ranking minimizes the ordering-sensitive contributions (3.2.9), summed over all $\{x, y\} \subset S$ pairs. Notice that the optimal ranking by a Class 1 detector does not depend on λ, a fact that is in contrast to the optimal ranking by a Class 2 detector (see Theorem 2.2).

In [12] we demonstrate that pd defined by $pd_r(x) = \theta_r(x)$ minimizes a penalty expression that, when a single θ is such that $g(\theta) = 1$, is equivalent to rsk_1. That rsk_1 is minimized by this pd_r when a single θ has mass 1 follows immediately from inspection of (the single nonzero term of) expression (3.2.9) or (3.2.10). It is apparent also that, when a single θ is such that $g(\theta) = 1$, pd_r is optimal not only within Class 1, but within Class 2, i.e., in this case, a pd that is permitted to depend on RS has no advantage, since in this case the posterior distribution on Θ obviously is the same as the prior distribution, namely $f(\theta|DB, RS) = g(\theta) = 1$.

When g is nonzero on more than one $\theta \in \Theta$, the minimization of risk within Class 1 can become significantly more difficult. Analysis of expressions (3.2.9) and (3.2.10) reveals important characteristics of the minimization. Observe, for example, that the optimal pairwise orderings -- that is, the collection of pairwise orderings that minimizes each $\{x, y\}$ pair's contribution to $rsk_1(pd, DB)$ -- need not induce a consistent order on S. For example, $x_1 < x_2$, $x_2 < x_3$, and $x_3 < x_1$ might each be the ordering minimizing the corresponding transaction pair's contribution to $rsk_1(pd, DB)$; hence, the optimal detector of Class 1 (which maps $pd\colon DB \times S \rightarrow \mathbf{R}$, and thus induces a single, consistent ordering on S) in such a case cannot be pairwise optimal.

Example 3.1: Consider again Example 2.1 of Section 2.5 used to demonstrate that the optimal member of Class 2 must rank $RS_1 = < x_1, x_2, x_3 >$ and $RS_2 = < x_1, x_1, x_2 >$ inconsistently. This same example demonstrates also that, when attention is restricted to Class 1 detectors, the pairwise orderings which minimize each $\{x, y\}$ pair's contribution to risk are inconsistent with each other. Specifically (assuming ε is sufficiently small), it is easy to demonstrate [e.g., in terms of (3.2.9) or (3.2.10)] that:

$x_1 < x_2$ contributes less to risk then does $x_2 < x_1$, since θ_1 dominates this pair's ranking differential

$x_2 < x_3$ contributes less to risk then does $x_3 < x_2$, since θ_2 dominates this pair's ranking differential

$x_3 < x_1$ contributes less to risk then does $x_1 < x_3$, since θ_2 dominates this pair's ranking differential

The optimal Class 1 detector hence cannot be pairwise optimal.

Further, expression (3.2.9) or (3.2.10) easily can be used to show that the optimal Class 1 detector orders S as $x_1 < x_2 < x_3$. To see this, note that the optimal Class 1 detector must order as $x_1 < x_2$, since the difference between the contribution of $x_1 < x_2$ and $x_2 < x_1$ under θ_1 dominates all other contributions. In determining where in the optimal ordering x_3 must be placed, we observe that only θ_2 is significant, since any term involving x_3 and θ_1 includes a factor of ε, and hence goes to 0. It is then easy to verify that, with $x_1 < x_2$ fixed, the best placement of x_3 is following x_2. Hence, we have established also the claim made at the conclusion of Example 2.1, that the optimal member of Class 1 must order S as $x_1 < x_2 < x_3$, while the optimal detector in Class 2 must order the rank stream $RS_1 = < x_1, x_2, x_3 >$ as $x_2 < x_3 < x_1$. In particular, we have demonstrated the suboptimality of $pd(\overline{DB}, x) = \underset{\Theta}{integral} Pr\{M | x; \theta\} * Pr\{RS = < x_1, \ldots, x_{|S|} > | \theta\} * Pr\{\overline{DB} | \theta_N\} g(\theta) d\theta$ with respect to the minimization of risk for a Class 1 detector, where $S = \{x_1, \ldots, x_{|S|}\}$. □

3.3 The Suboptimality of Posterior-Based Class 1 Detectors and the Divergence of Class 1 Ranking from Classification

Corollary 2.2 demonstrates that ranking by a Class 2 detector is optimal if and only if its ranking is consistent with $Pr\{M|x;\, DB, RS\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{RS}x\overline{DB})d\theta$. It therefore is natural to consider the Class 1 detector

$$pd_{Post}(DB, x) = Pr\{M|x; DB\} = \underset{\theta}{integral} Pr\{M|x;\theta\} f(\theta|x; DB)d\theta$$

that ranks based on the posterior of x being phenomena given DB (but not RS).

Example 2.1 can be used to demonstrate that pd_{Post} is not an optimal Class 1 detector. It can be verified that $Pr\{M|x_1;\overline{DB}\} < Pr\{M|x_3;\overline{DB}\} < P\{M|x_2;\overline{DB}\}$ [for the $\overline{DB}$, g, and λ specified in Example 2.1, and $\overline{DB}$ sufficiently large so as to make each x_i insignificant in the posterior $f(\theta|x_i;\overline{DB})$], yet, as observed previously, the Class 1 optimal order of S is $x_1 < x_2 < x_3$.

The result is of more than mere theoretical interest in that the decrease in risk incurred by the optimal ordering over the prior-based ordering is significant. In particular, the risk incurred by the optimal ordering $x_1 < x_2 < x_3$ is 0.0623751, while the risk incurred by the posterior-based ordering $x_1 < x_3 < x_2$ is 0.0935628, which is 1.5 times the former. It is interesting to note also that the average of the risks incurred by the six possible orderings is 0.109219, which is 1.75 times the risk incurred by the optimal ordering but only 1.17 times the risk incurred by the posterior-based ordering.

Therefore, even if we were able to determine exactly $Pr\{M|x; DB\}$, we could not necessarily construct an optimal, or near optimal, ranking for a Class 1 detector. As we shall demonstrate in the next section, optimal ranking by Class 1 detectors in fact is intractable, even given an efficient means of computing the probabilities $Pr\{M|x; DB\}$.

4. The Complexity of Constructing Optimal Detectors

4.1 Optimization Over Class 2

Theorem 2.2 and its corallary establish $pd^*(x) = Pr\{M|x\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{DB})d\theta$ as an optimal Class 2 detector. If there is a finite number of nonzero density θ (specified explicitly in a problem instance's input), or if the integral is evaluable in polynomial time, evaluating pd^* requires time bounded by a low order polynomial in the size of the problem instance's description, and hence the problem of optimal Class 2 detection is in P. (The size of the problem instance is measured in the number of information pieces encountered and the length of the representation of the distribution g over Θ.) Corollary 2.2 establishes that, under these same assumptions, optimal Class 2 classification is in P also, i.e., when the classifier may inspect all of RS before classifying any x. For comparison with the results derived in this section we restate these results as the following theorem.

Theorem 4.1: If g is nonzero on only finitely many $\theta \in \Theta$, or if the above integral is evaluable in time polynomial in the size of the problem instance, the problem of obtaining an optimal ranking [classification] by Class 2 detectors [classifiers] is in P.

4.2 Optimization Over Class 1

In contrast to the expression for an optimal Class 2 detector, we have for Class 1 not specified an optimal detector. We establish here that, even when a finite number of θ have nonzero density, an optimal Class 1 detector, evaluable in polynomial time, in fact cannot exist, unless $P = NP$.

Observe that the problems of finding an optimal Class 1 detector and of constructing an optimal ordering of S are of equivalent complexity: if an optimal detector exists which is evaluable in polynomial time it can be applied to each member of S to sort S; if an optimal ordering of S can be obtained in polynomial time, an optimal pd can be defined by $pd(x) = position(x)$. Our first step is to establish the NP-completeness of the following abstract problem, which we call **nontransitive sorting** (**NTS**).

NTS is defined as:

INSTANCE: Collection $S = \{x_1, \ldots, x_n\}$ of items, a list of values $F(x_i, x_j) \in Z$ for each $1 \le i \ne j \le n$, and an integer bound K.

QUESTION: Is there a permutation $\Pi(S)$ such that $Cost(\Pi(S)) = \sum_{(x_i, x_j) \ni x_i < x_j \text{ in } \Pi(S)} F(x_i, x_j) \le K$?

Theorem 4.2: NTS is NP-complete.

Proof: See Appendix

Corollary 4.2: NTS is NP-complete even when $F(\cdot,\cdot)$ is restricted to the five values in $\{0, 1, 3, B_2+3, B_1+3\}$, where B_1 and B_2 are defined in the proof of the previous theorem.

Proof: Modify the transformation to use $F'(a,b) = F(a,b)+3$ and a target minimization value of $3 * W - 4K - 3(m-K)$, where $W = Choose(6m, 2)$, which is the number of terms appearing in the cost of any Π. $\square$

We next demonstrate that **Optimal Risk 1 Ordering** (**OR1O**) defined as follows is NP-complete.

> **INSTANCE**: A set $S = \{x_1, \ldots, x_s\}$ of transactions, a set $\Theta = \{\theta_1, \ldots, \theta_z\}$ such that each θ_i assigns the rational pair ($\theta_{iN}(x_j), \theta_{iM}(x_j)$) of probabilities to each $x_j \in S$, a (rational) prior probability value $g(\theta_i)$ for each $\theta_i \in \Theta$, a sample $\overline{DB} =< count(x_1), \ldots, count(x_s) >$, and a rational bound K.
>
> **QUESTION**: Is there a permutation Π of $\{x_1, \ldots, x_s\}$ such that $rsk(\Pi; \Theta, g, \overline{DB}) \le K$, where $rsk(\Pi; \Theta, g, \overline{DB}) = \sum_{(x_j, x_k) \ni x_j < x_k \text{ in } \Pi} (\sum_{i=1}^{z} \theta_{iM}(x_j) \theta_{iN}(x_k) Pr\{\overline{DB} | \theta_i\} g(\theta_i))$.

