Performance of a Security Control Scheme for a Health Data Exchange System

Erol Gelenbe 
Fellow, IEEE  
Institute of Theoretical & Applied Informatics  
Polish Academy of Sciences (IIITIS-PAN)  
ul. Bałtycka 5, 44100 Gliwice, PL

Mihajlo Pavloski  
Facebook  
11 Rathbone Square  
London W1T 1EB, UK

Abstract—We propose an access control scheme to improve the security of a cross border health data access running under distributed systems such as OpenNCP that cooperate through message exchanges. The scheme allows messages that request data and services to cross different regional or national nodes after they are verified for security issues both by the sender and the receiver. It rejects those access requests that are detected to be malicious, and consequently throttles down the flow of permit messages via explicit feedback from threat detection software placed at the sender and the receiver. An analytical model is used to evaluate the message delay to determine the overhead caused by the enhanced security system, and numerical results illustrate the impact of attack rates and on the system’s performance.

Keywords—OpenNCP, E-Health Systems, Security, G-Networks, Performance Evaluation

I. INTRODUCTION

Across Europe, health data is stored and processed in diverse systems that have been designed separately at a regional or national level. These distinct systems have to be accessed in a secure manner, with health practitioners on a given system requesting and updating patients’ records on different systems. Access control for network security with regard to malicious attacks has not been widely investigated, though recent events, such as the WannaCry attack [1] which caused over $92 Million in damages just to the UK National Health Service, have shown the importance of improving access controls for system security.

In Europe, the OpenNCP system has been proposed as a distributed continent-wide architecture that can meet such needs [2], though cannot handle address threats to peer-to-peer communication between health services of different countries. Thus the ongoing EU project KONFIDO [3], [4] addresses security for the OpenNCP service-oriented infrastructure, where interoperating National Contact Point (NCP) nodes offer access to the e-health services of different nations and regions in Europe [5].

Each NCP node supports multiple stand-alone services, and different NCP nodes communicate bilaterally in a client-server manner, acting as a data requester (client) and data provider (server). There are five distinct services supported in the KONFIDO system: Patient Identification, Patient Summary, ePrescription, eDespensing and Consent Services [4], and within each NCP node. KONFIDO implements specific security services and encryption that protect each individual node [7]. Thus to complement node-based means of assuring system security [3], this paper proposes an access control scheme to detect attacks via malicious requests or messages directed between pairs of nodes, and we evaluate this scheme using an analytical model that quantifies the additional delay it will introduce for legitimate messages.

Robust techniques for secure access control have been investigated [8], on top of conventional security services such as authentication and auditing [9], and end-to-end security controls [10]. Such security schemes may be implemented at the access nodes themselves or handled by remote Cloud services [6].

Active queue management [11], [12] based on reducing queue sizes in network nodes by dropping arriving packets has been suggested to enhance security, once an attack has been detected [13]. Admission or access control schemes have been the subject of research via analytical models [31], and some were later adopted in commercial systems [32].

Other work has considered the mitigation of virus attacks [14] and of signalling storms in mobile networks using a statistical approach to detect and block the attacks [15]. Recent work [16] has discussed multi-domain networking environments, where malicious flows can contain repetitive attack patterns. Various approaches to improve security are discussed in [17], while earlier work [18] raised the question of how to control anomalies after they are detected, in order to prevent system congestion and failure.

In the sequel, Section II provides a high-level description of the access control system. The modeling methodology for the system is described in Section II-A. The analytical model for system performance is presented in Section III and numerical results are discussed in Section IV. Finally conclusions and extensions are discussed in Section V.

II. THE ACCESS CONTROL SYSTEM

Figure 1 shows an open queueing network with incoming requests at rate $r_0(i,j)$ and permits at rate $l(i,j)$, that describes the proposed access control system. To present a high-level view of the model, we can use a real-world example where...
a patient from one European country (home country) visits the hospital in another European country (visiting country), and the visited country health system needs to make a request to the patient’s home country in order to retrieve the needed data. The two network entities involved in establishing the connection and performing the data exchange are NCP nodes $S_i$ and $S_j$, that represent the visited and home NCP system, respectively.

