Guarding the Next Internet Frontier: Countering Denial of Information Attacks†

Mustaque Ahamad
Leo Mark
Wenke Lee
Edward Omicienski
Andre dos Santos
Ling Liu
Calton Pu

College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
1 (404) 894-2593

{mustaq, wenke, lingliu, leomark,
edwardo, calton,
andre}@cc.gatech.edu

ABSTRACT

As applications enabled by the Internet become information rich, ensuring access to quality information in the presence of potentially malicious entities will be a major challenge. Denial of information (DoI) attacks attempt to degrade the quality of information by deliberately introducing noise that appears to be useful information. The mere availability of information is insufficient if the user must find a needle in a haystack of noise that is created by an adversary to hide critical information. We focus on the characterization of information quality metrics that are relevant in the presence of DoI attacks. In particular, two complementary metrics are explored. Information regularity captures predictability in the patterns of information creation and access. The second metric, information quality trust, captures the known ability of an information source to meet the needs of its clients.

Categories and Subject Descriptors

D.4.6 [**Security and Protection**]: Information flow and access control.

General Terms

Algorithms, Management, Measurement, Security.

Keywords

Quality of information, countering information attacks.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
New Security Paradigms Workshop '02, September 23-26, 2002, Virginia Beach, Virginia.
Copyright 2002 ACM ISBN 1-58113-598-X/02/0009 ...$5.00

1. INTRODUCTION

Quality of Service (QoS) in networked systems is an active topic of research well recognized by the community. QoS is primarily concerned with end-to-end performance metrics (e.g., network bandwidth and latency). Denial of Service (DoS) attacks are the major malicious threats to network QoS [1]. DoS attacks aim to reduce the QoS available to legitimate users by saturating some system resource through a flood of syntactically correct requests. DoS is recognized as a major threat as well as an active topic of research. While we recognize the importance of QoS and DoS research, we are looking at the *next* research challenge that comes after QoS/DoS.

Many Internet applications, e.g., digital libraries and electronic commerce, are built around information flows. Their main goal is to transport the right information to the right user at the right time. From school children to experts who manage critical national scale systems, an increasing number of information consumers are depending on information content that is relevant, accurate and satisfactory in serving the request. We believe that providing *Quality of Information* (QoI) in large networked information flow applications is a research challenge that immediately follows the QoS research. In analogy to the many dimensions of QoS, there are also many dimensions of QoI, such as the consistency, timeliness, reliability, trustworthiness, and density/richness of information. In this position paper, we focus on DoI attacks and our initial approach for countering such attacks.

A fundamental assumption made by many information rich applications is the ability of the system to find and deliver information with satisfactory QoI when such information is needed. This assumption is vulnerable to the intentional introduction of noise in the information system to confuse (or lower the efficiency of) the mechanisms for finding resources and information. We call this noise introduction *denial of information (DoI)* attack, which is the information analog of denial of service (DoS) attacks. Similar to denial of service, which floods a particular service with massive syntactically correct requests, denial of information floods resource discovery and information services by diluting their content with massive syntactically correct noise data. A concrete example of DoI attack is the "mail bomb", which reduces email information content. Consequently, even when critical information is accessible in principle, it may be difficult or impossible to find it, resulting in lower or unacceptable level of QoI. Unlike QoS attacks which succeed by saturating the

† This work was supported in part by NSF ITR grant 0121643.

resources of a computer system, QoI attacks target the inability of humans to handle information beyond a certain limit. Thus, QoI attacks can succeed with lot less computational resources compared to QoS attacks.

Although DoI attacks are the information analog of DoS attacks, there are some striking differences between them. For example, DoS attacks must be massive attacks that overwhelm the server capacity, thus degrading the service level. In contrast, DoI attacks may be either massive or gradual. A massive DoI attack generates a large amount of spurious information, hoping that the information system will transmit a portion of it to the consumer and hamper the consumer's ability to discern the actual picture of reality, at least momentarily. A massive DoI attack is usually unconcerned with detection and is designed to produce immediate confusion before recovery can take place. In contrast, a gradual DoI attack attempts to remain below the detection threshold and continues to disseminate small amounts of spurious information surreptitiously. Over a period of time, a gradual DoI attack may be able to trick the information system to deliver increasingly misleading information. Although DoI attacks may be used as a defense mechanism (e.g., music industry floods music sharing systems with corrupted files), we consider attacks that attempt to degrade information quality for legitimate users and address how they can be countered.