Note that it follows from (3.2.9) of Section 3.2 that a Class 1 detector minimizes risk if and only if the ordering Π it induces minimizes $rsk(\Pi; \cdot,\cdot,\cdot)$ as defined in OR1O. (If pd minimizes risk and $pd(x) = pd(y)$ so too does pd' obtained by ordering x and y either way.)

Theorem 4.3: OR1O ordering is NP-complete.

Proof: See Appendix

The main results of this section are stated in the following theorem.

Theorem 4.4:

(A) OR1O is NP-complete even if there are exactly $|S|^2 - |S|$ nonzero density $\theta \in \Theta$, each giving nonzero probability to no more than 3 transactions.

(B) The optimization version is NP-hard, that is, finding the ordering of S that minimizes risk is NP-hard.

(C) No optimal, polynomial time evaluable Class 1 detector pd: $G \times \{DB\} \times S \to \mathbf{R}$ can exist unless P=NP.

Proof: (A) follows directly from the construction of Theorem 4.3. (B) follows trivially from a Turing reduction of the decision problem to the optimization problem (the cost of a permutation can be obtained in polynomial time). (C) follows since the existence of such a detector allows the NP-hard ordering problem to be solved in polynomial time. □

To complete the contrast with Class 2 methods (where ranking and classification are essentially equivalent), we state the following.

Theorem 4.5: If g is nonzero on only finitely many $\theta \in \Theta$, or if the integral for $Pr\{M|z; DB\}$ is evaluable in polynomial time, the problem of obtaining an optimal classification by Class 1 methods is in P.

Proof: As observed at the conclusion of Section 3, given penalties for Type I and Type II classification errors, the optimal classification of y is phenomena if and only if $pen_{II} * (1 - Pr\{M|y; DB\}) \leq pen_I * Pr\{M|y; DB\}$. The posterior can be computed in time polynomial in the size of the input under the conditions of the theorem. □

5. Minimizing Risk Under Simple Assumptions

In the remainder of the paper we focus on Class 1 detection methods (though, where noted, some of the general results apply as well to Class 2 detection). The main results of the current section are theorems which specify efficient optimal detectors for prior densities g on Θ that generalize **anomaly detection**, under either of the following assumptions: (i) g is symmetric, reflecting the absence of any prior semantic knowledge of the attributes; or (ii) A particular, known set of attributes completely characterizes the processes. We describe in a separate paper [11] heuristics for specifying good, though not necessarily optimal, detectors for a broader class of prior distributions which includes applications outside the realm of anomaly detection, and report on a suite of experimental results using these methods.

5.1 A Sufficient Condition for Class 1 Optimality

In this section we derive a condition sufficient for establishing the optimality of Class 1 detectors. Sections 5.2 and 5.3 apply the result to establish the optimality of a Class 1 detector for each of the aforementioned families of prior distributions on Θ.

It is convenient to extend the definition of $\theta_r(x)$ so that it is defined even when $\theta_N(x) = 0$.

Definition 5.1: For $\theta \in \Theta$ and $x \in S$, define $\theta_r(x) = \theta_M(x)/\theta_N(x)$ when $\theta_N(x) \neq 0$, $\theta_r(x) = 1$ when $\theta_N(x) = \theta_M(x) = 0$, and $\theta_r(x) = \mathit{infinity}$ when $\theta_N(x) = 0, \theta_M(x) \neq 0$. The *infinity* symbol behaves as: $v < \mathit{infinity}$ for all $v \in \mathbf{R}$.

Observation 5.1: Under this definition, $\theta_M(x) * \theta_N(y) < \theta_M(y) * \theta_N(x)$ if and only if $\theta_r(x) < \theta_r(y)$.

Definition 5.2: Let g be a fixed probability or density function† on Θ. For any fixed pair $\{x, y\}$ of transactions, we partition as follows the subset of Θ on whose members g is nonzero:

$\Theta_{x<y} = \{\theta \in \Theta | g(\theta) \neq 0 \text{ and } \theta_r(x) < \theta_r(y)\}$

$\Theta_{x>y} = \{\theta \in \Theta | g(\theta) \neq 0 \text{ and } \theta_r(x) > \theta_r(y)\}$

$\Theta_{x=y} = \{\theta \in \Theta | g(\theta) \neq 0 \text{ and } \theta_r(x) = \theta_r(y)\}$

Definition 5.3: Let $\hat{\Theta}$ be any subset of Θ, and x, y any pair of transactions. Φ is an x, y-reflection on $\hat{\Theta}$ if for each $\theta = \langle \theta_N, \theta_M \rangle \in \hat{\Theta}$, $\Phi(\theta) = \langle \theta'_N, \theta'_M \rangle$ is such that:

$$\theta_N(x) = \theta'_N(y),\ \theta_N(y) = \theta'_N(x),\ \theta_M(x) = \theta'_M(y), \text{ and } \theta_M(y) = \theta'_M(x).$$

We also apply Φ component-wise, that is, for $\theta = \langle \theta_N, \theta_M \rangle \in \hat{\Theta}$, we shall write $\Phi(\theta) = \langle \Phi(\theta_N), \Phi(\theta_M) \rangle$.

Note that there is no restriction on an x, y-reflection's effect on coordinates other than x and y.

We now prove a general result that is useful in establishing the form of an optimal detector for certain simple prior distributions g.

Theorem 5.1: Let g be a probability or density function over Θ and $\overline{DB}$ some training sample. Suppose

---

†As discussed in Section 2, our notation assumes g is a Riemann integrable density function or a discrete probability function; in the latter case, the integrals are to be interpreted as sums. The families of distributions considered here can be enlarged to include more general probability measures over Θ, in which case the Riemann integral is replaced by the Lebesgue integral.

that for each x, y transaction pair there exists a bijection (possibly depending on $\overline{DB}$)

$$\Phi_{x,y}: \Theta_{x<y} \rightarrow \Theta_{x>y},$$

which is an x, y-reflection on $\Theta_{x<y}$, such that either:

$$\forall \theta \in \Theta_{x<y},\ Pr\{\overline{DB}|\theta\} * g(\theta) \geq Pr\{\overline{DB}|\Phi_{x,y}(\theta)\} * g(\Phi_{x,y}(\theta)) \text{ or}$$

$$\forall \theta \in \Theta_{x<y},\ Pr\{\overline{DB}|\theta\} * g(\theta) \leq Pr\{\overline{DB}|\Phi_{x,y}(\theta)\} * g(\Phi_{x,y}(\theta)).$$

Then the collection of pairwise orders defined by:

$x<y$ when a bijection exists to satisfy the first condition, with the inequality strict for at least one $\theta \in \Theta_{x<y}$

$y<x$ when a bijection exists to satisfy the second condition, with the inequality strict for at least one $\theta \in \Theta_{x<y}$

$x = y$ when bijections exist to satisfy both conditions (and hence none of the inequalities is strict.)

is a collection of optimal pairwise orderings given $\overline{DB}$ (that is, each ordering in the collection minimizes each $\{x, y\}$'s contribution to rsk_1 given this fixed $\overline{DB}$).

Proof: See Appendix

Remark: As the following example demonstrates, the required bijections can exist and induce a collection of pairwise orders that is not consistent; in this case, the optimal Class 1 detector is not specified by this ordering rule (i.e., the optimal Class 1 detector is not pairwise optimal).

Example 5.1: First suppose that the following are the only four $\theta \in \Theta$ such that $g(\theta)>0$.

θ_1 is such that $\theta_{1_r}(x)<\theta_{1_r}(y)<\theta_{1_r}(z)$. Note $\theta_1 \in (\Theta_{x<y} \cap \Theta_{y<z} \cap \Theta_{x<z})$.

θ_2 is such that $\theta_{2_r}(y)<\theta_{2_r}(x)<\theta_{2_r}(z)$. Note $\theta_2 \in (\Theta_{y<x} \cap \Theta_{y<z} \cap \Theta_{x<z})$.

θ_3 is such that $\theta_{3_r}(z)<\theta_{3_r}(x) = \theta_{3_r}(y)$. Note $\theta_3 \in (\Theta_{z<y} \cap \Theta_{z<x})$.

θ_4 is such that $\theta_{4_r}(z)<\theta_{4_r}(x) = \theta_{4_r}(y)$. Note $\theta_4 \in (\Theta_{z<y} \cap \Theta_{z<x})$.

Suppose that $\overline{DB}$ and g are such that

$$Pr\{\overline{DB}|\theta_1\} * g(\theta_1) = Pr\{\overline{DB}|\theta_4\} * g(\theta_4) < Pr\{\overline{DB}|\theta_2\} * g(\theta_2) = Pr\{\overline{DB}|\theta_3\} * g(\theta_3).$$

Define bijections:

$\Phi_{x<y}: \Theta_{x<y} \rightarrow \Theta_{y<x}$ as $\Phi_{x<y}(\theta_1) = \theta_2$.

$\Phi_{y<z}: \Theta_{y<z} \rightarrow \Theta_{z<y}$ as $\Phi_{y<z}(\theta_1) = \theta_4$ and $\Phi_{y<z}(\theta_2) = \theta_3$.

$\Phi_{z<x}: \Theta_{z<x} \rightarrow \Theta_{x<z}$ as $\Phi_{z<x}(\theta_4) = \theta_1$ and $\Phi_{z<x}(\theta_3) = \theta_2$.

