Network Intrusion Detection with Minimal Communication Overhead

O. Patrick Kreidl and Alan S. Willsky
MIT Laboratory for Information and Decision Systems, Cambridge, MA 02139
{opk,willsky}@mit.edu

Abstract

We consider a simplest probabilistic model for network intrusion detection, assuming that (i) each individual computer node hosts a local detector with tunable thresholds and (ii) there exist explicit and severe constraints on the allowable inter-node communication during system operation. Optimal threshold selection is characterized as a function of the uncertain environment, the network constraints and the relative tolerances of false positives/negatives; more specifically, we deduce that all detectors' thresholds are globally-coupled through a system of nonlinear equations that admits an efficient iterative numerical solution. Illustrative examples expose a number of subtle design considerations, including that (i) there should typically be different thresholds across all nodes (even if all nodes employ identical local detection systems) and (ii) seemingly reasonable heuristics for coupling the thresholds can result in catastrophic network-wide performance (i.e, worse than that achieved when every node simply acts in isolation).

1. Introduction

The simplest formulation of the *single-computer* intrusion detection problem is arguably as a statistical binary hypothesis test [1], depicted in Fig. 1. Given a (noisy) Euclidean-valued observation $Y = y$ of a (hidden) binary-valued state process X, with joint process (X, Y) described by probability distribution $p_{X,Y}(x, y)$, the problem is to choose one of the two alternatives with minimum risk of errors i.e., a *false positive* or *false negative*, corresponding to the choice of $\hat{x} \neq x$ when $x = -1$ or $x = +1$, respectively. Its solution is the celebrated *likelihood-ratio threshold test*: given any observation $y \in \mathcal{Y}$, choose $\hat{x}$ according to

$$\underbrace{\frac{p_{Y|X}(y|+1)}{p_{Y|X}(y|-1)} = L(y)}_{\text{likelihood-ratio given } y} \quad \begin{array}{c} \hat{x}=+1 \\ > \\ < \\ \hat{x}=-1 \end{array} \quad \underbrace{\eta = \frac{p_X(-1)c^{FP}}{p_X(+1)c^{FN}}}_{\text{threshold}}, \quad (1)$$

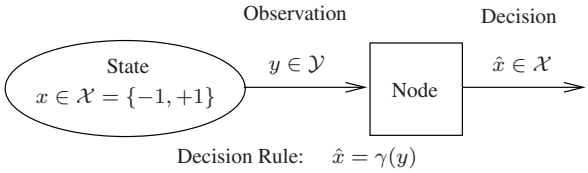

Figure 1. Single-node detection model

where (i) parameters c^{FP} and c^{FN} denote the costs assigned to the two error types and (ii) the *prior probabilities*

$$p_X(x) = \int_{\mathcal{Y}} p_{X,Y}(x, y)dy, \quad x \in \{-1, +1\}$$

and the so-called *likelihood function*

$$p_{Y|X}(y|x) = p_{X,Y}(x, y)/p_X(x), \quad (x, y) \in \{-1, +1\} \times \mathcal{Y}$$

are both defined from the given distribution $p_{X,Y}(x, y)$ [1].

This paper formulates (in Section 3) the *multi-computer* network generalization of this simplest intrusion detection model. Firstly, we assume there are n total computer nodes, each ultimately deciding upon its own (hidden) state variable based on its own local observation. Secondly, before making its local state-related decision, each node has the opportunity to exchange information with its one-hop neighbors in the network (as defined by the edge set of a given n-node undirected graph), but *only* its one-hop neighbors. Moreover, in contrast to other recent work in distributed intrusion detection (e.g., [4, 5]), this information-sharing is taken to be *severely* limited relative to the volume of information comprised in each local observation; as such, subject to design at each node is not just the rule for its final state-related decision, but also the rule by which to generate the "summarizing bits" to transmit to its neighbors and, in turn, interpret the bits received from those same neighbors. Fig. 2 illustrates a typical network (with $n = 100$) under our consideration, which we note is only sparsely-connected. Altogether, these severe communication constraints befit a requirement of having negligible security overhead, reserving ample resources for intended network services. Finally, our generalization captures the fact that multiple nodes are