The requests arrive at the request queue $U(i,j)$ and a flow of permits which can enable the transfer of requests will arrive at the permits queue $V(i,j)$. The requests leave the request queue in First-In-First-Out (FIFO) order when they are matched with a permit. Permits leave $V(i,j)$ in FIFO order and act as triggers to forward the requests in $U(i,j)$. A request that leaves $U(i,j)$ is then forwarded to the anomaly detection software $S(i,j)$ at node $S_i$ (the visited country), and the software $S^*(i,j)$ at the home country $S_j$. Thus the output of the permit queue $V(i,j)$ triggers messages that are first checked at $S(i,j)$ and then at $S^*(i,j)$ in order to detect attacks. If at least one of the security checks detects a problem with the request, they will stop the procedure, and stop or reduce the information flow from the home country to the visiting country. As shown in Figure 1, if a message or request is detected as being an attack, it is rejected by either one, or both, detection software, but it may come back later as a new request at the input queue at $U(i,j)$. Also, depending on how severe it is viewed to be, represented by an integer $L$, the number of permits in the queue $V(i,j)$ is reduced by $L$. Moreover, the information about requests can also be passed to a cross-border wide “higher level control” to share security related information with all communicating pairs $(i,j)$, and regulate the release of permits.

**A. Useful Queueing Models**

The modelling approach in the following sections is based on Queueing Network Theory [19] that treats the individual units of information flow, such as messages, data requests or data responses, simply as “jobs” in a network of service centers or queues. Simple forms of these systems was first described in [20], [21] that showed the remarkable property of Product Form Solutions, for distributions of the joint queue lengths of multiple service centres in open and closed systems in steady state with a single class of jobs. Later, product-form solutions for novel types of networks consisting of positive jobs and signals, known as G-networks were developed around 1990 [22], with research that currently continues [23]–[28]. Here we will use the batch-removal and triggering capability of G-networks [29], [30] to build a feedback control system for the flow of permits in response to the detection of threats of $M$ different levels.

**III. PERFORMANCE OF THE CONTROL ALGORITHM**

In this paper we study a permit-based system that is, enabled by a higher level control which has an overview over all the nodes in the network. Given NCP nodes $S_i$ and $S_j$ that belong to different nations’ (or regions’) health systems, requests must be matched by permits to authorize the data transfer between two nodes installed in different countries.

New requests for access to the patients’ data arrive to the NCP node $S_i$ for a connection to some other node $S_j$ at rate $r_0(i,j)$ and are stored in an input queue $U(i,j)$. A higher level control regulates the communication of node $S_i$ with another node $S_j$ by issuing permits at rate $t(i,j)$, which are stored in the permit queue $V(i,j)$. Denote by $u(i,j)$ the probability that there is at least one request in $U(i,j)$, and by $q(i,j)$ the probability that there is at least one permit in $V(i,j)$. Let $w(i,j) > 0$ be the maximum speed or rate at which permits are forwarded out of $V(i,j)$, provided that $q(i,j) > 0$, so that the effective rate at which permits allow request messages to be processed is $w(i,j)q(i,j)$.

Once a request for communication is admitted, originating from the health system at node $S_i$ towards node $S_j$, it is forwarded to the anomaly detection software at $S_i$ (denoted $S(i,j)$), and simultaneously to the anomaly detection software at node $S_j$ (denoted $S^*(i,j)$). Thus $S(i,j)$ and $S(i,j)^*$ are the software that effect the security checks at each of the nodes for the message that is forwarded $i \rightarrow j$. They carry out functions such as Message encapsulation (to avoid propagation of malware through the system), then description and interpretation, Signature verification, Pivoting, Machine Learning based anomaly detection [33], and finally Access control authorization. Or to the contrary, if the message is viewed
as an attack, the message is blocked and stored for further analysis, and the incoming messages are throttled as described below. The service rates at the two anomaly detection nodes are given with \( s(i,j) \) and \( s^*(i,j) \) respectively. Similarly, we can denote with \( Q(i,j) \) and \( Q^*(i,j) \) the probability that there is at least one message in \( S(i,j) \) and \( S^*(i,j) \).