It is easily verified that each of these bijections obeys the inequalities of Theorem 5.1 and together prescribe the orders $x<y$, $y = z$, and $z = x$. These three orders together are inconsistent, but do leave open the possibility that a consistent (and optimal) *pd* could be defined by resolving the equalities in a consistent way (any orderings of y and z and of z and x are pairwise optimal). However, we can mirror the cyclic structure from above, using disjoint collections of θ to augment each of the six sets $\theta_{v<w}$ so that the resulting 3 bijections (which extend the 3 specified above) prescribe the orders $x<y$, $y<z$, and $z<x$. Hence, the pairwise optimal orderings in this case cannot be used to construct an optimal pd. □

A theorem parallel to Theorem 5.1 holds for Class 2 detectors, where $Pr\{\overline{RS} \wedge \overline{DB} | \theta\}$ replaces $Pr\{\overline{DB} | \theta\}$ in the inequality conditions. The result for Class 2 detectors is established simply by showing that the pairwise orderings prescribed by the bijections are identical to those induced by the optimal Class 2 detector pd^* of Theorem 2.2 (and hence are consistent). The result, however, is of less utility for Class 2 detectors than Class 1 detectors, since the optimality of pd^* gives us a more convenient and broadly applicable tool for establishing the optimality of a given Class 2 detector.

5.2 Symmetric Anomaly Detection

An important illustration of how Theorem 5.1 can be applied is anomaly detection, when no prior information on attribute meaning is available.

Definition 5.3: $\theta = \langle \theta_N, \theta_M \rangle \in \Theta$ is N-**ordered** if, for each x and y pair, $\theta_r(x) < \theta_r(y)$ implies $\theta_N(x) \geq \theta_N(y)$. We say that distribution g on Θ is consistent with **anomaly detection** if g has mass 1 on $\Theta_{AD} = \{\theta \in \Theta \mid \theta$ is N-ordered$\}$.

Discussion: Anomaly detection has been characterized [12] as detecting events rarest in N, without explicit reference to the process M. Perhaps the most intuitive abstraction of a (N, M) relationship consistent with anomaly detection is *common in N implies rare in M*, meaning $\theta_N(x) > \theta_N(y)$ implies $\theta_M(x) \leq \theta_M(y)$.

Observe that the common-rare condition implies N-ordered, since if we had $\theta_r(x) < \theta_r(y)$ yet also $\theta_N(x) < \theta_M(y)$, the common-rare condition leads to a contradiction, since it forces $\theta_M(x) \geq \theta_M(y)$ making impossible the premise $\theta_r(x) < \theta_r(y)$. On the other hand, we can have N-ordered without obeying this common-rare condition: We can have $\theta_r(x) < \theta_r(y)$, $\theta_N(x) > \theta_N(y)$, and $\theta_M(x) > \theta_M(y)$. Further, if θ is N-ordered, the most suspicious x (highest $\theta_r(x)$ value) are indeed the events rarest in N (lowest $\theta_N(x)$ value). Hence, θ_{AD} is a precise characterization of anomaly detection.

The following operator is used to characterize symmetric distributions.

Definition 5.4: For any θ (in either Θ_N or Θ_M) let $\theta_{x/y}$ denote the θ' obtained by interchanging the coordinates $\theta(x)$ and $\theta(y)$. Extend to $\theta \in \Theta$ this exchange operator by $\theta_{x/y} = \langle (\theta_N)_{x/y}, (\theta_M)_{x/y} \rangle$.

Observe that for any $\theta \in \Theta$ and any transactions x, y, $\theta_N(x) = (\theta_N)_{x/y}(y)$, $\theta_M(x) = (\theta_M)_{x/y}(y)$, and $\theta_r(x) = (\theta_{x/y})_r(y)$.

Theorem 5.2: If g is consistent with anomaly detection and has the additional property that, for each $\theta \in \Theta_{AD}$ and any x, y transaction pair, $g(\theta) = g(\theta_{x/y})$, then the frequentist detector defined by $pd(x) = 1/count(DB, x)$ minimizes risk among Class 1 detectors.

Proof: See Appendix

Discussion: The symmetry property of g required by the theorem is implied when we have no prior information on attribute meaning, and hence no transaction can be favored over another. In particular, the condition is equivalent to the condition that all permutations of any given probability vector have the same prior density, since a sequence of interchanges can map any permutation to any other permutation. Hence, while the condition allows different shaped distributions (e.g., all transactions equally likely vs. distributions that makes some transactions more likely than other) to be assigned different priors, it prohibits assigning larger prior density to a distribution that favors x over y than to an identically-shaped distribution that favors y over x.

A simple prior distribution that satisfies the conditions of Theorem 5.2 associates each θ_N only with the flat (uniform) θ_M (denoted θ_{M_U}), and further assigns density so that $g(\theta_N, \theta_{M_U}) = g(\theta'_N, \theta_{M_U})$ for every $\theta_N, \theta'_N \in \Theta_N$. This prior distribution may be considered *anomaly detection with a completely uninformed prior on* Θ_N.

5.3 Minimizing Class 1 Risk Under the Hypothesis of Specific Relevant Attributes

Theorem 5.2 may seem to go against experience, which shows that using a *pd* defined in terms of raw frequency counts leads to poor results. In particular, it generally is the case that a great majority of the transactions in the training data have count 1, and transactions encountered in the data to be analyzed were unseen in the training data, making useless the raw frequency counts. Experience indicates that selection of subsets of the attributes is required before an effective detector can be defined. Basing a detector on projection sets, however, appears contrary to the result of Theorem 5.2. The reconciliation of theory and practice is that we do not lack totally information regarding attribute behavior. Rather, we believe *a priori* that some attributes will be irrelevant, though we don't know which. Since the modeling of such prior information results in a prior distribution g which violates the symmetry required for the applicability of Theorem 5.2 (see Observation 5.4 which follows), the use of raw, unprojected frequency counts is not supported in such cases.

In this section, we give a flavor of our feature selection methods by considering the case where we know the set of relevant attributes. Heuristics for treating the more realistic problem in which the set(s) of relevant attributes is not known in advance are described in [11].

Definition 5.5: Let $A \subseteq U$ be a subset of attributes and x any transaction. The **projection** $x[A]$ of x onto A is the transaction obtained by restricting x to the attributes in A. We write $x[A] = z[A]$ if and only if x and z agree on each attribute in A and observe that projection onto any $A \subseteq U$ induces on S an equivalence class structure, that is, $z \in [x]_A$ iff $z[A] = x[A]$. We define also the following aggregate class quantities:

$$\theta_N(x[A]) = \sum_{z \in [x]_A} \theta_N(z)$$

$$\theta_M(x[A]) = \sum_{z \in [x]_A} \theta_M(z)$$

$\theta_r(x[A]) = \theta_M(x[A])/\theta_N(x[A])$

$count(DB, x[A]) = \sum_{z \in [x]_A} count(DB, z)$

Consider now the following family of prior distributions on Θ which, intuitively, states that only the subset A of attributes is relevant to determining the relative frequencies of transactions generated by N and M and, consequently, to prioritization.

Definition 5.6: The subset Θ_A of Θ is defined as $\Theta_A = \{\theta \mid$ whenever $x[A] = z[A]$, $\theta_N(x) = \theta_N(z)$ and $\theta_M(x) = \theta_M(z)\}$. We say that g is consistent with H_A (the hypothesis that only A is relevant to detection) if g has mass 1 on Θ_A.

Observation 5.2: A sufficient condition for the validity of H_A is:

In each of N and M, attributes in $(U - A)$ are independent of those in A and N and M is each uniform on $(U - A)$ in the sense that for all x and z, $\theta_N(x[U - A]) = \theta_N(z[U - A])$ and $\theta_M(x[U - A]) = \theta_M(z[U - A])$.

In this case, if $x[A] = z[A]$ then $\theta_N(x) = \theta_N(x[A]) * \theta_N(x[U - A]) = \theta_N(z[A]) * \theta_N(z[U - A]) = \theta_N(z)$ and similarly for θ_M.

Observation 5.3: Suppose g is such that the inequality $g(\theta) > g(\theta')$ holds whenever $\theta \in \Theta_A$ for *some* proper subset $A \subset U$ and $\theta' \notin \Theta_B$ for *every* proper subset $B \subset U$. Then g violates the symmetry required by Theorem 5.2 since, as the following lemma demonstrates, $\theta \in \Theta_A \Rightarrow \theta_{x/y} \notin \Theta_B$ for any proper subset $B \subset U$. (The result assumes each attribute has more than 2 distinct values and $\theta_N(x) \neq \theta_N(y)$ or $\theta_M(x) \neq \theta_M(y)$.)

Lemma 5.1: Let $A \subset U$ be a proper subset of the set U of all attributes. Let $\theta \in \Theta_A$ and let x and y be transactions such that $\theta_N(x) \neq \theta_N(y)$ or $\theta_M(x) \neq \theta_M(y)$. Then $\theta_{x/y} \notin \Theta_B$ for any proper subset $B \subset U$. (The result assumes each attribute has more than 2 distinct values.)

Proof: See Appendix

The following map is used in the statement of Theorem 5.3, playing the role of the bijection required by Theorem 5.1.

Definition 5.7: Define $Q_A(x, y; \theta)$ for any $\theta \in \Theta_A$ as interchanging the (common) θ_N and θ_M values of members of $x[A]$ with the (common) values of members of $y[A]$.

Clearly Q_A is well defined on Θ_A and maps Θ_A to itself. Observe also that $Q_U(x, y; \theta)$ is exactly the x/y interchanges operation.

Theorem 5.3: If g has mass 1 on $\Theta_A \cap \Theta_{AD}$, and if for each x, y pair $g(\theta) = g(Q_A(x, y; \theta))$ for all $\theta \in \Theta_A \cap \Theta_{AD}$, then the detector $pd_A(x) = 1/count(DB, x[A])$ minimizes risk among Class 1 detectors.