We claim that QoI assurance is an important security requirement that needs to be addressed in the near future. Traditional information security deals with confidentiality, integrity and availability of information and neither of these capture QoI. For example, integrity typically means that information is not altered by unauthorized parties. Integrity captures QoI only if information is accessed from known and authenticated sources. In the case when applications must discover relevant information, spurious sources can easily create noise that can successfully obscure information produced by legitimate sources. Such attacks can succeed in diminishing the availability of good quality information. DoI attacks may be more damaging in the long term because they can succeed even with a perfect defense against DoS. Having all the resources to process requests (successful defense against DoS) still does not prevent spurious or maliciously false information from being injected into the system.

Our hypothesis is that the mere accessibility of critical information becomes insufficient if the user must find the needle in a haystack of noise introduced by DoI attacks that effectively hide the needed information. In an NSF ITR program funded project at Georgia Tech, we have begun to explore how to meet QoI needs in the presence of threats from DoI attacks. In this position paper, we motivate the need for DoI research and present our initial research approach for countering them. Section 2 presents several motivating examples. Section 3 outlines our initial approach. We discuss related work in Section 4 and conclude the position paper in Section 5.

2. DoI Examples

DoI attacks aim to create bogus information that appears to be legitimate. They can also attempt to create noise that effectively hides useful information. Several attacks that have been described in the literature are examples of DoI attacks. We use such examples to illustrate the harm that can be done by DoI attacks. Our research aims to develop techniques that can be used to counter a broad set of attacks, and its focus is not on a specific example that is presented here.

Example of Massive Attack. A good example of active DoI attacks is the so-called "mail bomb". On August 10, 1996, Mr. Dave Methvin [7], an executive editor of Windows magazine at that time, was one of more than a dozen persons who suffered a typical mail bomb attack. The list of persons attacked included former President Bill Clinton and Bill Gates. The attacker subscribed the victims to over 1000 mail lists. Mr. Methvin is a good example of how a massive DoI attack can impair someone from getting information, in this case e-mail. Just 10 hours after the attack started, he was flooded with 1600 mail messages. There are many variations of this attack that can seriously affect a user's information reception capability in critical situations.

Example of Passive Attack. An interesting historical example of a passive DoI attack was the early Red Herring web page [4], which simply threw back a dictionary when it sensed a probe by a web robot such as WebCrawler and Lycos. Although this was designed as an offensive reaction against web robot probing, it shows the power of DoI attacks against naïve web robots. Early robots would include the Red Herring URL in every search involving a combination of single keywords. The author of the Red Herring home page voluntarily withdrew the page after making his point, but it shows the potential threat.

Example of Gradual Attack. A clear trend in the web robot (a.k.a. search engine) efficiency is the downward spiral of the signal-to-noise ratio of search results. Granted, a fundamental contributing factor is the natural expansion of the documents accessible through the web. At the same time, we see an aggressive insertion of many keywords in HTML headers to increase the chance of the page being listed by more search results. We see this effort, which could be called "search engine spoofing", as a gradual version of the Red Herring attack, or alternatively, a passive version of spam. Search engine designers try to improve their heuristics to filter out the noise, but it is a race against the spammers, since it is a matter of time before they figure out the new heuristics and tailor their pages accordingly.

Likelihood of DoI Attacks. We believe that the probability of seeing effective DoI attacks is increasing. On the producer side, necessary conditions for mounting an effective DoI attack probably include fast CPU, large memory as well as disk, and high network bandwidth. All are needed in the creation of a large amount of noise. Given the continuous decline in computer prices, the potential attacker can generate and disseminate a much larger amount of noise than previously possible. On the consumer side, more and more systems are evolving towards an increasing number of dynamic links to sources of fresh information (e.g., sensors). This makes the ad hoc defense against DoI attacks more difficult and expensive.

3. Our Research Approach

There are many metrics that can be used to judge the quality of information [10]. For example, relevance, accuracy, consistency and timeliness are some metrics that are frequently used. In open environments such as the Internet, search engines attempt to locate information that is highly relevant. Such relevance is based on the match of a user's query to the content of a document that is available at a certain source. Because of the unprecedented growth in content, more sophisticated techniques are used to narrow the sources to

only those that are viewed to provide the highest quality information. For example, the Google search engine takes into account the number of other documents that have links to a given document in determining the relevance ranking of the document's content [8]. For dynamic content that changes frequently, consistency becomes an important quality metric. Techniques such as time-to-live (TTL) fields are employed in the Web to ensure that the likelihood of applications accessing stale information is low.

The systems that currently exist primarily focus on techniques that meet data quality needs such as relevance and consistency. Although these techniques are extremely useful, their effectiveness can be severely impacted by DoI attacks. For example, a malicious client can create a document D that appears to contain highly relevant information for a topic that is of interest to large number of people. To make the document appear even more relevant, the attacker can create other documents that have links to D. The information in D and other such bogus documents can make it difficult for applications to locate legitimate sources of information even when they exist because they get masked by the noise that is created by the DoI attack. Thus, we claim that these QoI metrics need to be augmented with new ones that are particularly relevant in the presence of DoI attacks.

