Information Awareness: A Prospective Technical Assessment

David Jensen
Computer Science Department
University of Massachusetts
Amherst, MA 01003
413-545-9677
jensen@cs.umass.edu

Matthew Rattigan
Computer Science Department
University of Massachusetts
Amherst, MA 01003
413-545-1519
rattigan@cs.umass.edu

Hannah Blau
Computer Science Department
University of Massachusetts
Amherst, MA 01003
413-545-1519
blau@cs.umass.edu

ABSTRACT

Recent proposals to apply data mining systems to problems in law enforcement, national security, and fraud detection have attracted both media attention and technical critiques of their expected accuracy and impact on privacy. Unfortunately, the majority of technical critiques have been based on simplistic assumptions about data, classifiers, inference procedures, and the overall architecture of such systems. We consider these critiques in detail, and we construct a simulation model that more closely matches realistic systems. We show how both the accuracy and privacy impact of a hypothetical system could be substantially improved, and we discuss the necessary and sufficient conditions for this improvement to be achieved. This analysis is neither a defense nor a critique of any particular system concept. Rather, our model suggests alternative technical designs that could mitigate some concerns, but also raises more specific conditions that must be met for such systems to be both accurate and socially desirable.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining.

Keywords

Information awareness, relational data mining, social network analysis, ranking classifiers, iterative classification, collective classification, TIA, privacy, technology assessment.

1. INTRODUCTION

Proposals to apply data mining techniques to preventing domestic terrorism have received unprecedented attention in the past year. Perhaps the most widely reported proposal arises out of Terrorism (formerly "Total") Information Awareness (TIA), an ongoing research program at the U.S. Defense Advanced Research Projects Agency. News reports about the potential application of technologies developed under the TIA research program have emphasized the potential size and scope of such a system. For example, Robert O'Harrow Jr., of the Washington Post [13], characterized TIA as developing "new technologies to sift through 'ultra-large' data warehouses and networked computers in search of threatening patterns among everyday transactions, such as credit card purchases and travel reservations." The article identifies a potential end point of the work as "a global computer-surveillance system to give U.S. counterterrorism officials access to personal information in government and commercial databases around the world."

In addition to the hypothetical system discussed in the media reports on TIA, a number of other potential systems for broad-scale analysis of data on U.S. citizens have been proposed since the terrorist attacks of September 11, 2001. For example, Oracle CEO Larry Ellison [4] argued strongly for a national database in a Wall Street Journal article in October of 2001: "Do we need more databases? No, just the opposite. The biggest problem today is that we have too many. The single thing we could do to make life tougher for terrorists would be to ensure that all the information in myriad government databases was integrated into a single national file."

These proposals for "information awareness" systems share several common characteristics. First, all are hypothetical. No such system has actually been developed or deployed, and implementation and ongoing government use of such systems would require large changes in U.S. law. However, serious *prospective* examinations are underway, both inside and outside of government, to assess the potential efficacy and impacts of information awareness systems. Second, the scope of these hypothetical systems is extraordinarily broad. They would correlate and examine extremely large numbers of records, including some collected without pri picion about the individuals or activities represented in the records. Third, they are attempting to detect an extremely rare phenomenon. For example, even extremely high estimates place the number of potential terrorists in the U.S. at less than one in 10,000.

Proposals for information awareness systems have drawn widespread media attention, particularly with respect to their accuracy and potential impact on privacy. Concerns about accuracy have focused on whether such systems would falsely label innocent persons as terrorists (false positives) and whether they would they would miss terrorists amid the vast numbers of innocent persons (false negatives). Concerns about privacy have focused on whether increased data collection or the fusion of existing databases would pose serious threats to the privacy of U.S. citizens, and whether the creation and use of extremely large, centralized databases presents unacceptable risks of theft or unauthorized access.

SIGKDD '03, August 24-27, 2003, Washington, DC, USA.
Copyright 2003 ACM 1-58113-737-0/03/0008...$5.00.

The privacy impact of information awareness systems could be extraordinarily large, as could the potential impact of future terrorist attacks. As a result, the issues surrounding these systems deserve wide, active, and informed public debate. Unfortunately, this public debate is hampered by the technical nature of the proposed systems. The accuracy and privacy impacts of any real system will depend critically on the specific technologies employed, and the impact of those design choices is difficult to assess without technical expertise. As a result, the active participation of experts in data mining technologies is crucial to well-informed debate.[1]

We intend this paper as an initial step in that direction. We examine two of the most common technical critiques of information awareness systems, that: 1) proposed systems will have unacceptably high numbers of false positives (e.g., innocent individuals identified as terrorists); and 2) extremely large, centralized databases will be necessary for these systems to function effectively. These two critiques lie at the heart of questions about accuracy and privacy impacts. An extremely high number of false positives implies an inaccurate system that will be of little use to law enforcement agencies. In addition, a high number of false positives implies the potential for serious violations of civil liberties, due to protracted investigations or arrests of entirely innocent persons. Similarly, the necessity of a large, centralized database implies a host of privacy risks due to data theft or unauthorized access.

Specifically, we argue that these two critiques are based on a naive model of systems for knowledge discovery and data mining. This naive model assumes that data consist of statistically independent instances, that analysis uses a simple binary classifier, and that analysis consists of applying that classifier in a single pass through the data. Modern approaches to learning and applying classifiers can alter each of these assumptions, and thus the naive model is a misleading tool for informing public debate.

We provide an enhanced model that alters these key assumptions. Specifically, the enhanced model assumes that data consist of instances connected by meaningful transactions, that analysis uses a ranking classifier, and that analysis consists of applying that classifier as part of a larger iterative algorithm. We construct a simulation using this enhanced model and use the simulation to explore the characteristics of the model. Specifically, we examine the accuracy and the data requirements of the enhanced model. We show that accuracy can be greatly improved, that data can be accessed in stages, and that substantially smaller amounts of data can be accessed in later stages of inference.