Proof: See Appendix

Discussion:

1. The theorem provides a risk-based justification of the common practice of attribute projection (known also as *attribute selection*).
2. The condition on g is the natural generalization of that for Theorem 5.2, when density is restricted to Θ_A; all permutations within Θ_A of any given probability vector have equal prior density, while a permutation outside of Θ_A has zero density.
3. If $g(\Theta_A) = 1$ then $g(\Theta_{A'}) = 1$ for all $A' \supset A$ (since then $\Theta_{A'} \supset \Theta_A$). However, in this case, if g satisfies the symmetry condition on Q_A it generally will not satisfy it on $Q_{A'}$, since for $\theta \in \Theta_{A'} - \Theta_A$, $g(\theta) = 0$, yet, for some such θ, $\theta' = Q_{A'}(x, y; \theta)$ is a member of Θ_A and $g(\theta') \neq 0$. Consequently, the theorem's premise is generally satisfied only by the minimal A such that $g(\Theta_A) = 1$.
4. A simple prior distribution that satisfies the condition of Theorem 5.3 associates each $\theta_N \in \Theta_A$ only with θ_{M_U}, and further assigns density so that $g(\theta_N, \theta_{M_U}) = g(\theta'_N, \theta_{M_U})$ for every $\theta_N, \theta'_N \in \Theta_{A_N}$.

6. Conclusions

We have defined the information prioritization problem and considered two classes of phenomena detectors. We have demonstrated that the two classes have rather different theoretical properties, including that optimal detection by one class is consistent with optimal classification, while such is not the case for the other class. The computational complexity of optimal detection diverges for the two classes as well. In particular, optimal detection by algorithms in one class is shown to be solvable in polynomial time (assuming a finite prior distribution or, more generally, a polynomial time oracle for computing the posterior $Pr\{M|x\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{DB})d\theta)$ while optimal detection by the other class (under the same assumptions) is NP-hard. We have demonstrated also that prioritizing by the posterior probability $Pr\{M|x\} = \underset{\Theta}{integral} Pr\{M|x;\theta\} * f(\theta|\overline{DB})d\theta$ can incur significantly greater error than the optimal Class 1 detector. This suggests that Class 1 prioritization is a problem fundamentally different from Class 1 classification (where classifying by this posterior in fact is optimal), and that techniques which prioritize by approximating this posterior have inherent limitations.

We presented detectors for solving optimally restricted versions of the detection problem, including symmetric anomaly detection. Other work [11] reports heuristic approachs to more general problem versions; results of an experimental software system appear to indicate that the approachs hold promise for many important problem domains. Recently introduced approximation techniques based on the probabilistic structure of the model may prove helpful in the calculation of the detectors.

ACKNOWLEDGEMENT: The authors wish to thank the referee for his comments which greatly improved the quality of this work. We wish to thank also Jessie Bhangoo for suggesting the phrase *phenomena detector*, and Henry Shapiro for his help in debugging a preliminary version of the reduction used in the proof of Theorem 4.2.

Appendix:

Theorem 2.2: pd is an optimal Class 2 detector if and only if, for each $(\overline{DB}, \overline{RS})$, pd orders $\overline{RS}$ consistently with $pd^*(\overline{DB}, \overline{RS}, x) = \underset{\Theta}{integral} Pr\{M|x;\theta\} * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\}g(\theta)d\theta$.

Proof: Inspection of $rsk(pd, \overline{DB}, \overline{RS})$ reveals that the contribution of the pair $\{x, y\}$ to $pd's$ risk is minimized when $pd(\overline{DB}, \overline{RS}, x) < pd(\overline{DB}, \overline{RS}, y)$ if and only if

$$\underset{\Theta}{integral} Pr\{M|x;\theta\} * (1 - Pr\{M|y;\theta\}) * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\}g(\theta)d\theta \le \underset{\Theta}{integral} Pr\{M|y;\theta\} * (1 - Pr\{M|x;\theta\}) * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|$$

or, equivalently, if and only if

$$integral_{\Theta} Pr\{M|x;\theta\} * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\} g(\theta)d\theta \leq integral_{\Theta} Pr\{M|y;\theta\} * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\} g(\theta)d\theta.$$

Similarly, the contribution of the pair $\{x, y\}$ to $pd's$ risk is minimized when $pd(\overline{DB}, \overline{RS}, x) = pd(\overline{DB}, \overline{RS}, y)$ if and only if

$$integral_{\Theta} Pr\{M|x;\theta\} * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\} g(\theta)d\theta = integral_{\Theta} Pr\{M|y;\theta\} * Pr\{\overline{RS}|\theta\} * Pr\{\overline{DB}|\theta_N\} g(\theta)d\theta.$$

$\square$

Theorem 3.1: Let pd be any member of Class 1. Then for each fixed $\theta \in \Theta$ and training sample DB, the inner term

$$[1/k(k-1)] * (\sum_{RS}(\sum_{\{x,y\} \subset RS} L_{x,y}(pd, DB; \theta)) * Pr\{RS|\theta\})$$

of $rsk(pd)$ is equivalent to

$$\sum_{\{x,y\} \subset S} \theta_\lambda(x) * \theta_\lambda(y) * L_{x,y}(pd, DB; \theta).$$

Proof: For any pair z, w of transactions in S, consider the total weight, over all rank streams RS of size k, attached to the term $L_{z,w}$. Consider first the weight contributed by a fixed $\overline{RS}$ of size k. The number of times the term $L_{z,w}$ occurs in the sum

$$\sum_{\{x,y\} \subset \overline{RS}} L_{x,y}(pd, DB; \theta)$$

is $n(z) * n(w)$, where $n(z)$ and $n(w)$ are, respectively, the number of occurrences of z and w in $\overline{RS}$. Hence, the contribution to risk of terms $L_{z,w}$ for such an $\overline{RS}$ is $n(z) * n(w) * (L_{z,w}(pd, DB; \theta) * Pr\{\overline{RS}|\theta\})$.

Consider now the total weight attached to $L_{z,w}$ by all RS with the same counts $n(z)$ and $n(w)$. To compute this quantity, we require the total probability of this collection of rank streams, that is, $Pr\{RS \ with \ counts \ n(z) \text{ and } n(w) \ |\theta\}$. Let V denote the set $S - \{z, w\}$ of remaining transactions. Then $n(V) = k - (n(z) + n(w))$ is the number of occurrences in RS of members of V and $\theta_\lambda(V) = 1 - (\theta_\lambda(z) + \theta_\lambda(w))$ is the probability that any single transaction generated by θ is other than z or w. Hence, $Pr\{RS \ with \ counts \ n(z) \text{ and } n(w) \ |\theta\}$

$$= \frac{k!}{n(z)! * n(w)! * n(V)!} * \theta_\lambda(z)^{n(z)} * \theta_\lambda(w)^{n(w)} * \theta_\lambda(V)^{n(V)}.$$

The total weight, over all RS of size k, attached to the term $L_{z,w}$ hence is

$$\sum_{(n(z),n(w),n(V))} [(n(z) * n(w)) * \frac{k!}{n(z)! * n(w)! * n(V)!} * \theta_\lambda(z)^{n(z)} * \theta_\lambda(w)^{n(w)} * \theta_\lambda(V)^{n(V)}],$$ where the sum is taken over all 3-tuples $(n(z), n(w), n(V))$ such that $n(z), n(w) \geq 1, n(V) \geq 0, n(z) + n(w) + n(V) = k$. (Note that $n(z), n(w) > 0$ since, otherwise, an RS with these counts does not contribute any z, w pairs.) This weight equivalently can be written as

$$k * (k-1) * \theta_\lambda(x) * \theta_\lambda(y) * (\sum_{(a,b,c)} [\frac{(k-2)!}{a! * b! * c!} * \theta_\lambda(x)^a * \theta_\lambda(y)^b * \theta_\lambda(z)^c]),$$ where the sum is taken over all 3-tuples (a, b, c) such that $a, b, c \geq 0, a + b + c = k - 2$. But this quantity is equal to

$k * (k-1) * \theta_\lambda(x) * \theta_\lambda(y) * (\theta_\lambda(x) + \theta_\lambda(y) + \theta_\lambda(z))^{(k-2)} = k * (k-1) * \theta_\lambda(x) * \theta_\lambda(y)$, since $(\theta_\lambda(x) + \theta_\lambda(y) + \theta_\lambda(z)) = 1$.

Since this is the total weight, over all RS of size k, attached to each term $L_{z,w}$ we have established that

$$[1/k(k-1)] * (\sum_{RS}(\sum_{\{x,y\}\subset RS} L_{x,y}(pd, DB; \theta)) * Pr\{RS|\theta\}) = \sum_{\{x,y\}\subset S} \theta_\lambda(x) * \theta_\lambda(y) * L_{x,y}(pd, DB; \theta). \quad \square$$

Theorem 4.2: NTS is NP-complete.

Proof: Clearly, NTS is in **NP**, as we can guess a permutation $\Pi(S)$ and in polynomial time compute its cost. We establish the NP-completeness of NTS by reducing Maximum 2-Satisfiability (Max 2-SAT) [6;LO5] to it.

Let the instance I of Max 2-SAT be:

$U = \{x_1, \ldots, x_n\}$

$C = \{c_1, \ldots, c_m\}$

K

Create as follows an instance I' of NTS:

S contains a total of $6m$ elements: Each clause $c_j = \{l_{ij}, l_{kj}\}$, where l_{ij} and l_{kj} denote complemented or uncomplemented occurrences of the distinct[†] variables x_i and x_k, generates four members of S, namely

[†]Any instance of Max 2-SAT containing a clause $\{x, x'\}$ trivially can be transformed to an equivalent instance without such a clause, since all truth assignments satisfy such a clause.

x_{ij}, x_{ij}', x_{kj}, and x_{kj}'. Each c_j additionally generates the two enforcer elements u_j and z_j in S. Intuitively, ordering as $x_{ij} < x_{ij}'$ in I' will correspond to setting x_i true in I, while ordering as $x_{ij}' < x_{ij}$ in I' will correspond to setting x_i false.

$F(\cdot, \cdot)$ is defined as follows: Unless explicitly assigned a value below, $F(a, b) = 0$.