DoI attacks can interfere with access to high quality information by manipulating attributes related to the content of the data objects and their access patterns. Examples of such attributes are the states of data objects, the nature of their changes, and the rates of change. Furthermore, we expect that relationships across such attributes and their rates of change will also be important in countering DoI attacks. Information regularity, the first QoI measure proposed by us, captures normal patterns in such attributes and is used to construct models that can detect anomalies in creation, update or access rates of certain data objects. Thus, it can be used to alert a user of a potential DoI attack. When such an alarm is sounded, information consumers could rely on the history of their past interactions with the sources from where information comes. QoI-trust, the second QoI metric that we define, captures past interactions with information sources.

3.1 QoI Example: Information Regularity

In order to detect and counter DoI attacks, we need to define appropriate QoI metrics that can help distinguish normal and attack situations. We assume that we cannot know all possible DoI attacks a priori and hence will focus on *anomaly detection*, which uses models of normal behavior to detect deviations (anomalies) as possible attacks. The basic premise for anomaly detection is that there is intrinsic and observable characteristic (or regularity) of normal behavior that is distinct from that of abnormal behavior. The task of developing an anomaly detector therefore involves first studying the normal characteristics and then building a model that best utilizes the characteristic. We propose to use a set of information-theoretic based regularity measures, namely, entropy [29] and conditional entropy [27], for detecting DoI attacks. The reasons for using these generic measures are two folds. First, there are many different kinds of information flow applications that may require different kinds of (specific) regularity measures. We need to start with these general measures and study how to use them as the building blocks for application-specific regularity measures. Second, we as well as other researchers have begun using these measures to build and evaluate anomaly detection models (in intrusion detection domain) and have obtained very encouraging results [17,21]. We can build on our past experience to meet the research challenges here.

Entropy, or Shannon-Wiener Index, is defined as $H(X) = -\sum_x P(x)\log P(x)$ where $P(x)$ is the probability of record x in dataset X. The smaller the entropy, the fewer the number of different records (i.e., the higher the redundancies), and we say that the more regular the audit dataset. High-regularity data contains redundancies that help predict the future because the fact that many observations are repeated (or redundant) in the current dataset suggests that they are likely to appear in the future. In other words, for anomaly detection, we need data with low entropy. Because of the temporal nature of information flows, we need to characterize the regularity of sequential data. *Conditional entropy* is defined as $H(X \mid Y) = -\sum_{x,y} P(x,y)\log p(x \mid y)$ where $P(x, y)$ is the joint probability of x and y and $p(x|y)$ is the conditional probability of x given y. If y *is* preceded by x in a sequence, then the lower the conditional entropy, with more certainty we can determine x after we have seen y. We can use classifiers as anomaly detection models. For example, we can use a classifier, trained using normal data, to predict the (normal) next event based on the previous n events. When the prediction is not same as the actual event, there is an anomaly. Given a record described by a set of features (e.g., the names of the n previous events), a classifier determines the class label of the record (e.g., the name of the next event). When constructing a classifier, the algorithm searches for features with high *information gain* (or reduction in entropy) [28], defined as $Gain(X, A) = H(X) - \sum_v |X_v|/|X| H(X_v)$, where X_v is the subset with attribute A having value v. That is, a classifier needs feature value tests to partition the original dataset (with mixed classes and hence high entropy) into pure subsets (each ideally with one class and hence low entropy). Therefore, there is a direct connection between entropy measures and classification accuracy. For example, we can show that if we "collapse" the n previous events into a single feature when building a classifier to predict the next event, then the second term in the above *Gain* formula is essentially the conditional entropy of the next event given the previous n events. The lower the conditional entropy, the higher the Gain, and the more accurate the classifier can predict the next event.

In using these information-theoretic based regularity measures to detect DoI attacks, we first assume that attributes of an information source such as number of objects, their updates and access rates can be defined and logged. We can then select (or partition) log data so that the normal dataset has entropy (or conditional entropy) as low as possible, and then perform appropriate data transformation according to the entropy measures (e.g., constructing new features with high information gain) and apply classification algorithm to learn a classifier. For example, suppose we want to predict the next information delivery based on the previous n. We first compute the conditional entropy measures using various values of n, and then select the one, say n_0, which has the

lowest entropy as the "sequence length". Then we transform the data where the features are the first n_0 deliveries and the class labels are the next delivery. The classifier essentially describes what is normally the next delivery given the first n_0. In real-time monitoring, if there are a significant number of deliveries that do not agree with the classifier, e.g., due to noises being injected in the information flow, then we can raise an alarm. An alarm signifies that attributes of the information source or the documents supplied by it have changed. Although this may be due to degradation of QoI, in some application a user may only be interested in knowing when a change occurs in the regularity metric of QoI.