We believe that the enhanced model could lead to better assessments of systems for information awareness. It offers more meaningful critiques of information awareness systems, highlighting both spurious objections and serious potential deficiencies. These critiques apply to a broad range of systems, including systems for detecting money laundering [19], stock fraud, and cellular phone fraud [5].

That said, two caveats are in order. First, the enhanced model is still extremely simple. We have kept the model simple intentionally, so it is theoretically tractable and relatively easy to convey in a single technical paper. It ignores a large number of practical complexities of real data, and thus it is intended for illustrative purposes only.

Second, this paper is not a defense of any proposed system or any particular technical approach. Rather, it critiques an overly simplistic conceptual model of information awareness systems that is being used for policy discussions, and it attempts to create an improved conceptual model. It considers only two of the many issues surrounding proposed systems. Other issues include whether attempts to identify terrorists will lead to racial and ethnic profiling, whether attendant administrative changes would erode the current separation between intelligence and law enforcement agencies, and whether the actual use of information awareness systems would violate Fourth amendment protections against unreasonable searches and seizures. These issues are both valid and extremely important, but they are outside the scope of this paper.

2. TECHNICAL CRITIQUES

As already noted, two technical critiques of proposed information awareness systems are common. First, critics charge that, given the extremely low incidence of positive cases, nearly any error rate in the classifier will produce an extremely large number of false positives. Second, critics charge that, to obtain a highly accurate classifier, an enormous amount of information will need to be stored in a centralized database. In this section, we examine these two critiques in more detail.

2.1 False Positives

The vast numbers of individuals that could potentially be screened by an information awareness system can lead simple classifiers to produce a vast number of false positives. For example, in an open letter to Congress regarding DARPA's TIA research program [1], the U.S. Public Policy Committee of the Association for Computing Machinery explained that:

> "Any type of statistical analysis inevitably results in some number of false positives — in this case incorrectly labeling someone as a potential terrorist. As the entire population would be subjected to TIA surveillance, even a small percentage of false positives would result in a large number of law-abiding Americans being mistakenly labeled. For example, suppose the system has an 99.9% accuracy rate. We believe that having only 0.1% of records being misclassified as belonging to potential terrorists would be an unachievable goal in practice. However, if records for everyone in the U.S. were processed monthly, even this unlikely low rate of false positives could result in as many as 3 million citizens being wrongly identified each year. More realistic assumptions about the percentage of false positives would drive the number even higher."[2]

[1] Public debate about information awareness systems faces an additional challenge — the legitimate need for secrecy to avoid informing terrorist groups about the specific details of counter-terrorism efforts. However, that topic is beyond the scope of this paper.

[2] ACM's calculations rest on an unrealistic assumption unrelated to those discussed elsewhere in this paper. The calculations assume monthly monitoring and then cite an annual magnitude of false positives that is approximately ten times the monthly magnitude. This assumes that errors of the classifiers used in each month will be largely independent, an unlikely scenario. This assumption inflates the magnitude of false positives as much as 10 times.

A similar critique was made in a recent editorial in Scientific American [18]:

> "...terrorism is very rare — which is good for us but bad for data miners. Even with a low error rate, the vast majority of red flags will be red herrings. Suppose that there are 1,000 terrorists in the U.S. and that the data mining process has an amazing 99 percent success rate. Then 10 of the terrorists will probably slip through — and 2.8 million innocent people will also be fingered."

This critique is not unique to information awareness systems; it has been made of a large number of screening systems. For example, it was noted as a potential problem with systems for screening wire transfers for evidence of money laundering [12]: "As a result [of the false positive problem], a group of [wire] transfers identified by the system as illegitimate would consist almost entirely (99 percent) of transfers that are actually legitimate." It has also been noted as a problem for polygraph systems [10] and screening protocols for relatively rare diseases.

2.2 Centralized Database

Another common critique concerns the necessity of a massive, centralized database. If an information awareness system uses a binary classifier with single-pass inference, it appears difficult to escape the notion that the system would need either a single massive database or complete access to many smaller distributed databases that could approximate a single database.[3] Without such access, the classifier might lack values for critical model components, and accuracy could suffer.

This critique is central to most of the prominent objections to the TIA research program. For instance, one early editorial [17] painted a particularly graphic view: "Every purchase you make with a credit card, every magazine subscription you buy and medical prescription you fill, every Web site you visit and e-mail you send or receive, every academic grade you receive, every bank deposit you make, every trip you book and every event you attend — all these transactions and communications will go into what the Defense Department describes as 'a virtual, centralized grand database.'"

The potential existence of such a database leads directly to concerns about security and privacy. According to the ACM letter [1]: "Immense databases, such as are being proposed by TIA — whether operated by governmental or commercial organizations — represent substantial security and privacy risks in their own right. An all-encompassing database, compiled from private and governmental databases including financial, medical, educational, telephone, and travel records, will contain large quantities of sensitive information." ACM goes on to note potential problems of theft, unauthorized access, and institutional misuse, all things made much easier when a database is massive and centralized.

3. MODELING INFORMATION AWARENESS SYSTEMS

As mentioned in the introduction, the current critiques of information awareness systems use a simplistic model of data mining. This model is largely implicit in published critiques, although three elements are relatively easy to identify: 1) propositional data; 2) binary classifiers; and 3) single-pass inference. In this section, we examine each of these assumptions, and contrast it with an alternative that forms an element of the enhanced model. In later sections we will show how the elements can interact to obviate critiques about the number of false positives and the necessity of a single massive database.