We first specify F so as to enforce the requirement that, for each i, the pairs x_{ij}, x_{ij}' are ordered the same way across the j, and hence that the induced assignment to x_i is consistent. To this end, we first force the permutation to be consistent with the clause ordering, that is, we ensure that, for each $i = 1, 2, \ldots, n$, $x_{ij} < x_{ik}$ and $x_{ij}' < x_{ik}'$ for every pair $j < k$ such that $x_{ij}, x_{ik} \in S$. To accomplish this, set

$F(x_{ik}, x_{ij}) = B_1$, for each i and every pair $k > j$ such that $x_{ij}, x_{ik} \in S$

$F(x_{ik}', x_{ij}') = B_1$, for each i and every pair $k > j$ such that $x_{ij}', x_{ik}' \in S$,

where $B_1 = 7nm + 7m + n + 1$. The only way to avoid incurring a cost of B_1 is to order the variables by clause, as described above. As we shall see, B_1 is sufficiently large that if a single B_1 cost is incurred by an ordering, the total cost of this ordering is guaranteed to be positive, and hence greater than the bound we shall select for instance I'.

Next, we force, for any i, all the x_{ij} variables to come either before or after all the x_{ij}' variables. Let $f(i)$ and $l(i)$ respectively denote the subscripts of the first and last clause in which x_i or x_i' appears. Set

$F(x_{\Leftarrow}', x_{il(i)}) = B_2$, for each i

$F(x_{il(i)}', x_{\Leftarrow}) = -B_2$, for each i,

where $B_2 = 7m + 1$. There are exactly two ways to avoid a B_2 cost, assuming all B_1 costs are avoided:

If, for any i, all the x_{ij}' follow all the x_{ij}, the above function values are not activated and neither B_2 nor $-B_2$ costs are incurred.

If, for any i, all the x_{ij}' precede all the x_{ij}, each of the above 2 function values are activated and the costs cancel.

Notice that any interleaving necessarily will incur a B_2 cost (assuming B_1 is not incurred and hence the clause ordering holds) which won't be cancelled. As we shall see, if an ordering incurs even a single

B_2 cost that is not cancelled by a $-B_2$ cost, the cost of the ordering will be positive. Observe also that even if all n $-B_2$ costs are obtained, this would not compensate for incurring even a single B_1 cost.

Now define F to check satisfiability. Let clause $c_j = \{l_a, l_b\}$, where literals l_a and l_b are complemented or uncomplemented occurrences of the distinct variables x_a and x_b of U. l_{aj} denotes $x_{aj} \in S$ if l_a is uncomplemented and $x_{aj}' \in S$ if l_a is complemented; similarly for l_{bj}. Also, if l_{aj} represents x_{aj}, $\neg l_{aj}$ represents x_{aj}' and if l_{aj} represents x_{aj}', $\neg l_{aj}$ represents x_{aj}; similarly for l_{bj}.

We use F to force, for each clause $c_j = \{l_a, l_b\}$:

$u_j < l_{aj}$, by setting $F(l_{aj}, u_j) = B_1$

$u_j < l_{bj}$, by setting $F(l_{bj}, u_j) = B_1$

$\neg l_{aj} < z_j$, by setting $F(z_j, \neg l_{aj}) = B_1$

$\neg l_{bj} < z_j$, by setting $F(z_j, \neg l_{bj}) = B_1$

Hence, we force both

$u_j < l_{aj}, l_{bj}$ and $\neg l_{aj}, \neg l_{bj} < z_j$,

without specifying how the "in clause" variables l_{aj} and l_{bj} are ordered, nor how the "out clause" variables $\neg l_{aj}$ and $\neg l_{bj}$ are ordered, nor how the two chains interact. Note, for example, that any order of the form $u_j < \pi < z_j$ (π denotes any ordering of the four complemented and uncomplemented variables) does not incur any of these B_1 costs. However, there also are orderings not incurring a B_1 cost in which u_j follows z_j.

Clause c_j specifies also:

$F(l_{aj}, \neg l_{bj}) = -2$

$F(l_{bj}, \neg l_{aj}) = -2$

$F(z_j, u_j) = -3$

Observe that if an ordering is credited with -3 for $z_j < u_j$, the ordering avoids a B_1 cost only if it is of the form:

$\neg l_{aj}, \neg l_{bj} < z_j < u_j < l_{aj}, l_{bj}$.

Hence, in this case, the ordering is not credited with either of the -2 terms, and as we shall argue, the corresponding assignment will not satisfy c_j.

The transformation is completed by setting the the target minimization value in instance I' of NTS to $-4K - 3(m - K)$.

We now demonstrate that if instance I of Max 2-SAT has an assignment A satisfying J clauses then the constructed instance I' of NTS has a permutation achieving cost $-4J - 3(m - J)$; hence, I is a YES instance implies I' is a YES instance.

Consider any clause $c_{j=}\{l_a, l_b\}$. We demonstrate that if assignment A satisfies c_j, we can order the corresponding elements of S so that we obtain a value of -4 on these pairs, and that if A does not satisfy c_j, we can order the corresponding elements of S so that we obtain a value of -3 on these pairs.

Case 1: A satisfies c_j.

1a: Both l_a and l_b are assigned true. Then we order as

$u_j < l_{aj} < l_{bj} < \neg l_{bj} < \neg l_{aj} < z_j$.

We obtain 2 pairs of -2, hence achieving the -4.

1b: l_a is assigned true and l_b assigned false. Then we order as

$u_j < l_{aj} < \neg l_{bj} < l_{bj} < \neg l_{aj} < z_j$

As in Case 1a, we obtain 2 pairs of -2.

1c: l_a is assigned false and l_b assigned true. Then we order as

$u_j < l_{bj} < \neg l_{aj} < l_{aj} < \neg l_{bj} < z_j$

Again we obtain 2 pairs of -2.

Case 2: A does not satisfy c_j. Then we order as

$\neg l_{aj} < \neg l\_bj < z_j < u_j < l_{aj} < l_{bj}$

We obtain -3 on $F(z_j, u_j)$.

We now extend to a total order on S the orderings specified above, while incurring zero additional (cumulative) cost. Let $y_{1j} < y_{2j} < y_{3j} < y_{4j}$ denote the ordering of the "literals" $l_{aj}, l_{bj}, \neg l_{aj}, \neg l_{bj}$ specified by clause c_j. Then the orders already specified are consistent with the total order

$$y_{11} < y_{21} < y_{12} < y_{22} < \cdots < y_{1m} < y_{2m} < y_{31} < y_{41} < y_{32} < y_{42} < \cdots < y_{3m} < y_{4m}$$

of $U - \{u_j, z_j \mid j = 1, 2, \ldots, m\}$. (The variables u_j and z_j are placed into the above chain in a manner depending on the case for clause c_j).

To see that the new pairwise orders specified by this total order incur zero additional cost, observe first that the order specifies $x_{ij} < x_{ik}$ and $x_{ij}' < x_{ik}'$ for all i and $j < k$ such that $x_{ij}, x_{ik}, x_{ij}', x_{ik}' \in S$. Hence, the B_1 costs of $F(x_{ik}, x_{ij})$ and $F(x_{ik}', x_{ij}')$ for $j < k$ are avoided.

Next, note that all the x_{ij} precede all the x_{ij}' if A assigns x_i true, while all the x_{ij} follow all the x_{ij}' if A assigns x_i false, i.e., there is no interleaving. To see this, observe that the orders prescribed by each c_j places the two literals assigned true before the two literals assigned false, i.e., in all cases, the pattern is $T < T < F < F$. Hence, any B_2 cost incurred by $F(x_{\Leftarrow}', x_{il(i)})$ is paired with a $-B_2$ cost obtained by $F(x_{il(i)}', x_{\Leftarrow})$.

Finally, the remaining pairwise orderings are of the form $\{l_{ij}, l_{qk}\}$, $i \neq q$ and $j \neq k$ (the variables may be primed or unprimed, in any combination), and u_j, z_j variables paired with variables from clauses other than j. Since $F(\cdot, \cdot) = 0$ for either ordering of all such pairs, the cost of the total ordering remains $-4J - 3(m - J)$.

To complete the proof, we must demonstrate that the constructed instance I' of NTS is a YES instance only when the original instance I of Max 2-SAT is a YES instance. Suppose we have an order $\Pi(S)$ of S such that $Cost(\Pi(S)) \leq -4K - 3(m - K)$. Since this cost is negative: (i) all B_1 costs are avoided, and hence (ii) B_2 and $-B_2$ costs are incurred in pairs. Consequently, for each i, all x_{ij} variables come either before or after all x_{ij}' variables and we may equate $\Pi(S)$ unambiguously with an assignment A_Π to the variables of U, that is, A_Π assigns x_i true if all the x_{ij} precede all the x_{ij}', while A_Π assigns x_i false if all the x_{ij} follow all the x_{ij}'.

By (i) and (ii), the cost obtained by $\Pi(S)$ is obtained on the clause specific function values and, since the cost is negative, the B_1 costs are avoided here as well. Observe that no single clause can achieve a cumulative cost of less than -4. To see this, observe that if a clause obtains a -3 value for ordering $z_j < u_j$, it cannot additionally obtain a -2 value without incurring an B_1 cost. Hence, it must be the case that K or more clauses contribute a cost of -4 each. Now observe that if c_j contributes -4, it must be the case that assignment A_Π satisfies c_j. To see this, suppose that A_Π does not satisfy $c_j = \{l_{aj}, l_{bj}\}$ and hence A_Π assigns false to both of these literals. But then by the way A_Π is constructed from $\Pi(S)$, $\Pi(S)$ orders $\neg l_{aj} < l_{aj}$ and $\neg l_{bj} < l_{bj}$. Hence, all consistent orderings of $\neg l_{aj}, l_{aj}, \neg l_{bj}, l_{bj}$ must end in either l_{aj} or l_{bj}, meaning that the clause can contribute a -2 value for at most one of $F(l_{aj}, \neg l_{bj})$ and $F(l_{bj}, \neg l_{aj})$. Since we have already argued that c_j cannot obtain a -3 value for ordering $z_j < u_j$, and additionally obtain a -2 value without incurring an B_1 cost, c_j can contribute at best -3.