We have begun a preliminary study of using these information-theoretic measures for intrusion detection and have obtained very encouraging results [17]. The domain here is different and poses new challenges. First, DoS attacks are "massive" and can normally be observed in lower layers (e.g., TCP/IP packets or operating system events) whereas DoI attacks can also be gradual and may only be observable at the application layer. For example, the DoI attacker can inject a lot of noise while keeping the network traffic volume constant. Such massive DoI attacks can be detected using our framework (as discussed in the example above). Gradual attacks may require data transformation, e.g., sorting the data using different keys (rather than time-stamps). The question is whether to build static (a priori) models, i.e., with several models according to different transformations, or to build a dynamic model that can pick up very small evidence and perform a series of transformations and modeling according to some heuristics. Second, while the detection algorithms are application-independent, the log formats or data schema are necessarily application-specific. We need to design our algorithms to use schema (e.g., what are the key fields of the data) as parameters. Third, QoI measures are also application-specific while the information-theoretic measures are generic. We will study how to link these two kinds of measures together. Our conjecture is that if the application-specific audit data schema can represent what QoI measures are important (e.g., timeliness of information delivery), then the information-theoretic measures of the audit data basically describe the normal characteristics of QoI measures (e.g., how the timeliness of relevant information deliveries fluctuates).

3.2 QoI Example: Trust

An interesting area of research is the measure of reliability, authority, or trustworthiness of information sources. This is an important problem that has received considerable attention in the Web community, and remains an active area of research with many open problems. Zagat Survey, for example, is considered an authoritative source on the quality of restaurants due to the variety of information sources included in the survey. Popular web sites such as CNET, for example, use similar methods in their classification of computer equipment. Our goal is to explore novel methods that can be used to rank information sources by QoI metrics that are appropriate in the presence of malicious behavior.

Although the notion of trust has been used widely in many contexts that range from network authentication [23,26] to e-commerce [31,32], there is no universal definition that is appropriate for all domains. QoI-trust, which is a notion of trust we define as a QoI metric, captures the assurance that information contained in certain objects meets the requirements of an application. For example, if an e-commerce application purchases certain goods on the web based on information made available by vendors, QoI-trust reflects the ability of an online vendor to supply the merchandise of agreed upon quality at the advertised price in a timely manner. Thus, trust in this context is similar to seller reputations of Ebay. In another setting, if a scientist is searching for information relevant to an experiment that she is conducting, QoI-trust captures the belief that the data found in an article is scientifically valid. Such trust reflects the level of rigor of the reviewing and editing process of the source where the article appears. Clearly, a dishonest or compromised source, or an imposter can attempt to mount DoI attacks by creating noise that appears as useful information. The goal of the QoI system is to filter out the noise by associating very low values of QoI-trust with it, while maintaining high level of trust with information objects that are produced by reliable sources. Although trust has been used in many contexts, our goal is to explore QoI-trust which captures the quality of information in the presence of QoI attacks.

The QoI-trust metric has a value between 0 and 1. A QoI-trust value close to 0 means that either the information is known to be useless or little is known about its source and its ability to meet a consumer's need. On the other hand, a QoI-trust value close to 1 indicates that the information will meet the needs of the consumer with a very high likelihood. Although we choose a range from 0 to 1, other values of QoI-trust are also possible. For example, a value of –1 can be used to denote complete distrust in a source because it is known to supply misleading or false information. Clearly, one challenge that must be addressed is how QoI-trust values can be associated with information objects and sources. If channels over which information will be communicated can tamper with the information then trust values must be associated with both the sources of information as well as the channels over which it is sent.

Many models are possible for building an infrastructure that allows QoI-trust values to be associated with information. On one extreme are architectures where QoI-trust is obtained from a small number of fully trusted authorities or trust managers. For example, consider the financial domain where credit worthiness ratings are associated with creditors. A small number of companies (e.g., Equifax) maintain credit history of consumers and rate consumers based on it. Creditors use such ratings, which are akin to the QoI-trust metric we have discussed. In the QoI domain, certain trusted authorities could certify information made available by sources. For example, the Wall Street Journal Site www.wsj.com could be deemed as a trusted source of information related to business news. The information must be digitally signed or transported over secure channels to ensure that it comes from the claimed sources. In this approach, the burden of getting a QoI-trust value associated with its information is on the sources that create the information.