3.1 Propositional vs. Relational Data

Most published critiques of information awareness systems have made simplistic assumptions about the type of data that would be employed. Specifically, they have assumed that data will consist of *propositional instances*, in which each instance is characterized by a set of simple propositions (e.g., *age=32*, *gender=male*). In propositional data, each individual is assumed to be statistically independent of any other; knowing something about one individual is assumed to tell you nothing about any other individual. For example, a medical diagnosis system using propositional data would diagnose each patient independently, without using information about the relationships between one patient and other potential patients.

This assumption of propositional data is implicit in many of the critiques of information awareness systems. An identical assumption was identified in an early study of proposed systems for identifying money laundering in large databases of financial transactions [12].[4] Simplistic conceptual models of systems for identifying illicit wire transfers assumed that each transfer would be examined independently, without considering information about the related bank accounts, account holders, or transactions.

An alternative is to assume that the relevant data will be *relational* rather than propositional. Relational data are contrasted with propositional data in the schematic shown in Figure 1. Relational data provide connections (or *relations*) between individual data records. For example, in medical diagnosis, we know that the assumption of independence is a poor one for many diseases and conditions. If a patient is suspected of having a genetic disorder, then the medical history of close relatives could be predictive. If a potential disease is communicable, then knowledge of recent contacts with family members or c m that disease could be predictive. Relational data can represent these relationships and make them available to algorithms for learning statistical models and making inferences with those models.

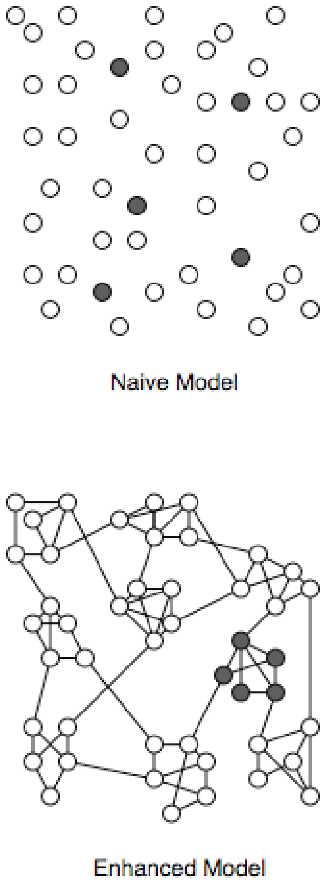
Relational data lie at the heart of approaches to understanding organizations and social groups [23], to analyzing organized crime and terrorism [21], and to identifying financial fraud [7,12,19,20]. Despite the assumptions about propositional data that underlie nearly all technical critiques of information awareness systems, the examples cited in these critiques nearly always mention relational data in the form of commercial

[3] Alternative approaches — often categorized as "privacy preserving data mining" — exist that can safeguard individual records while still allowing the creation and use of statistical models. However, these technologies are not our focus here.

[4] One of the authors of the present paper (Jensen) coauthored that study while serving as an analyst with the Congressional Office of Technology Assessment.

transactions and communications. For example, early editorials criticizing the potential use of TIA technologies [13,17] specifically cite many examples of relational data, including financial transactions, web navigation, email messages, and travel arrangements.

Analysis of relational data is a rapidly growing area within the larger research community interested in machine learning, knowledge discovery, and data mining. Several recent workshops [3,6,8] have focused on this precise topic, and another DARPA research program — Evidence Extraction and Link Discovery (EELD) —focuses on extracting, representing, reasoning with, and learning from relational data.[5] A growing list of algorithms has been developed that learn probabilistic models from relational data, and important new research results in this field are emerging every month.


Naive Model

Enhanced Model

Figure 1: Data representation for naive and enhanced models

3.2 Binary vs. Ranking Classifiers

Much of the attention devoted to systems for information awareness has focused on the accuracy and data requirements of *classifiers*. For the purposes of this discussion, a classifier receives input about each data instance (typically a vector of values for a given set of variables) and produces output in the form of the value for a single discrete or continuous variable. For example, a classifier for medical diagnosis might receive input in the form of a vector of values indicating the results of diagnostic tests, and output a number indicating the probability that the patient has a particular disease.

Much of the discussion of systems for information awareness has made particularly limiting assumptions about the type of classifier that could be used. Specifically, many accounts have assumed that a single *binary* classifier would be applied to a set of individuals. That is, the classifier is assumed to output only a binary class label indicating whether the system predicts a positive class label (e.g., terrorist) or a negative label (e.g., non-terrorist). The output of such a classifier can be completely summarized by a contingency table, such as the one shown in Table 1, that cross-tabulates the actual and predicted class labels. The classifier is assumed to output no additional information that could distinguish among instances assigned to each class.

Table 1: Contingency table for a binary classifier

		Actual Class Label	
		+	–
Predicted Class Label	+	True Positive (TP)	False Positive (FP)
	–	False Negative (FN)	True Negative (TN)

An alternative to this simplistic assumption is a *ranking classifier*. Such a classifier assigns a real-valued *score* to each instance, where a higher value implies a greater probability of having a positive class label. For example, such a score might estimate the probability of a positive label, although for the purposes of this discussion, the ranking classifier need only produce a score that establishes a relatively fine-grained partial order over all instances. Any given score can serve as a threshold that converts a ranking classifier into a binary classifier. A ranking classifier thus defines a family of binary classifiers, where each specific binary classifier corresponds to one or more thresholds.

The performance of a ranking classifier can be can be visualized by using a Receiver Operating Characteristic (ROC) curve [14,15]. The two-dimensional *ROC space* is defined by the false positive rate on the x-axis and the true positive rate on the y-axis. The true positive rate of the classifier is $TP/(TP+FN)$ and the false positive rate is $FP/(FP+TN)$, where the elements of the equation correspond to the contingency table entries shown in Table 1. Example ROC curves are shown in Figure 2.