Therefore, we have established that K clauses must contribute a cost of -4 each, and that each of these clauses must be satisfied by A_Π. □

Theorem 4.3: OR1O ordering is NP-complete.

Proof: Clearly, OR1O is in **NP**, as we can guess a permutation $\Pi(S)$ and in polynomial time compute its cost. We establish the NP-completeness of OR1O by reducing the restricted version of NTS to it.

Let the instance I of NTS be specified by $S = \{x_1, \ldots, x_n\}$, $F(\cdot,\cdot)$, and bound K_{NTS}. We construct the following instance I' of OR1O.

The set of elements to be ordered is $S' = S \cup \{x_0\}$.

$\overline{DB}$ contains only 1 transaction, and this is a single occurrence of x_0, that is, $count(x_0) = 1$ and $count(x_i) = 0$, $1 \le i \le n$.

$\Theta = \{\theta_{ij} \mid 1 \le i \ne j \le s\}$, where for each ordered pair (i, j) $i \ne j$, $\theta_{ij,N}(x_0) = \theta_{ij,N}(x_j) = 1/2$, $\theta_{ij,M}(x_i) = 1$, and $\theta_{ij} = 0$ elsewhere. Notice $Pr\{\overline{DB} | \theta_{ij}\} = 1/2$ for each $\theta_{ij} \in \Theta$.

Probability function g on Θ is defined as follows. Let $sum\_F = \sum_{(x_i, x_j) \ni 1 \le i \ne j \le n} F(x_i, x_j)$ and define $g(\theta_{ij}) = F(x_i, x_j) / sum\_F$. Clearly the g values sum to 1.

Set the target minimization value $K_{OR1O} = (1/(4 * sum\_F)) * K_{NTS}$.

Let Π' denote any ordering of S'. Then we have

$$rsk(\Pi'; \Theta, g, \overline{DB})$$

$$= \sum_{(x_i,x_j) \ni x_i < x_j \; in \; \Pi' \text{ and } i,j>0} \theta_{ij,M}(x_i)\theta_{ij,N}(x_j) Pr\{\overline{DB}|\theta_{ij}\} g(\theta_{ij})$$

$$+ \sum_{(x_k,x_0) \ni x_k < x_0 \; in \; \Pi'} \left(\sum_{\substack{j=1 \\ k \neq j}}^{s} \theta_{kj,M}(x_k)\theta_{kj,N}(x_0) Pr\{\overline{DB}|\theta_{kj}\} g(\theta_{kj}) \right)$$

$$= 1/4 * \sum_{(x_i,x_j) \ni x_i < x_j \; in \; \Pi' \text{ and } i,j>0} g(\theta_{ij}) \; + \; 1/4 * \sum_{(x_k,x_0) \ni x_k < x_0 \; in \; \Pi'} \left(\sum_{\substack{j=1 \\ k \neq j}}^{s} g(\theta_{kj}) \right)$$

$$= 1/(4 * sum\_F) * \sum_{(x_i,x_j) \ni x_i < x_j \; in \; \Pi' \text{ and } i,j>0} F(x_i, x_j) \; + \; 1/(4 * sum\_F) * \sum_{(x_k,x_0) \ni x_k < x_0 \; in \; \Pi'} \left(\sum_{\substack{j=1 \\ k \neq j}}^{s} F(x_k, x_j) \right)$$

$$= 1/(4 * sum\_F) * Cost_{NTS}(\Pi(S)) \; + \; 1/(4 * sum\_F) * \sum_{(x_k,x_0) \ni x_k < x_0 \; in \; \Pi'} \left(\sum_{\substack{j=1 \\ k \neq j}}^{s} F(x_k, x_j) \right),$$

where $\Pi(S)$ is formed from Π' by removing x_0.

Observe that if x_0 appears in the first position of Π', then the second term, which represents the contribution to risk of pairs involving x_0, is zero; further, this second term is never negative. Hence, if I a YES instance of NTS with $Cost(\Pi) \leq K$, then Π' formed by placing x_0 at the beginning of Π has $rsk(\Pi') \leq 1/(4 * sum\_F) * K$ and I' hence is a YES instance of OR1O. Similarly, if I' is a YES instance of OR1O with $rsk(\Pi') \leq 1/(4 * sum\_F) * K$, then Π obtained by removing x_0 has $Cost(\Pi) \leq K$ and I hence is a YES instance of NTS. □

Observation A.1: Since g is assumed to be a valid probability or density function (in the discrete and continuous cases), it follows that each of $\sum_{\Theta} Pr\{DB|\theta\} * g(\theta)$ and $integral_{\Theta} Pr\{DB|\theta\} * g(\theta) d\theta$ converges to $Pr\{DB\}$.

The following lemmas are used in the proof of Theorem 5.1.

Lemma A.1: Suppose Θ contains an infinite sequence $\theta_1, \theta_2, \ldots, \theta_i, \ldots$ (with each θ_i distinct) nondecreasing in $Pr\{DB|\theta_i\} * g(\theta_i)$, (i.e., $Pr\{DB|\theta_i\} * g(\theta_i) \leq Pr\{DB|\theta_{i+1}\} * g(\theta_{i+1})$, for all $i \geq 1$) with $Pr\{DB|\theta_j\} * g(\theta_j) > 0$ for some $j \geq 1$. Then:

(i) When g is nonzero on a countably infinite number of $\theta \in \Theta$, the convergence of $\sum_{\Theta} Pr\{DB|\theta\}g(\theta)$ is contradicted.

(ii) When g is a probability density function, the convergence of $\underset{\Theta}{integral} Pr\{DB|\theta\}g(\theta)d\theta$ is contradicted.

Proof:

(i) g is nonzero on a countably infinite number of $\theta \in \Theta$: Then the existence of the monotone sequence contradicts the convergence of $\sum_{\Theta} Pr\{DB|\theta\}g(\theta)$, since the series contains the infinite subseries $\sum_{i} Pr\{DB|\theta_i\}g(\theta_i)$ whose terms do not converge to zero.

(ii) g is a probability density function: The existence of the divergent series in (i) over discrete points on which the integral is evaluated implies the integral diverges. □

Lemma A.2: For any transaction pair $\{x, y\}$ and training sample $\overline{DB}$, the existence of bijections

$\Phi_{x,y}: \Theta_{x<y} \rightarrow \Theta_{x>y}$

$\Phi_{y,x}: \Theta_{y<x} \rightarrow \Theta_{y>x}$

such that for every $\theta \in \Theta_{x<y}$, $\theta' \in \Theta_{y<x}$ we have both

$[Pr\{\overline{DB}|\theta\} * g(\theta) \leq Pr\{\overline{DB}|\Phi_{x,y}(\theta)\} * g(\Phi_{x,y}(\theta))]$ and $[Pr\{\overline{DB}|\theta'\} * g(\theta') \leq Pr\{\overline{DB}|\Phi_{y,x}(\theta')\} * g(\Phi_{y,x}(\theta'))]$,

where the inequality is strict for at least one θ or θ':

(a) Is impossible when g is nonzero on a finite number of $\theta \in \Theta$.

(b) Contradicts the convergence of $\sum_{\Theta} Pr\{DB|\theta\}g(\theta)$ when g is nonzero on a countably infinite number of $\theta \in \Theta$.

(c) Contradicts the convergence of $\underset{\Theta}{integral} Pr\{DB|\theta\}g(\theta)d\theta$ when g is a probability density function.

Proof: Assume the existence of such a pair of bijections and, WLOG, assume there exists $\theta_1 \in \Theta_{x<y}$ such that $[Pr\{\overline{DB}|\theta_1\} * g(\theta_1) < Pr\{\overline{DB}|\Phi_{x,y}(\theta_1)\} * g(\Phi_{x,y}(\theta_1))]$. (The statement of the theorem guarantees that at least one strict inequality holds; the argument is symmetric if we must choose $\theta_1 \in \Theta_{y<x}$).

Form an infinite sequence $\theta_1, \theta_2, \ldots, \theta_i, \ldots$ as:

$\theta_{i+1} = \Phi_{x,y}(\theta_i)$, if $\theta_i \in \Theta_{x<y}$ (i.e., if i odd)

$\theta_{i+1} = \Phi_{y,x}(\theta_i)$, if $\theta_i \in \Theta_{y<x}$ (i.e., if i even)

Observe that this sequence is nondecreasing, that is, $Pr\{DB|\theta_i\} * g(\theta_i) \le Pr\{DB|\theta_{i+1}\} * g(\theta_{i+1})$, $i \ge 1$, with strict inequality when $i = 1$.

a) g is nonzero on a finite number of $\theta \in \Theta$. Since only θ_1 could be such that $g(\theta_1) = 0$, some θ_k must repeat in the sequence. If θ_1 is not the first to repeat, the injective property of one of the bijections is violated. If θ_1 is the first to repeat, we have the inconsistent chain

$Pr\{DB|\theta_1\} * g(\theta_1) < Pr\{DB|\theta_2\} * g(\theta_2) \le \cdots \le Pr\{DB|\theta_{k-1}\} * g(\theta_{k-1}) \le Pr\{DB|\theta_1\} * g(\theta_1)$

of inequalities, a contradiction.

b-c) If some θ_k repeats in the sequence, we have the same inconsistent chain of inequalities as in (a). Otherwise, the θ_i in the sequence are distinct and the previous lemma contradicts the convergence of the sum and integral. □

Theorem 5.1: Let g be a probability or density function over Θ and $\overline{DB}$ some training sample. Suppose that for each x, y transaction pair there exists a bijection (possibly depending on $\overline{DB}$)

$\Phi_{x,y}: \Theta_{x<y} \to \Theta_{x>y}$,

which is an x, y-reflection on $\Theta_{x<y}$, such that either:

$\forall \theta \in \Theta_{x<y}$, $Pr\{\overline{DB}|\theta\} * g(\theta) \ge Pr\{\overline{DB}|\Phi_{x,y}(\theta)\} * g(\Phi_{x,y}(\theta))$ or

$\forall \theta \in \Theta_{x<y}$, $Pr\{\overline{DB}|\theta\} * g(\theta) \le Pr\{\overline{DB}|\Phi_{x,y}(\theta)\} * g(\Phi_{x,y}(\theta))$.