Since the security checks are performed on both data requests and responses, in the following we can refer to both requests and responses simply as “messages”. The details of the above security checks are out of the scope of this model, but the detector’s performance metrics are important. The following performance metrics of the attack detection are of interest:

- **Probability of a true positive decision** \( \pi_{D_i} \) at node \( i \), the portion of correctly detected anomalies among the malicious messages (probability of correct detection).
- **Probability of false positive decisions** \( \pi_{F_i} \) at node \( i \), or the portion of incorrectly detected anomalies among the real normal or non-malicious messages (probability of false alarm).

The above metrics can be estimated from a detector’s performance or “ground truth”. Given a message that is passed from \( S_i \) to \( S_j \), we denote by:

- \( a(i,j,L) \) the probability that the detector \( S(i,j) \) at node \( S_i \) makes the decision that the message that reaches \( i \) and is meant for \( j \), is malicious, i.e. it is the probability that the detector detects an attack.
- \( L \) is an integer, \( 1 \leq L \leq M \) which denotes how severe the threat is (i.e. the threat level), where \( M \) is the highest threat level. A larger value of \( L \) denotes a more severe threat. \( L \) will be used to slow down the rate at which permits allow the messages to be forwarded from \( i \) to \( j \).
- Similarly, \( b(i,j,L) \) is the corresponding probability for \( S(i,j) \) at node \( S_j \),
- Since the decision is taken jointly by both nodes, the overall probability of detection is

\[
a(i,j,L) = \sum_{L=1}^{M} a(i,j,L), \quad b(i,j,L) = \sum_{L=1}^{M} b(i,j,L).
\]  

(1)

If \( a(i,j,L) \) is the fraction of malicious messages of threat level \( L \) among all messages exchanged from \( S_i \) to \( S_j \), the overall probability of detection of an anomaly, including a wrong detection, at node \( i \) is:

\[
a(i,j,L) = a(i,j,L)\pi_{D_i} + (1 - a(i,j,L))\pi_{F_i},
\]  

(2)

while if node \( j \) carries out the same analysis independently of \( i \), the corresponding quantity at \( j \) is:

\[
b(i,j,L) = a(i,j,L)\pi_{D_j} + (1 - a(i,j,L))\pi_{F_j}.
\]  

(3)

Let \( c(i,j,L) \) be the probability of detection decision jointly for both \( S_i \) and \( S_j \) nodes i.e., the probability that at least one of the two detectors decides that it has detected a malicious message of threat level \( L \):

\[
c(i,j,L) = 1 - (1 - a(i,j,L))(1 - b(i,j,L)).
\]  

(4)

Finally, we can calculate the utilizations of the anomaly detection nodes \( Q(i,j) \) and \( Q^*(i,j) \) as follows:

\[
Q(i,j) = \frac{r(i,j)}{s(i,j)}, \quad Q^*(i,j) = \frac{(1 - a(i,j))r(i,j)}{s^*(i,j)}.
\]  

(5)

where \( a(i,j) \) is the sum of \( a(i,j,L) \) probabilities over \( L \). Having defined the parameters of the anomaly detectors, we can feed their decision back into the access control scheme as feedback that results in the following action:

- If a threat of level \( L \geq 1 \) is detected in a message \( i \to j \), the message is discarded. Also, the access control scheme removes \( L \) permits from queue \( V(i,j) \) so as to throttle the flow of possible future malicious messages;
- It will also provide information to the higher level controls on the decision made for communicating nodes \( i \to j \). This aspect is not considered any further in the present paper.
- However, if a message is identified as being malicious and stopped from proceeding beyond the input points of node \( S_i \) and \( S_j \), with some probability \( d(i,j,L) \) it will return to the request input queue at \( i \) and the same procedure will be repeated. Note that the returning message is not necessarily malicious, since the detector is subject to false alarms with probability \( \pi_{F} \).