The point (0,1) in ROC space corresponds to a perfect classifier that correctly classifies all instances. The point (0,0) corresponds to a classifier that labels all instances negative, and (1,1) corresponds to a classifier that labels all instances positive. The line $x=y$ (shown as a dotted line in Figure 2a) corresponds to a classifier that assigns class labels at random. Given a ranking classifier C_r, we can derive a series of binary classifiers $C_r(T_0)$, $C_r(T_1)$, ... by choosing the threshold value T_i. Each binary classifier corresponds to a point in ROC space. As we vary the threshold T throughout its range, we get a series of points that form the ROC curve for the ranking classifier C_r.

[5] The authors' research is partially supported by EELD.

The ROC curve provides a visualization of C_r's performance across all possible cost and class distributions. The choice of a particular threshold for C_r depends on a particular pair of cost and class distributions. The threshold that is appropriate if negatives and positives occur in roughly equal proportion will not be appropriate if the positives are rare compared with the negatives. If false positives have higher cost than false negatives, we would choose a higher threshold to avoid capturing instances that should be labeled negative.

The three curves shown in Figure 2a correspond to different ranking classifiers. All perform better than random, and thus lie above the dotted line $x=y$. Regardless of the cost and class distribution, C_1 performs worse than C_2. Depending on the cost and class distributions, C_3 can be either the best classifier or the worst.

If the cost and class distributions are known, a particular point on an ROC curve can be selected as the optimal classifier [16]. The cost and class distributions define an iso-performance line in ROC space, shown as dotted lines in Figure 2b. All the classifiers whose points lie on the iso-performance line have identical expected cost under the specified class and cost distributions. We draw a line having the specified slope in the upper left-hand corner of the ROC plot and move it in a direction perpendicular to its slope until the line comes into contact with the ROC curve for a ranking classifier C_r. The point at which the line is incident to the ROC curve identifies the optimal binary classifier $C_r(T_{opt})$ for that cost and class distribution.

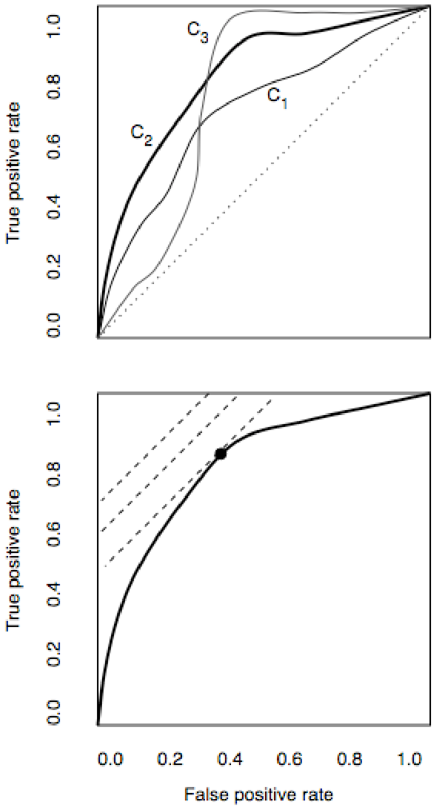


Figure 2: ROC curves

When the class distributions and error costs are known, we can use the iso-performance line to identify the best ranking classifier and the optimal threshold for that classifier. However, in many practical situations, the class and cost distributions are unknown or dynamic. In some of these s still relatively easy to identify a preferred classifier. If one ranking classifier C_r has an ROC curve that *dominates* all others, then we can declare C_r uniformly superior to its competitors. Here *dominance* indicates that the given curve is closer to the perfect performance point (0,1) over the entire range of values. For example, in Figure 2b, C_2 dominates C_1 but not C_3.

In practice, however, it is relatively rare for a single ranking classifier to dominate all competitors. This returns us to the problem of determining the costs of misclassification. In the case of terrorism,, both casting suspicion on an entirely innocent persons (false positives) and failing to prevent terrorist attacks (false negatives) have serious consequences, but reasonable observers can disagree on the relative costs of these two types of misclassification.

If there is no dominating ROC curve and there is uncertainty about cost and class distributions, one method for assessing a ranking classifier is by the total area under its ROC curve. This area represents the performance of the classifier, averaged across all possible cost and class distributions. The area-under-the-curve (AUC) represents a reasonably good measure of performance when cost or class distributions are uncertain, and we will use AUC as an evaluation criterion for the remainder of this paper.

Why concern ourselves with ranking classifiers? First, the use of a ranking classifier emphasizes that most practical classifiers are flexible tools that can reflect a range of cost and class distributions. Ranking classifiers allow explicit tradeoffs in the types of errors incurred, rather than allowing only a single tradeoff implicitly encoded within a binary classifier. Second, it is possible to combine several ranking classifiers into a new, hybrid classifier that outperforms any of the base classifiers. This approach is known as the ROC convex hull (ROCCH) [16]. Finally, the use of a ranking classifier, when combined with more realistic assumptions about the data and inference techniques, makes it possible to reduce the number of false positives and to reduce data requirements of systems for information awareness. This final point will be discussed in more detail below.

3.3 Single-Pass vs. Multi-Pass Inference

A final assumption of many technical critiques of information awareness systems is that a single classifier is applied once to all instances. We characterize this as “single-pass” inference. If an instance is misclassified in this single pass, it cannot be corrected.

At least two alternatives exist to single-pass inference. In the first, a system could use predictions of one pass to inform the predictions made by a subsequent pass. Various types of multi-pass algorithms have shown good results when applied to relational data [2,11,22].