Then the collection of pairwise orders defined by:

$x<y$ when a bijection exists to satisfy the first condition, with the inequality strict for at least one $\theta \in \Theta_{x<y}$

$y<x$ when a bijection exists to satisfy the second condition, with the inequality strict for at least one $\theta \in \Theta_{x<y}$

$x = y$ when bijections exist to satisfy both conditions (and hence none of the inequalities is strict.)

is a collection of optimal pairwise orderings given $\overline{DB}$ (that is, each ordering in the collection minimizes each $\{x, y\}$'s contribution to rsk_1 given this fixed $\overline{DB}$).

Proof: Lemma 5.2 implies that, when a bijection Φ exists to satisfy one of the two conditions and the inequality is strict for at least one $\theta \in \Theta_{x<y}$, a bijection cannot exist to satisfy the other condition; hence the ordering rule is well defined.

We now establish that this ordering of each $\{x, y\}$ minimizes this pair's contribution to rsk_1. θ with $g(\theta) = 0$ have no effect on risk contribution. For a fixed x, y, the loss incurred on $\theta \in \Theta_{x=y}$ by any ranking of x, y is identical, since in this case $\theta_M(x) * \theta_N(y) = \theta_M(y) * \theta_N(x)$. Hence, in the following analysis, we suppress θ such that $g(\theta) = 0$ or $\theta \in \Theta_{x=y}$.

It follows from (1) of Section 3.2 that the contribution to $rsk_1(\cdot, \overline{DB})$ of each $\{x, y\}$ transaction pair assumes one of the following values:

If we order as $x<y$:

$$\underset{\Theta_{x<y}}{integral} \theta_M(x) * \theta_N(y) * Pr\{DB|\theta\} g(\theta) + \Phi(\theta_M)(x) * \Phi(\theta_N)(y) * Pr\{DB|\Phi(\theta)\} g(\Phi(\theta))\ d\theta$$

If we order as $x>y$:

$$\underset{\Theta_{x<y}}{integral} \theta_M(y) * \theta_N(x) * Pr\{DB|\theta\} g(\theta) + \Phi(\theta_M)(y) * \Phi(\theta_N)(x) * Pr\{DB|\Phi(\theta)\} g(\Phi(\theta))\ d\theta$$

If we order as $x = y$: average of the above two quantities.

Since $\theta_N(x) = \Phi(\theta_N)(y)$ and $\theta_M(x) = \Phi(\theta_M)(y)$, we can write the contributions as

If we order as $x<y$:

$$\underset{\Theta_{x<y}}{integral} \theta_M(x) * \theta_N(y) * Pr\{DB|\theta\} g(\theta) + \theta_M(y) * \theta_N(x) * Pr\{DB|\Phi(\theta)\} g(\Phi(\theta))\ d\theta$$

If we order as $x>y$:

$$\underset{\Theta_{x<y}}{integral} \theta_M(y) * \theta_N(x) * Pr\{DB|\theta\} g(\theta) + \theta_M(x) * \theta_N(y) * Pr\{DB|\Phi(\theta)\} g(\Phi(\theta))\ d\theta$$

If we order as $x = y$: average of the above two quantities.

Since, by definition, $\theta_r(x)<\theta_r(y)$ (and hence $\theta_M(x) * \theta_N(y) < \theta_M(y) * \theta_N(x)$) for each $\theta \in \Theta_{x<y}$: $Pr\{DB|\theta\} * g(\theta) \geq Pr\{DB|\Phi(\theta)\} * g(\Phi(\theta))$ for each $\theta \in \Theta_{x<y}$ implies the first contribution is no more than the second, while

$Pr\{DB|\theta\} * g(\theta) \le Pr\{DB|\Phi(\theta)\} * g(\Phi(\theta))$ for each $\theta \in \Theta_{x<y}$ implies the second contribution is no more than the first. (When there is equality for each $\theta \in \Theta_{x<y}$, the contributions are identical.) Hence, the order specified for the pair $\{x, y\}$ orders the transaction pair in a manner that minimizes its contribution to rsk_1. $\Box$

Theorem 5.2: If g is consistent with anomaly detection and has the additional property that, for each $\theta \in \Theta_{AD}$, $g(\theta) = g(\theta_{x/y})$, the frequentist detector defined by $1 = 1/count(DB, x)$ minimizes risk among Class 1 detectors.

Proof: We establish that, when the conditions of the current theorem are satisfied, bijections exist to satisfy the conditions of Theorem 5.1 and that $pd(x) = 1/count(DB, x)$ orders each $\{x, y\}$ pair identically with the order prescribed by Theorem 5.1. (Except in the trivial special case handled at the proof's conclusion.) Since pd maps S to $\mathbf{R}$, the resulting collection of pairwise orderings necessarily is consistent and hence Theorem 5.1 implies the optimality of pd among Class 1 detectors.

Observe first that each member θ of the partition cell $\Theta_{x<y} \subset \Theta_{AD}$ has the property that $\theta_N(x) > \theta_N(y)$, or $\theta_N(x) = \theta_N(y)$ and $\theta_M(x) < \theta_M(y)$. Similarly, each member θ of the partition cell $\Theta_{x>y} \subset \Theta_{AD}$ has the property that $\theta_N(x) < \theta_N(y)$, or $\theta_N(x) = \theta_N(y)$ and $\theta_M(x) > \theta_M(y)$. We define Φ to satisfy Theorem 5.1 as the x/y exchange operator from Definition 5.4 (clearly this is an x, y-reflection) and claim this Φ is a bijection from $\Theta_{x<y}$ to $\Theta_{x>y}$. The first step is to establish that if $\theta \in \Theta_{AD}$, then also $\Phi(\theta) \in \Theta_{AD}$. The condition for $\Phi(\theta)$'s membership in Θ_{AD} obviously is preserved on w and z different from x and y, since these components are identical in θ and $\Phi(\theta)$. It holds also if one of w and z is x or y; WLOG, consider the pair x and z, z different from y. Since in θ the condition must hold for y and z, it must hold for x and z in $\Phi(\theta)$. Finally, consider x, y and suppose $\Phi(\theta)_r(y) < \Phi(\theta)_r(x)$. Then it must be that $\theta_r(x) < \theta_r(y)$ and hence that $\theta_N(x) \ge \theta_N(y)$. Therefore $\Phi(\theta_N)(y) \ge \Phi(\theta_N)(x)$, as required. It then is apparent that Φ maps from $\Theta_{x<y}$ to $\Theta_{x>y}$. To see Φ is a bijection, define Φ^{-1} to operate on members of $\Theta_{x>y}$ in the same manner that Φ operates on members of $\Theta_{x<y}$; it is apparent that Φ and Φ^{-1} are inverses.

To see that Φ obeys the inequality conditions of Theorem 5.1, consider three cases for $\overline{DB}$:

1. $count(\overline{DB}, x) < count(\overline{DB}, y)$: It follows from the Theorem 2.1 that $Pr\{\overline{DB}|\theta\}$ and $Pr\{\overline{DB}|\Phi(\theta)\}$ differ

only in that $Pr\{\overline{DB}|\theta\}$ contains the terms $\theta_N(x)^{count(\overline{DB},x)}$ and $\theta_N(y)^{count(\overline{DB},y)}$ while $Pr\{\overline{DB}|\Phi(\theta)\}$ contains instead $\theta_N(x)^{count(\overline{DB},y)}$ and $\theta_N(y)^{count(\overline{DB},x)}$. Since $\theta_N(x) \geq \theta_N(y)$ and $g(\theta) = g(\Phi(\theta))$, we have $Pr\{\overline{DB}|\theta\} * g(\theta) \leq Pr\{\overline{DB}|\Phi(\theta)\} * g(\Phi(\theta))$ for every $\theta \in \Theta_{x<y}$.

2. $count(\overline{DB}, x) > count(\overline{DB}, y)$: We similarly obtain $Pr\{\overline{DB}|\theta\} * g(\theta) \geq Pr\{\overline{DB}|\Phi(\theta)\} * g(\Phi(\theta))$ for every $\theta \in \Theta_{x<y}$.
3. $count(\overline{DB}, x) = count(\overline{DB}, y)$: We similarly obtain $Pr\{\overline{DB}|\theta\} * g(\theta) = Pr\{\overline{DB}|\Phi(\theta)\} * g(\Phi(\theta))$ for every $\theta \in \Theta_{x<y}$.

We now establish that *pd* orders each $\{x, y\}$ pair in the manner prescribed by Theorem 5.1. (This assumes at least one $\theta \in \Theta_{x<y}$ is such that $\theta_N(x) \neq \theta_N(y)$; the special case where this assumption fails is handled below.) To see this, we need only observe that when $count(\overline{DB}, x) \neq count(\overline{DB}, y)$, the inequality obeyed by the bijection $\Phi_{x<y}$ is strict on any $\theta \in \Theta_{x,y}$ such that $\theta_N(x) \neq \theta_N(y)$, and the direction of the inequality prescribes the same ordering of x, y as that induced by *pd*; when $count(\overline{DB}, x) = count(\overline{DB}, y)$ both Theorem 5.1 and *pd* order $x = y$.

In the trivial case that $\theta_N(x) = \theta_N(y)$ for all $\theta_N \in \Theta_{x<y}$, the correspondence induced by Φ (i.e., as used in the proof of Theorem 5.1) makes it easy to see that any of the three orderings of x, y results in the same contribution to risk, since in this case $Pr\{\overline{DB}|\theta\} * g(\theta) = Pr\{\overline{DB}|\Phi(\theta)\} * g(\Phi(\theta))$ for every $\overline{DB}$ and every $\theta \in \Theta_{x<y}$.