Although an approach based on trust authorities or identities of information sources can be useful, QoI needs across different consumers can vary significantly and the QoI-trust metric values associated with certain information by the authority may not match the needs of all users. The scalability of such a trust architecture in the presence of massive amount of information and sources is also a problem. These problems are already evident from the lack of widely deployed public key infrastructures for service and user authentication in the Internet. Similarly, identity based QoI-trust suffers from

problems. Apart from the problem of identity establishment, QoI-trust values may depend on the nature of the information. For example, legal information provided by a lawyer may be trusted but medical information from the same source may have limited value. We need to explore a decentralized architecture, where QoI-trust values are based on observed quality of information.

Our architecture is based on the general notion of observers that build and provide QoI-trust values for various information sources based on attributes of information. Observers could be information consumers or intermediaries that are in close vicinity of the consumers. We motivate the decentralized architecture based on the following simple scenario. Consider a scheme in which a browser pops up a question that asks the user if the displayed information meets her needs. If the user clicks a yes, this represents a positive experience. On the other hand, a no represents a negative experience. A QoI-trust metric can be derived based on positive and negative experiences [25]. Although the recording of such experiences my be burdensome or too expensive [31], we believe that this is necessary because QoI is fundamentally a semantic property and must be built based on feedback from consumers of the information.

Although actual experiences of consumers who access information are highly valuable, this kind of approach for building QoI-trust suffers from several problems. First, in an open system where sources of information change continuously, a user may not have had any direct experiences with a source to associate a QoI-trust value with an information object supplied by it. Second, such an approach relies on user's willingness to provide accurate information about his or her experiences. We address these problems by developing a decentralized architecture for QoI-trust as follows.

We call the observers that monitor user experiences trust authorities (TA) because they build QoI-trust for various sources of information. The TAs, either periodically or on access by a consumer, log the source of information S, salient attributes of the information, A, and the time T at which the access is completed. In addition to logging such (S,A,T) tuples, a TA also has a component that derives a QoI-trust value for a given source of information based on logged tuples. Such trust value computation can be based on correlation of S's information with similar information that can be acquired from other sources by the TA. The TA may also be able to validate the information based on its observations or user experiences. For example, the quality of weather forecast can be ascertained based on observed weather conditions. The correlation will result in an increase or decrease in the QoI-trust level value of the source S for information that has attributes A. Furthermore, TAs can coordinate with each other and compose or combine QoI-trust values that are computed by them for common information for which tuples are logged by each TA.

A consumer can request information from multiple sources. The consumer will maintain another list of tuples that include the TA, information attributes and the QoI-trust as returned by the TA. The QoI-trust values received from different TAs can also be composed by the consumer similar to a TA composing QoI-trust based on interactions with other TAs. Specifically, periodically, the consumer will correlate information provided by a given source with its own experiences or from sources that have information with similar attributes, weighted by their QoI-trust levels. Such correlation can result in a dynamic change in the level of QoI-trust that the consumer assigns to a TA. A consumer can implement its own TA or a TA can provide QoI-trust values for information accessed by consumers in a neighborhood.

There are several issues that are being investigated in this research. The logging of QoI relevant attributes, derivation of QoI-trust based on tuples logged by the TA or consumers, and composition of QoI-trust from multiple TAs are some of the problems that are being explored by us. Another issue is the potential vulnerabilities and threats that exist in this architecture. These include monotonically increasing the QoI-trust level associated with an information source or TA through training or impersonation. A simple approach against training is periodic resetting of trust levels, which is performed by resetting the QoI-trust level to its initial value. Another benefit of resetting is the aging of experiences from which QoI-trust levels are derived. A solution to impersonation is to limit the maximum QoI-trust levels that can be associated with sources or TAs. These levels can be determined based on claimed or known identity of the sources or the availability of secure communication channels with them. Similarly, lower initial QoI-trust can be associated with sources or TAs that are vulnerable to compromise.

The QoI-trust architecture can be used to build filters for information sources that will be personalized for each consumer, based on the consumer's ability to handle potentially unreliable information. These filters will have to consider several characteristics, including the risks of accepting incorrect information or receiving no information at all. It will be up to the information consumer to set the filter, based on QoI-trust levels, to the threshold that best matches the risk and benefit levels acceptable to the consumer.