The removal of permits from the queue \( V(i,j) \) will modify the permit queue’s utilization \( q(i,j) \), and therefore influence the whole message admission process and the delays both for normal and malicious messages. The resulting utilization can be computed using the model of G-networks with batch removal [29]:

\[
q(i,j) = \min(1, \frac{t(i,j)}{w(i,j) + r^*(i,j)} - \frac{1}{\frac{1}{\sum_{L=1}^{M} p(i,j,L)q(i,j)^L} - \frac{1}{\sum_{L=1}^{M} p(i,j,L)q(i,j)^L}}),
\]  

(6)

where \( p(i,j,L) \) is the probability distribution for the tokens removal, such that \( \sum_{L=1}^{M} p(i,j,L) = 1 \), and \( r^*(i,j) \) is the rate of returning messages i.e., the portion of the total rate of messages which is detected as malicious, and is given with:

\[
r^*(i,j) = a(i,j)Q(i,j) + b(i,j)Q^*(i,j) = [a(i,j) + b(i,j)(1 - a(i,j))]r(i,j),
\]  

(7)

where again \( a(i,j) \) and \( b(i,j) \) are summed up over \( L \). The rejected and then re-emitted input messages will also increase the total incoming rate of message requests, which then becomes:

\[
r(i,j) = r_0(i,j) + r(i,j) \sum_{L=1}^{\infty} c(i,j,L)d(i,j,L),
\]  

(8)

\[
r_0(i,j) = \frac{r_0(i,j)}{1 - \sum_{L=1}^{\infty} c(i,j,L)d(i,j,L)}.
\]

IV. NUMERICAL RESULTS

The model can be used to calculate important performance metrics from both user and system perspectives, and to spot
any bottlenecks in the system due to poor design in the communication procedures, such as congestion due to processing time. For example, time consumption of encryption/decryption procedures can be inserted in the model as a random delay that modifies the service rate of the anomaly detectors. Its effect can then be examined for request arrival rates, or different attack rates. This analysis would help us test the system, and adjust its design to optimize the communication, processing and memory resources that are available at the nodes.

To illustrate this approach we calculate a user-perspective performance metric: the average waiting time \( W(i, j) \) that a request experiences before it is admitted into the system prior to being processed by the attack detectors at nodes \( i \) and \( j \).

We use G-network theory to write:

\[
W(i, j) = \frac{1}{q(i, j)w(i, j) - r(i, j)}. \tag{9}
\]

To obtain numerical results, we must select some of the parameters regarding the flows in the network, and the anomaly detectors. Note that, in practice, these values would come from historical analysis of the corresponding components and functions. To normalize the rates in the system, we take \( t(i, j) = 1 \) and \( w(i,j) = 1 \). Then, assuming we have \( M = 5 \) threat levels, and the probability of detecting a threat is uniform across different levels i.e. \( c(i, j, L) = \frac{c(L)}{M} \) for \( L = 1, \ldots, M \). The probability of an anomalous request returning in the system is \( d(i, j, L) = 0.5 \). The detectors can be modelled using \( \pi_D = 0.9 \) and \( \pi_F = 0.05 \) for both \( S_i \) and \( S_j \) nodes, meaning that on average they detect 90% of all attacks, and mistakenly identify as attacks 5% of the messages which are benign.

Note that applying the results of G-networks with batch removal [29], the system has three tandem servers, the permit controller, followed by the two anomaly detection queues, with batch removal “orders” flowing back to the batch removal queue of permits. Assuming Poisson external arrivals of message processing requests for node \( S_i \) directed at node \( S_j \), and that the system is stable, i.e. \( r(i, j) < q(i, j)w(i, j), r(i, j) < s(i, j) \) and \( r(i, j)(1 - a(i, j)) < s^*(i, j) \), each of the two detectors introduces an average queueing delay of:

\[
W_d(i, j) = \frac{1}{s(i, j) - r(i, j)}, \tag{10}
\]

\[
W_d^*(i, j) = \frac{1}{s^*(i, j) - r(i, j)(1 - a(i, j))}, \tag{11}
\]

because only a proportion \( (1 - a(i, j)) \) of messages that arrive at \( S(i, j) \) are then forwarded to \( S^*(i, j) \). In this case, the total average admission wait time will be:

\[
W^+(i, j) = \frac{1}{q(i, j)w(i, j) - r(i, j)} + \frac{1}{s(i, j) - r(i, j)} + \frac{1}{s^*(i, j) - r(i, j)(1 - a(i, j))}. \tag{12}
\]