Hence, *pd* induces a collection of optimal pairwise orderings which, since *pd* maps from $S \rightarrow \mathbf{R}$, is consistent. □

Lemma 5.3: Let $A \subset U$ be a proper subset of the set U of all attributes. Let $\theta \in \Theta_A$ and let x and y be transactions such that $\theta_N(x) \neq \theta_N(y)$ or $\theta_M(x) \neq \theta_M(y)$. Then $\theta_{x/y} \notin \Theta_B$ for any proper subset $B \subset U$. (The result assumes each attribute has more than 2 distinct values.)

Proof: First we argue that unless $A \cup B = U$, we cannot have $\theta_{x/y} \in \Theta_B$. Suppose $A \cup B \neq U$. Then class $x[A]$ contains z different from x, but with $x[A \cup B] = z[A \cup B]$. Since $z \in [x]_A$, we have $\theta_N(x) = \theta_N(z)$ and $\theta_M(x) = \theta_M(z)$. Since z is neither x nor y, we must have ${\theta_{x/y}}_N(x) \neq {\theta_{x/y}}_N(z)$ or ${\theta_{x/y}}_M(x) \neq {\theta_{x/y}}_M(z)$, by the lemma's premise. But $x[B] = z[B]$ and hence $\theta_{x/y} \notin \Theta_B$.

Now we argue that unless $A \subset B$, we cannot have $\theta_{x/y} \in \Theta_B$. Suppose $a \in (A - B)$. Choose any z such that $x[A \cup B - \{a\}] = z[A \cup B - \{a\}]$, and $x[a] \neq z[a]$ and $z[a] \neq y[a]$. Choose any x' different from x such that $x[A] = x'[A]$. Choose $z' \in [z]_A$, different from z, so that $x'[A \cup B - \{a\}] = z'[A \cup B - \{a\}]$. Observe that:

since $x[A] = x'[A]$, $\theta_N(x) = \theta_N(x')$ and $\theta_M(x) = \theta_M(x')$

since $z[A] = z'[A]$, $\theta_N(z) = \theta_N(z')$ and $\theta_M(z) = \theta_M(z')$

Further, since $x[B] = z[B]$ and $x'[B] = z'[B]$, if $\theta_{x/y} \in \Theta_B$, we must have

$\theta_{x/y_N}(x) = \theta_{x/y_N}(z)$ and $\theta_{x/y_M}(x) = \theta_{x/y_M}(z)$

$\theta_{x/y_N}(x') = \theta_{x/y_N}(z')$ and $\theta_{x/y_M}(x') = \theta_{x/y_M}(z')$

But $\theta_N(x) \neq \theta_{x/y_N}(x)$ or $\theta_M(x) \neq \theta_{x/y_M}(x)$ by the lemma's premise. Further, none of x', z, or z' is x or y; hence the θ_N and θ_M values for each of these transactions is the same under θ and $\theta_{x/y}$. Consequently, the above equalities cannot be satisfied simultaneously. □

Theorem 5.3: If g has mass 1 on $\Theta_A \cap \Theta_{AD}$, and if for each x, y pair $g(\theta) = g(Q_A(x, y; \theta))$ for all $\theta \in \Theta_A \cap \Theta_{AD}$, then the detector $pd_A(x) = 1/count(DB, x[A])$ minimizes risk among Class 1 detectors.

Proof: As in the proof of Theorem 5.2, we establish that, when the conditions of the current theorem are satisfied, bijections exist to satisfy the conditions of Theorem 5.1 and that $pd_A(x) = 1/count(DB, x[A])$ orders each $\{x, y\}$ pair identically with the order prescribed by Theorem 5.1. (Except in the trivial special case handled at the proof's conclusion.) Since pd_A maps S to $\mathbf{R}$, the resulting collection of pairwise orderings is necessarily consistent and hence Theorem 5.1 implies the optimality of pd_A among Class 1 detectors.

Observe first that each member θ of the partition cell $\Theta_{x<y} \subset \Theta_A \cap \Theta_{AD}$ has the property that $\theta_N(x) > \theta_N(y)$, or $\theta_N(x) = \theta_N(y)$ and $\theta_M(x) < \theta_M(y)$. Similarly, each member θ of the partition cell $\Theta_{x>y} \subset \Theta_A \cap \Theta_{AD}$ has the property that $\theta_N(x) < \theta_N(y)$, or $\theta_N(x) = \theta_N(y)$ and $\theta_M(x) > \theta_M(y)$. Of course these relationships hold between all members of $x[A]$ and $y[A]$. We define Φ as the function $Q_A(x, y; \cdot)$ from above (clearly this is an x, y-reflection) and claim this Φ is a bijection from $\Theta_{x<y}$ to $\Theta_{x>y}$. The argument that Φ is a bijection closely follows that from Theorem 5.2; the three cases here being w and z are in classes other than $x[A]$ and $y[A]$, one of w and z is in one of the classes $x[A]$ and $y[A]$, and $w \in [x]_A$ and

$z \in [y]_A$.

To see that Φ obeys the inequality conditions of Theorem 5.1, consider three cases for $\overline{DB}$:

1. $count(\overline{DB}, x[A]) < count(\overline{DB}, y[A])$: It follows from the Theorem 2.1 that $Pr\{\overline{DB}|\theta\}$ and $Pr\{\overline{DB}|\Phi(\theta)\}$ differ only in that $Pr\{\overline{DB}|\theta\}$ contains the terms $\theta_N(x)^{count(\overline{DB},x[A])}$ and $\theta_N(y)^{count(\overline{DB},y[A])}$ while $Pr\{\overline{DB}|\Phi(\theta)\}$ contains instead $\theta_N(x)^{count(\overline{DB},y[A])}$ and $\theta_N(y)^{count(\overline{DB},x[A])}$. (Note that since in each of θ_N and $\Phi(\theta_N)$ components for transactions agreeing on A are identical to each other, multiple terms of the product collapse to single terms such as the above.) Since $\theta_N(x) \geq \theta_N(y)$ and $g(\theta) = g(\Phi(\theta))$, we have $Pr\{\overline{DB}|\theta\} * g(\theta) \leq Pr\{DB|\Phi(\theta)\} * g(\Phi(\theta))$ for every $\theta \in \Theta_{x<y}$.
2. $count(\overline{DB}, x[A]) > count(\overline{DB}, y[A])$: We similarly obtain $Pr\{\overline{DB}|\theta\} * g(\theta) \geq Pr\{\overline{DB}|\Phi(\theta)\} * g(\Phi(\theta))$ for every $\theta \in \Theta_{x<y}$.
3. $count(\overline{DB}, x[A]) = count(\overline{DB}, y[A])$: We similarly obtain $Pr\{\overline{DB}|\theta\} * g(\theta) = Pr\{\overline{DB}|\Phi(\theta)\} * g(\Phi(\theta))$ for every $\theta \in \Theta_{x<y}$. Note that this case holds whenever $x[A] = y[A]$.

The argument that the collection of orders induced by pd_A is pairwise optimal is completely analogous to that of Theorem 5.2. When at least one $\theta \in \Theta_{x<y}$ is such that $\theta_N(x) \neq \theta_N(y)$, $\{x, y\}$ is ordered in the manner prescribed by Theorem 5.1; otherwise, any of the three orderings of x, y results in the same contribution to risk. Hence, pd_A induces a collection of optimal pairwise orderings which, since pd maps from $S \rightarrow \mathbf{R}$, is consistent. □

References

1. Angluin, D. and Smith, C., "Inductive inference: Theory and methods," *ACM Computing Surveys*, vol. 15, no. 3, pp. 237-269, 1983.

2. Baim, P. W., "A method for attribute selection in inductive learning systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 888-896, 1988.

3. Blumer, A., Ehrenfeucht, A., Haussler, D, and Warmuth, M., "Learnability and the Vapnik-Chervonenkis dimension," *J. ACM*, vol 36, no. 4, pp. 929-965, 1989.

4. Doak, J., "An evaluation of feature selection methods and their application to computer security," University of California at Davis, Technical Report CSE-92-18, 1992.

5. Duda, R. and Hart, P., *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.

6. Fukunaga, K., *Introduction to Statistical Pattern Recognition*,Second Edition,Academic Press, Inc., San Diego, 1990.

7. Garey, M. and Johnson, D., *Computers and Intractability*, W. H. Freeman, San Fransisco, 1979.

8. Gelman, A., Carlin, J., Stern H. and Rubin D., *Bayesian Data Analysis*, Chapman & Hall, London, 1995.

9. Gonzalez, R.C. and Thomason, M.G., *Syntactic Pattern Recognition, an Introduction*, Addison-Wesley, Reading, Mass., 1978.

10. Haussler, D., "Learning conjunctive concepts in structured domains," *Machine Learning*, vol. 4, no. 1, 1989.

11. Helman, P. and Bhangoo, J., "A statistically based system for prioritizing information exploration under uncertainty," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 27, no. 4, pp. 449-466, 1997.

12. Helman, P. and Liepins, G., "Statistical foundations of audit trail analysis for the detection of computer misuse," *IEEE Trans. on Software Engineering*, vol. 19, no. 9, pp. 886-901, 1993.

13. Littlestone, N., "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm," *Machine Learning*, vol. 2, no. 2, pp. 285-318, 1988.

14. Sobel, M. J., "Complete Ranking Procedures With Appropriate Loss Functions", *Communications in Statistics - Theory and Methods*, vol. 19, no. 12, pp. 4525-4544, 1990.

15. Tanner, M. A., *Tools for Statistical Inference - Observed Data and Data Augmentation Methods*, Springer-Verlag, New York, 1991.

16. Tou, J.T. and Gonzalez, R.C., *Pattern Recognition Principles*, Addison-Wesley, Reading, Mass., 1974.

17. Valiant, L., "A theory of the learnable," *Comm. ACM*, vol. 27, no. 11, pp. 1134-1142, 1984.

18. Watanabe, S., *Pattern Recognition: Human and Mechanical*, John Wiley and Sons, New York, 1985.