3.3 Combining Information Regularity and Trust

The two QoI metrics, information regularity and QoI-trust, together provide a powerful mechanism for accessing information that is of high quality. Information regularity, which captures the normal and regular patterns in the way information evolves, can be used as the basis for building detection models that can alert the system about a potential DoI attack. On the other hand, accessing information based on QoI-trust levels filters out sources that are suspected to be untrustworthy. Thus, if there is suddenly a significant increase in information irregularity, it can be the case that the system is under a massive DoI attack. In this case, consumers can raise the trust threshold to filter out potentially compromised sources of information. On the other hand, gradual attacks may not introduce a significant amount of anomaly within the relatively short time window for which anomaly detectors examine the logs. However, in this case, over time, the QoI-trust associated with compromised or malicious sources will degrade as the quality of information from such sources becomes poor. Such sources will be excluded by the consumers based on low QoI-trust values. Also, for given attributes of information, information regularity measures can be computed and anomaly detectors constructed for information that is available at all sources vs. only those sources that have QoI-trust levels above a certain threshold. If the anomaly detector signals no alarms for information that comes from sources with and without the trust filter, it is likely that the system is not under attack. If there are anomalies detected on data from sources without the QoI-trust

filter but there are no anomalies detected for sources that have certain QoI-trust levels, the system may be under attack but the trust filter is working properly. However, if the indication that the consumer is potentially under a DoI attack is strong enough, this fact should be logged. If anomalies are detected for sources with as well as without QoI-trust filters, this is an indication that the QoI-trust level assigned by certain TAs is wrong and should be revised. The revision can be performed automatically through the composing of QoI-trust with other TAs or by resetting of some or all QoI-trust levels, possibly with the intervention of an expert entity. The expert entity can be either a human being or a highly trusted information source to which the consumer has a secure channel. We will investigate how the mechanisms that are used to compute the two QoI metrics can be coordinated to obtain better values for them.

4. Comparison with Related Work

Secure and Survivable Systems. Traditional work on database security (see [5] for many pointers to work in the area) has focused on the protection of information within a database, usually from outside attackers. In contrast, DoI attacks primarily influence query results by adding new information to an open information system. Denial-of-service attacks have been well publicized. They differ from DoI attacks as they focus on the means (the server) rather than on the information. Access control models in information security such as multi-level security or information flow models primarily focus on confidentiality and to a limited extent on integrity. They do not address quality of information and dealing with noise and misinformation.

Managing Distributed Data. Digital libraries is one of the most active areas of data management research. It differs from our research because digital libraries assume that information is primarily historical and slow changing. The area of real-time decision support has been primarily of interest to applications classified under the electronic commerce area, for example, logistics and supply management. Unfortunately, existing solutions primarily work under the assumption of closed supply chains, where participants are well known and authenticated beforehand. The series of International Conference on Information and Knowledge Management (CIKM) [2] have provided a forum for discussing common research interests in databases and information retrieval. However, despite some development of methods and software to find information from both structured (closed) and unstructured (open) sources [6], there has been little concrete progress on security and survivability of such systems. There is also a series of conferences on information quality [3] sponsored by MIT's Sloan School. Some progress can be seen in the data quality and information quality areas, but there is little work on denial of information.

Delivery of Fresh Information: Several important emerging classes of distributed applications are inherently information-driven. Instead of occasionally dispatching remote computations, such information-driven systems tend to transfer and process streams of information continuously. Member of this class range from applications that primarily transfer information over the wires such as digital libraries, teleconferencing and video on demand, to applications that require information-intensive processing and manipulation, such as distributed multimedia, Web search and cache engines. Other applications such as electronic commerce combine heavy-duty information processing (e.g., during the discovery and shopping phase, querying a large amount of data from a variety of data sources) with occasional remote computation (e.g., buying and updating credit card accounts as well as inventory databases).

In the Infosphere project, we are particularly interested in *fresh* information that changes the way we interact with our environment. We envision the fresh information providing more details about the current state of our physical world than ever imaginable, impartially to all human beings. Our past work that contributes to this vision has focused on extraction of information, its transport with quality-of-service guarantees and monitoring of updates to information that is of interest to consumers in a large scale distributed environment like the Internet [30].

Decentralized Trust Architectures: Although QoI-trust is an information quality measure, there is work in the area of trust models that is relevant to our research. Several models of trust and their applications are surveyed in [23]. In the security area, past trust models have primarily been developed for authenticating users in distributed environments. For example, Reiter et. al. [24] explore metrics that are well suited when authentication is based on information provided by multiple sources (e.g., certification authorities). Other techniques for combining trust are presented in [25]. The PolicyMaker system makes use of a decentralized trust architecture to define an authorization framework [26]. Other trust based authorization systems also exist and trust composition is used in systems like PGP. In e-commerce systems, clients must have trustworthy relations with vendors with whom they want to do business. Trust models, metrics and their effectiveness has been addressed in [31,32]. Trust models have also been explored in other contexts such as social sciences. Although we will benefit from such work, our focus is on notions of trust and its dynamic maintenance and composition when it is used as a QoI metric to counter DoI attacks.