In the second alternative, a system could access different amounts or types of data on each pass. For example, assume that an initial prediction could be made based on relatively innocuous data (i.e., data which are not considered highly sensitive). Then those initial results could be used to limit the types of more sensitive data examined in subsequent passes, either resulting in fewer data items being accessed per object or in the same data items being accessed for fewer individuals. This approach is a part of several advanced systems for information awareness [19,20]. Below we examine how multi-pass inference can be used to achieve higher accuracy with lower data utilization rates than single-pass inference.

4. ENHANCED MODEL

Given the alternative elements mentioned in the previous section, can we do better? Specifically, can we design an enhanced model of an information awareness system, and can that model provide new technical understanding of the potential capabilities and weaknesses of such a system? In this section, we consider one candidate for an enhanced conceptual model. We emphasize that the model is still quite naive. Many of its assumptions are almost certainly not justified in practice. The model is intended to demonstrate an alternative to the widely used naive model and to illustrate the range of design alternatives. An enormous amount of additional work remains if we are to fully understand the potential design space for information awareness systems. Still, we hope that the enhanced model will lead to more informed debate over the effectiveness and impacts of such systems.

Table 2 summarizes the parameters of the model, which are introduced and explained below.

Table 2: Parameters of the enhanced model

Parameter	Meaning
N	Number of clusters per data set
n	Number of entities per cluster
p	Probability of a positive cluster
d	Difference between means of the score distributions of positive and negative entities, in units of standard deviations
r	Number of relations per entity
h	Probability that a given relation will terminate within the originating cluster (homophily)

4.1 Modeling Relational Data

As with the naive model, we assume a large population of entities, where each entity has a true class label. However, in contrast to the naive model, those entities are joined by relations. The relations are relatively sparse, representing only an extremely small fraction of all possible relations among the entities. Relations might model communications between entities, or financial transactions, joint ownership of some asset, or joint residence at some location.

Both the generation of true class labels and generation of the relational structure of data are controlled by an underlying clustering of entities. Each entity is assumed to be a member of one of *N* clusters, where each cluster has a fixed size *n*. Clusters might model families, social groups, or terrorist cells. All members of a given cluster are assigned the same true class label. Clusters are only used to generate data, and cluster membership is hidden from all classifiers.

For each entity, a given number of relations *r* is generated based on *h*, the probability that any given relation originating at an entity e_1 will terminate in another member of e_1's cluster. Thus, the probability that a given relation will terminate outside the originating cluster is *1-h*. The entity where a given link will terminate is selected randomly, given a population of candidate entities (i.e., within-cluster entities or outside-cluster entities).

This clustered relational structure is similar to the "small world" structure that has been observed in a wide variety of contexts, including social networks [25]. Given the homogeneity of class labels among cluster members, the probability *h* corresponds to the "homophily" of entities — the tendency of entities to connect to entities with the same class label. Low *h* produces many relations terminating outside the originating cluster; high *h* keeps relations within a cluster, and thus connects objects of the same class. Homophily has been observed in a wide variety of contexts, including social and professional acquaintance, scientific citations, and web page links [9,24].

Our experiments use relatively small clusters ($2 \leq n \leq 10$). Unless otherwise noted, *n*=5. Cluster size for the 9/11 hijacking cells was five (four, in one case, although there is evidence that the original plan called for a fifth member). In addition, we assume that there is no tendency for out-of-cluster links from positive clusters to go to other positive clusters. This latter assumption is probably unrealistic, but it would only reinforce results show below.

4.2 Modeling Ranking Classifiers

A ranking classifier is simulated by drawing scores randomly from one of two normal distributions. Scores for negative entities are drawn from one distribution, with μ=0 and σ=1. Scores for positive entities are drawn from the other distribution, with μ=*d* and σ=1. We vary *d* to simulate ranking classifiers of differing quality.

Again, this model is intended to be illustrative, not a valid reflection of actual systems. It is intended to show how the naive model is flawed, to suggest additional critiques with more technical validity, and to suggest future research. It is not intended to provide a realistic assessment of the characteristics of implemented systems or to prove that a proposed information awareness system could be effective.

4.3 Modeling Multi-Pass Inference

Given the graph structure described in Section 4.1 and the first-round classifier described in Section 4.2 that associates a score with each entity, we can apply a second-round classifier that averages an object's first-round score and the scores of all its neighbors. That is:

$$s'(e_i) = \frac{\sum\limits_{e_j \in E_i} s(e_j)}{|E_i|}$$

Where score $s(e_i)$ is the first-round score for entity e_i, score $s'(e_i)$ is the second-round score for the entity, and E_i is the set containing e_i and all of its neighbors.

This second-round classifier thus "smooths" the score estimates for each entity, in a similar way to ensemble classifiers and smoothing probability estimators. This procedure for producing second-round scores provides a new ranking classifier for entities.

5. REDUCING FALSE POSITIVES

Given relational data, a ranking classifier, and multi-pass inference, our experiments indicate that a system can be configured that greatly reduces the number of false positives while retaining the number of true positives. The opportunity to reduce false positives comes from the combination of all three elements of the enhanced model. Given moderate or high homophily, any entity is likely to be linked to other entities with identical true class labels. This comes partially from an entity's outgoing links, but also from its incoming links (links that terminate in the given object), because these are likely to originate from members of its own cluster. Negative entities with unusually high scores are likely to be surrounded by other negatives (with low scores, on average), and positive entities with unusually low scores are likely to be surrounded by other positives (with high scores, on average). Thus, each second round score *s'* is likely to be a more reliable indicator of the true class of the entity than the first-round score *s*.