\( W^+(i, j) \) is plotted in both Figures 2 and 3, respectively against the arrival rate \( r_0(i, j) \) and the permit generation rate \( t(i, j) \) with the other parameter values as stated in the figure captions.

We will vary the percentage of anomalies \( \alpha(i, j) \), and also vary the arrival rate \( r_0(i, j) \) of fresh incoming messages, i.e. those that do not result from the repetition of a message that was detected as being an attack. The results in Figure 2 show curves for three values of anomaly percentage: 5%, 10% and 20% percent, and a case without anomalous requests (0%). As expected, the admission wait time increases with the increase of the incoming request arrival rate, until a point where the system becomes unstable. This happens when the system with the above parameters, is unable to keep up with the rate of arrival of requests. Furthermore, Figure 2 also shows the total average admission wait time \( W^+(i, j) \) for the same set percentages of anomalies. Here the anomaly detectors \( S(i, j) \) and \( S^*(i, j) \) are both modelled as single server queues with exponential service times and service rates \( s(i, j) \), \( s^*(i, j) \) respectively.

Fig. 2. Average admission wait time \( W(i, j) \) as a function of incoming request rate of messages for transfers \( r(i, j) \). We assume that the permit rate is fixed at \( t(i, j) = 1 \) and the maximum speed of the permit controller is also set at \( w(i, j) = 1 \). The total average wait time for requests \( W^+(i, j) \) is also shown, including the effect of the queues at the two anomaly detectors, when the average time it takes each detector to processes a single message is 2.5 so that \( s(i, j) = s^*(i, j) = 0.4 \).

Figure 3 shows the same four cases for the percentage of anomalies, with the incoming request arrival rate set to \( r_0(i, j) = 0.1 \), but we vary the rate of permit arrivals \( t(i, j) \). For example, we see that if permits are generated at rate 0.25, an introduction of 10% malicious messages can increase the average admission wait time by 100%. Faster permit generation significantly reduces admission wait times, but it can overload the detection systems as in the total average wait times.
admission wait time $W^+ (i, j)$. 

![Fig. 3. Average admission wait time $W^+ (i, j)$.

Fig. 3. Average admission wait time $W^+ (i, j)$ as a function of the permit arrival rate $r(i, j)$. We assume that the external message arrival or request rate is fixed at $r(i, j) = 0.1$ and the maximum speed of the permit controller is set at $w(i, j) = 1$, and both $S_i$ and $S_j$ have a correct detection probability $0.9$ and a false alarm probability $0.05$. The total average wait time $W^+ (i, j)$, including the effect of the anomaly detectors, is also shown when $s(i, j) = s^*(i, j) = 0.4$.

V. CONCLUSIONS

We have addressed the use of access control schemes to enhance the security of a distributed system such as OpenNCP that allows the interoperability of distinct regional or national health systems that cooperate through message exchanges. We have proposed a permit based admission scheme that allow messages that request services to cross different regional or national nodes after they are verified for security issues both by the sender and the receiver.

The scheme rejects those access requests that are detected to be malicious, and throttles down the permit rate via explicit feedback from software placed both at the sender and the receiver, by eliminating a number of permits proportionally to the threat level.

Using a queueing model, we have computed the effect of this security driven control scheme on system performance, by analyzing the delay incurred on all incoming messages that is introduced by this scheme.

In future work it would be interesting to see whether security and performance can be improved with two interacting but separate permit based sending and receiving access schemes that could strengthen security while allowing both senders and receivers to regulate the flow of requests so as to provide better performance.

REFERENCES