Intrusion Detection: In detecting DoI attacks, we can draw lessons and experiences from intrusion detection, a closely related research area that has been active for near two decades. A critical step in building an intrusion detection system is to identify the *features*, i.e., the evidence extracted from the audit data that distinguishes normal and intrusive activities. For *misuse detection* systems, for example, IDIOT [13] and STAT [12], the features are components of the attack patterns (or "signatures") of known intrusions. For *anomaly detection* systems, for example, IDES [20] and EMERALD [22], the features are system activities measures that constitute the normal profiles. We have developed a systematic framework that first computes activity patterns from audit data, identifies consistent normal and unique intrusion patterns, constructs features to capture the meanings of these patterns, and then computes classifiers as intrusion detection models using data formatted according to the features [14,15,16]. Results from the 1998 DARPA Intrusion Detection Evaluation showed that the detection models produced by our framework had one of the best performance of all participating systems [19].

Intrusion detection is traditionally concerned with attacks that exhibit clear evidence in network packet data (e.g., *tcpdump*), operating system event data (e.g., *BSM* data), and low-level generic application data (e.g., system calls [11], C/C++ library calls [18], etc.) These data sources are well studied and abundant data processing tools are available. However, many

DoI attacks may have evidence manifested only in high-level and application-specific audit data. The research challenge is therefore to develop a general, rather than application-specific, reasoning framework for analyzing audit data and construct appropriate features and effective detection models. We propose to use information-theoretic based *information regularity* measures, i.e., entropy and conditional entropy, to study and express the intrinsic characters of normal data flow and to guide the process of building detection models. We as well as other researchers have begun preliminary studies of using information-theoretic measures to build and evaluate anomaly detections models (in intrusion detection domains) and have obtained very encouraging results [17,21].

5. Conclusions

The critical nature of the information infrastructure has been well documented by recent studies. We claim that as applications become information rich, in addition to QoS requirements, they will also depend on the timely availability of high quality information. In this position paper we motivate the need to provide QoI assurance in the face of denial-of-information (DoI) attacks. Examples of DoI attacks already exist and they will become more frequent as information rich applications are deployed over the Internet.

We present an initial approach for countering DoI attacks. This approach includes the use of mechanisms similar to intrusion detection to find anomalies in the behavior of information sources. We also define a notion of trust that can be associated with such sources to capture the usefulness of information that is supplied by them.

There are many problems that need to be defined precisely for countering DoI attacks. We need to characterize the attacks that can be mounted by malicious entities and need to relate and compare them with other attack models. At the same time, we need to discuss the feasibility of the QoI metrics that are proposed by us and need to identify other metrics. Finally, we need to address the problem of evaluating the effectiveness of defenses that are mounted against DoI attacks.

Acknowledgements

We would like to thank Mary Ellen Zurko and anonymous referees for their comments that were helpful in improving the paper.

6. References

[1] CERT, *TCP SYN Flooding and IP Spoofing Attacks,* CERT Advisory CA-1996-21, 9/19/96, available online at CERT http://www.cert.org/advisories/CA-1996-21.html.

[2] Conferences on Information and Knowledge Management (CIKM). *Proceedings of CIKM.* Pointers available from the conference central site http://www.cs.umbc.edu/cikm/.

[3] Conferences on Information Quality (IQ). *Proceedings of IQ.* Pointers are available from the conference central site http://web.mit.edu/tdqm/www/IQConference.html. Conferences held between 1996-2001.

[4] Dellinger, Joe. *Three risks of web robots.* The Risks Digest, vol. 17, issue 70, 2/8/96, available online at the Risks archives http://catless.ncl.ac.uk/Risks/17.70. The actual page was available at http://www.graviton.com/red/ but it seems to be missing as of November 2000.

[5] IFIP WG 11.3 on Database and Application Security. *Home Page.* Pointing to papers and conferences in Database Security. Currently at http://homes.dsi.unimi.it/~ifip113.

[6] Liu, Ling, Calton Pu, and Wei Tang . *Continual Queries for Internet Scale Event-Driven Information Delivery,* IEEE Transactions on Knowledge and Data Engineering, Special issue on Web Technologies, Vol. 11, No. 4, July/August 1999.

[7] Methvin, Dave. *Mailbomb Maelstrom on the Internet.* Windows magazine August 1996, available online at <http://www.winmag.com/people/dmethvin/mailbomb.htm>. Summary in email posting entitled: *Re: Denial of service ... Netcom listservers,* The Risks Digest, vol. 18, issue 39, August 1996, available on line at the Risks archives <http://catless.ncl.ac.uk/Risks/18.39.html#subj9.1>.