As shown in Figure 3, the second-round classifier provides a substantial improvement in accuracy. These results assume data with 1000 clusters (N=1000), five entities per cluster (n=5), a 50% chance that a relation originating within a cluster will terminate outside of the cluster (h=0.5), four relations originating at each entity (r=4), a distance between score distributions of 1.5 (d=1.5), and a 2% probability that any given cluster has positive entities (p=0.02).

The ROC curve for the first-round classifier is substantially better than random, but still produces a large number of errors. As we know from the critiques above, many of these errors will be false positives under at least some cost and class distributions. The second-round classifier dominates the first-round classifier, providing uniformly superior performance at any cost and class distribution.

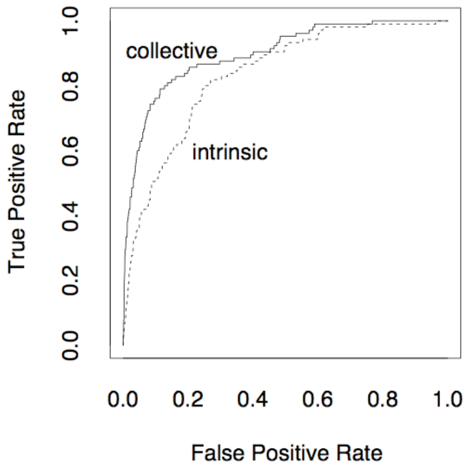


Figure 3: ROC curves for first-pass (intrinsic) and second-pass (collective) classifiers

Are the results robust over a range of parameter values? Figures 4, 5, and 6 show the performance of the second-round classifier as parameters of the data and the first-round classifier vary. In Figures 4 and 5 the cluster size *n* is fixed at 5. Cluster size varies in Figure 5. In Figures 4 and 5, each object has 5 outgoing links. Outgoing links per object varies in Figure 5.

All three graphs plot on the vertical axis the area under the ROC curve for the second-round classifier. The AUC is the average of 100 trials for each combination of parameter values. All three show cluster homophily *h* on the x-axis, ranging from 0.0 (no cluster structure) to 1.0 (all of an object's outgoing links connect to other cluster members). Figure 4 plots class separation *d* on the *y*-axis. The class separation is the distance (in units of standard deviations) between the means of the two normal distributions from which we simulate the first-round classifier's scores. Higher values of *d* simulate a more accurate first-round classifier because the scores assigned to positive objects are well separated from the scores assigned to negative objects. As we would anticipate, the AUC for the second-round classifier increases as the class separation (first-round classifier accuracy) increases. AUC also inc the cluster homophily increases, because with higher homophily each object is more likely to be connected to other objects from its cluster, which have the same true class label and similar first-round classifier scores.

Figure 5 shows the change in AUC as the size of clusters varies from 1 to 10. For any given trial all clusters have the same size, but the cluster size changes from one trial to the next. Class separation *d* is fixed at 1.5 for this figure, as well as Figure 6. Cluster size does not have much effect on AUC when cluster homophily is low, because the clusters are only loosely connected. Homophily does not have much effect on AUC when cluster size reaches its lower extreme of 1, since an object in a singleton cluster has no cluster partners to link with.

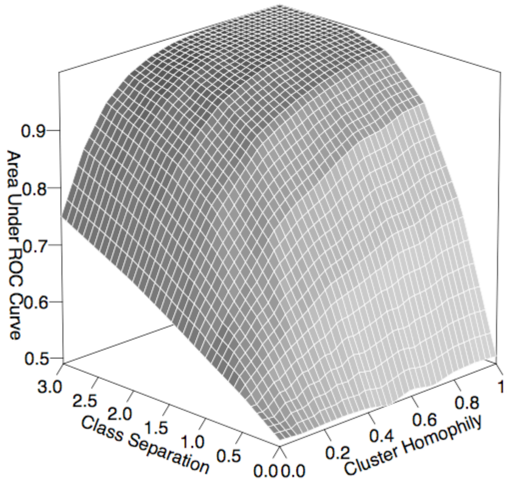


Figure 4: AUC of second-pass classifier varies with characteristics of the first-pass classifier and relational structure of the data

Figure 6 holds cluster size constant at 5, and varies the number of outgoing links each object has. At high homophily, AUC is high regardless of the number of links. At lower homophily values, the number of outgoing links per object becomes important because the clusters are not as well connected internally, so every link counts. The cluster size is fixed at 5; when the number of links per object drops below 5 the cluster cohesion is affected, and AUC drops off sharply as homophily decreases.

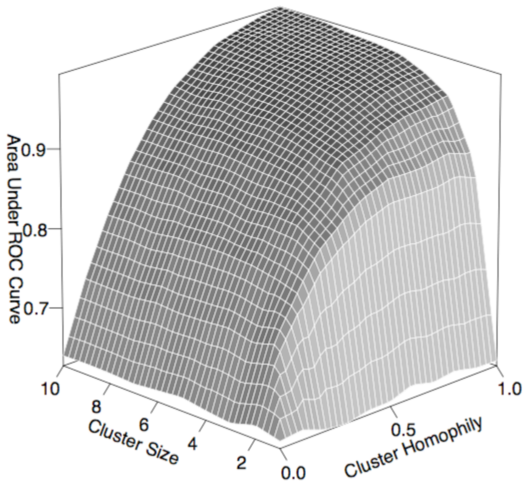


Figure 5: AUC of second-pass classifier varies with size and homophily of relations

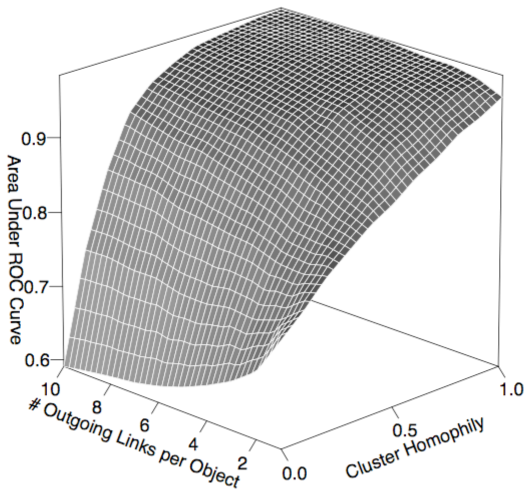


Figure 6: Second-pass classifier AUC varies with the density and homophily of relations

6. REDUCING INFORMATION REQUIREMENTS

Based on the results in Section 5, combining a ranking classifier, relational data, and a multi-pass classifier can substantially improve accuracy. Can these same elements be combined in a way that decreases the information collected about individuals?