[8] Google Search Technology. http://www.google.com/technology/index.html.

[9] Spafford, Eugene. *The Internet Worm: Crisis and Aftermath.* Communications of the ACM, Volume 32, Number 6, pp. 678-688, 1989.

[10] Wand, Yair and Wang, Richard. Anchoring Data Quality Dimensions in Ontological Foundations, CACM, November 1996.

[11] Forrest, S., Hofmeyr, S. A., Somayaji, A., and Longstaff, T. A. *A sense of self for Unix processes.* Proceedings of the 1996 IEEE Symposium on Security and Privacy, 1996.

[12] Ilgun, K., Kemmerer, R. A., and Porras, P. A. *State transition analysis: A rule-based intrusion detection approach.* IEEE Transactions on Software Engineering, 21(3), 1995.

[13] Kumar, S. and Spafford, E. H. *A software architecture to support misuse intrusion detection.* Proceedings of the 18th National Information Security Conference, 1995.

[14] Lee, W., Stolfo, S. J., and Mok, K. W. *Mining Audit Data to Build Intrusion Detection Models.* Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD '98), 1998.

[15] Lee, W., Stolfo, S. J., and Mok, K. W. *A data mining framework for building intrusion detection models.* Proceedings of the 1999 IEEE Symposium on Security and Privacy, 1999.

[16] Lee, W. and Stolfo, S. J. *A Framework for Constructing Features and Models for Intrusion Detection Systems.* ACM Transactions on Information and System Security, 3(4), 2000.

[17] Lee, W. and Xiang D. Information-Theoretic Measures for Anomaly Detection. Proceedings of the 2001 IEEE Symposium on Security and Privacy, 2001.

[18] Lin Y. and Jones A. *Application Intrusion Signatures Using Library Calls*. Department of Computer Science, University of Virginia, 2001.

[19] Lippmann R., Fried D. J., Graf I., Haines J. W., Kendall K. R., McClung D., Weber D., Webster S. E., Wyschogrod D., Cunningham R. K., and Zissman M. A. *Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation.* Proceedings of the 2000 DARPA Information Survivability Conference and Exposition, 2000.

[20] Lunt, T., Tamaru, A., Gilham, F., Jagannathan, R., Neumann, P., Javitz, H., Valdes, A., and Garvey, T. *A real-time intrusion detection expert system ({IDES}) – final technical report.* Technical report (February), Computer Science Laboratory, SRI International, Menlo Park, California, 1992.

[21] Maxion, R. A. and Tan K. M. C. *Benchmarking anomaly-based detection systems.* Proceedings of the 1st International Conference on Dependable Systems and Networks, 2000.

[22] Porras, P. A. and Neumann, P. G. *EMERALD: Event monitoring enabling responses to anomalous live disturbances.* Proceedings of the National Information Systems Security Conference, 1997.

[23] Vipin Swarup and Thayer Fabrega ,Trust: Benefits, Models and Mechanisms, Lecture Notes in Computer Science: Security Issues for Mobile and Distributed Objects, 1999.

[24] M. Reiter and S. Stubblebine. Towards acceptable metrics of authentication. IEEE Symposium on Privacy and Security, 1997.

[25] T. Beth, R. Yahalom and B. Klein. Trust relationships in secure systems. IEEE Symposium on Security and Privacy, 1993.

[26] M. Blaze, J. Feigenbaum and J. Lacy. Decentralized trust management. IEEE Symposium on Security and Privacy, 1996.

[27] Cover, T. M. and Thomas, J. A. Elements of Information Theory. Wiley, 1991.

[28] Mitchell, T. Machine Learning. McGraw-Hill, 1997.

[29] Shannon, C. E. and Weaver, W. The Mathematical Theory of Communication. University of Illinois Press, 1949.

[30] Ling Liu, Calton Pu, Karsten Schwan and Jonathan Walpole, "InfoFilter: Supporting Quality of Service for Fresh Information Delivery", *New Generation Computing Journal* (Ohmsha, Ltd. and Springer-Verlag), Special issue on Advanced Multimedia Content Processing, Vol.18, No.4, August 2000.

[31] Jens Riegelsberger and Angela Sasse, "Trustbuilders and Trustbusters: The Role of Trust Cues in Interfaces to e-Commerce Applications", First IFIP Conference on e-commerce, e-business and e-government, 2001.

[32] Daniel W. Manchala, "E-Commerce Trust Metrics and Models", IEEE Internet Computing, April 2000.