Information requirements can be reduced in at least one way, if a decrease in performance is acceptable. We can apply a multistage approach to information gathering. In the first stage, we gather only *intrinsic* variables on all entities. With this information, we can apply the first-pass classifier to the entities. Given the ranking of entities produced by this classifier, we select a small subset of entities with the highest scores (i.e., those judged most likely to be positive). Then we gather additional *relational* data on this small subset. We designate that set of entities *A*. In the experiments reported below, this set consisted of 2% of the entire set of entities.

By gathering relations in which entities in *A* participate, we also pull in a set of other entities, because many of the relations connect entities in *A* to entities outside of *A*. We designate these other entities *B*. Finally, we gather all relations in which entities in *B* participate, pulling in a third set of entities, designated *C*. For entities in *A* and *B*, we have all intrinsic and all relation information. For entities in *C*, we have intrinsic information and only those relations they share with *B*. There remain a potentially large set of entities on which no relational information has been gathered. Based on their intrinsic information, and the fact that they are more than two links away from any entity with a high intrinsic score, they are excluded from further analysis. This process is similar to the process of following bibliographic citations or a criminal investigation involving financial transactions or telephone calls.

With the intrinsic and relational information on A and B (and some relational information on C), we can make revised inferences about entities in *A* and *B* (but not for entities in *C* or the remaining entities). Those entities retain the score they received from the first-pass classifier. Given this data collection process, and the revised scores, what is the performance of the system?

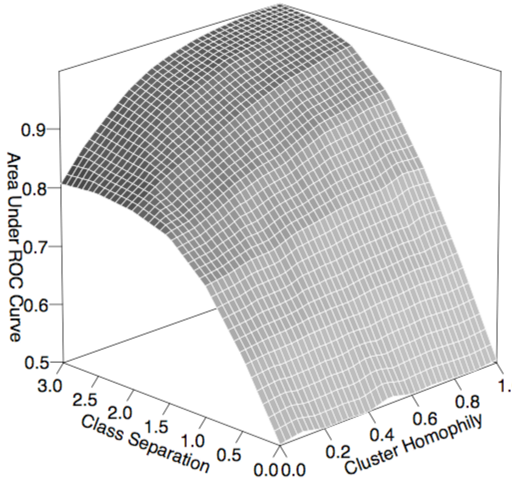


Figure 7: AUC of second-pass classifier with limited access to relations

Figure 7 shows how AUC changes with this scheme for limited access to relation information. Figure 7 is equivalent to Figure 4 in axes, though its AUC rises more slowly in the region of high homophily and large class separation.

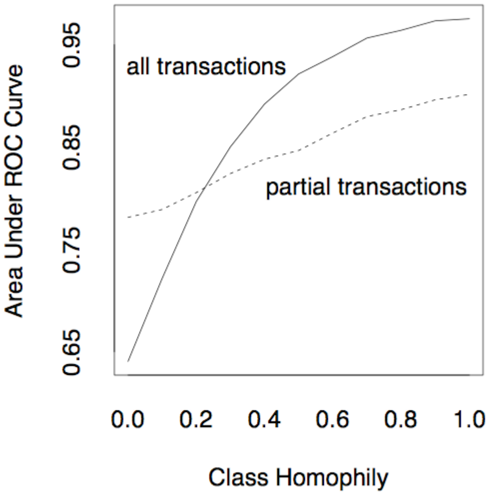


Figure 8: AUC for *d*=1.5 with limited access to relation information

Figure 8 shows a slice through Figure 4 and Figure 7 at a class separation of *d*=1.5. Clearly, overall classifier performance is lower. However, what have we gained for this lower performance? Figure 9 shows what is gained. Lines in the figure indicate the cumulative percentage of all objects in sets *A*, *A*+*B*, and *A*+*B*+*C*. If we consider only those entities for which we have all relations, they account for only about 10% of the total number of entities in the data sample. That is, this approach is able to achieve moderately high accuracy while only accessing the complete records of 10% of all entities.

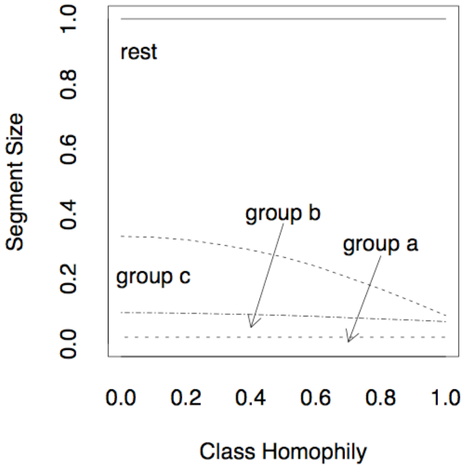


Figure 9: Percentage of all relations collected varies with class homophily

7. NEW CRITIQUES

The results in previous sections suggest that critiques of information awareness systems that are based on the naive model are not valid for more realistic systems that assume relational data, ranking classifiers, and multi-pass inference. What new critiques are suggested by the enhanced model?

First, the enhanced model shows that the results of a first-pass classifier can be substantially improved *given* the availability of relations that produce moderate to high levels of homophily among the class labels of entities. Without such homophily among related entities, the multi-pass classifier has little added utility. Thus, the enhanced model suggests a standard that relational data must meet if the approach outlined here is to work effectively.

Second, the system described here has a relatively obvious flaw — positive entities are assumed to exist in clusters. Recent warnings from Federal law enforcement officials regarding “lone terrorists” provide a strong counter-example. The basic approach outlined here would provide almost no advantage over the naive model if positive entities do not interact at all with other positive entities.

Third, the analysis done here assumes that the errors made by the first-pass classifier are independent of the relational structure of the data. This assumption is almost certainly false for many specific cases, given that other characteristics besides class label are likely to be homogeneous among members of the same cluster. As a result, first-pass classifiers may be likely to make the same errors on multiple members of the same cluster. This could produce clusters where most or all members have a high probability of being positive, but all are false positives. For a practical system to use the concepts described here, it would have to employ variables in the first-pass classifier that are known to be independent of the relational structure of the data.

Finally, the existence and use of these relations must be resistant to adversarial conduct. The generation of relational records should be difficult to avoid, even after an individual becomes aware that those records are used by an information awareness system. Some types of relational data are clearly not resistant to adversarial conduct. For example, an individual terrorist could refrain from initiating or receiving email messages, telephone calls, or financial transactions with other terrorists. Alternatively, an individual could purposely attempt to reduce his or her homophily, intentionally generating records that constitute “noise” with the intent of hiding the existence of a particular cluster. They could also use false identities to reduce the apparent number of relations tied to a particular identity. This issue is of concern for almost any information awareness system, regardless of design.

8. ACKNOWLEDGMENTS

This research is supported by DARPA and NSF under contract numbers F30602-01-2-0566 and EIA9983215, respectively. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, NSF, or the U.S. Government.

9. REFERENCES

[1] Association for Computing Machinery, U.S. Public Policy Committee, Letter to the Senate Armed Services Committee, January 23, 2003.

[2] Chakrabarti, S., Dom, B., & Indyk, P. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD-98 International Conference on Management of Data*, ACM Press, New York, 307-318, 1998.

[3] Dzeroski, S., De Raedt, L., and Wrobel, S. (Eds). Papers of the Workshop on Multi-Relational Data Mining. The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2002.

[4] Ellison, L. Digital IDs can help prevent terrorism. The Wall Street Journal. October 8, 2001.

[5] Fawcett, T., and Provost, F. Adaptive Fraud Detection. Data Mining and Knowledge Discovery 1(3), 91-316, 1997.

[6] Getoor, L., and Jensen, D. (Eds). Learning Statistical Models from Relational Data: Papers from the AAAI 2000 Workshop, AAAI Press, Menlo Park CA, 2000.

[7] Goldberg, H., and Senator, T. Restructuring databases for knowledge discovery by consolidation and link formation. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD 1995)*, AAAI Press, Menlo Park CA, 136-141, 1995.

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

[9] Jensen, D. and Neville, J. Linkage and autocorrelation cause feature selection bias in relational learning. In Proceedings of the 19th International Conference on Machine Learning (ICML 2002), Morgan Kaufmann, 259-266.

[10] National Research Council, Committee to Review the Scientific Evidence on the Polygraph. The Polygraph and Lie Detection, National Academy Press, Washington DC, 2003.

[11] Neville, J., and Jensen, D. Iterative classification in relational data. Learning Statistical Models From Relational Data: Papers from the AAAI Workshop, AAAI Press, Menlo Park, CA, WS-00-06, 42-49, 2000.

[12] Office of Technology Assessment, U.S. Congress, Information Technologies for the Control of Money Laundering. OTA-ITC-630, U.S. Government Printing Office, Washington DC, September 1995.

[13] O'Harrow, R. Jr. U.S. hopes to check computers globally. Washington Post, November 12, 2002, A04.

[14] Provost, F., and Fawcett, T. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), AAAI Press, 43-48, 1997.

[15] Provost, F., Fawcett, T., and Kohavi, R. The case against accuracy estimation for comparing classifiers. Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98), 445-553.

[16] Provost, F. and Fawcett, T. Robust Classification for Imprecise Environments. Machine Learning 42, 203-231, 2001.

[17] Safire, W. You Are a Suspect. New York Times. November 14, 2002.

[18] Scientific American (editorial). Total information overload. Scientific American, March 2003, 12.

[19] Senator, T., Goldberg, H., Wooton, J., Cottini, A., Umar, A., Klinger, C., Llamas, W., Marrone, M., and Wong, R. The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions. Proceedings of the 7th Conference on Innovative Applications of AI (IAAI 1995).

[20] Senator, T., and Goldberg, H. Break detection systems. *Handbook of Data Mining and Knowledge Discovery*, W. Klösgen and J. Zytkow (eds.), Oxford University Press. 863-873, 2002.

[21] Sparrow, M. The application of network analysis to criminal intelligence: An assessment of the prospects. Social Networks 13, 251-274, 1991.

[22] Taskar, B., Abbeel, P., and Koller, D. Discriminative Probabilistic Models for Relational Data. Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI 2002), Edmonton Canada.

[23] Wassermann, S. and Faust, K. *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994.

[24] Watts, D., Dodds, P., and Newman, M. Identity and search in social networks. *Science* 296, 1302-1305, 2002.

[25] Watts, D. and Strogatz, S. Collective dynamics of 'small-world' networks. *Nature* 393, 440-42, 1998.