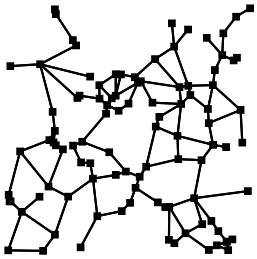

Figure 2. A typical sparsely-connected network in our multi-node detection model

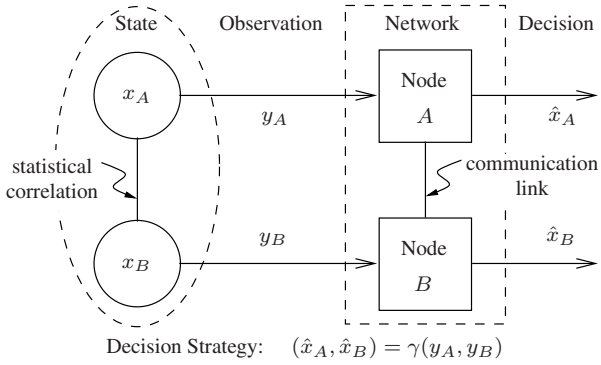

Figure 3. A two-node detection model

likely to fall under attack simultaneously, such correlation between any pair of nodes in direct relation to their proximity (e.g., distance in hops) in the network.

Our analysis shows that, under rather practical assumptions, every node should continue to generate decisions via (local) likelihood-ratio tests of the form in (1), but suitably generalized to involve a *collection* of thresholds. Specifically, for each communication-related decision, selecting one of say d symbols to send to its neighbors (e.g., $d = 2, 4, 8, ...$ when allotted one, two, three,... bits of overhead), there can be $d-1$ distinct thresholds; in turn, for each final state-related decision, there can be a unique threshold per possible value of the *composite* symbols received from all neighbors (e.g., in the case of three neighbors, each sending one of d symbols, this final rule could involve d^3 distinct thresholds). The minimum-cost settings of all thresholds turn out to be globally coupled through a system of nonlinear equations, which takes into account the uncertain environment, the network constraints and the specified error tolerances across *all* nodes. This system of equations lends itself to an iterative solution algorithm, which is both convergent and computationally efficient, scaling linearly with the number of nodes n (but exponentially with maximal node degree, so it's best suited for sparsely-connected networks). With respect to overall detection performance, simulation experiments verify that our network-constrained solutions fall short of that achievable in the absence of such constraints, but the gap is surprisingly small in many cases.

In the next section, before proceeding to describe the problem formulation and solution algorithm for the general case (in Section 3), we shall introduce the key concepts by way of a small example. Throughout, for a random variable V that takes its values in a discrete (or Euclidean) set $\mathcal{V}$, we let $p_V : \mathcal{V} \to [0, \infty)$ denote its probability mass (or density) function; similarly, let $c_V : \mathcal{V} \to [0, \infty)$ denote a cost function associated to V. We shall suppress the subscript notation when the random variable involved is implied by the functional argument; that is, we let $p(v) \equiv p_V(v)$ and $c(v) \equiv c_V(v)$ for every $v \in \mathcal{V}$. Also note that $p(V)$ and $c(V)$ are themselves well-defined random variables, each taking values in $[0, \infty)$ according to a distribution derived from V and the functions p_V and c_V, respectively. The expectation of random variable V is denoted by $\mathbf{E}[V]$.

2. Illustrative example: a two-node network

This section focuses on a simplest instance of the network intrusion detection problem, namely the case of two nodes (labeled A and B) depicted in Fig. 3. The two-node model relates to the classical setup of Fig. 1 in a number of ways. From the perspective external to the network, it is equivalent to the classical model with (hidden) state $x = (x_A, x_B)$ and (noisy) observation $y = (y_A, y_B)$ taking values in product sets $\{-1, +1\}^2$ and $\mathcal{Y}_A \times \mathcal{Y}_B$, respectively. The network can also be viewed as two distinct instances of the classical setup, each node observing (and ultimately responsible for deciding the value of) only its own binary-valued state variable; however, the two state variables can be correlated and, as such, the nodes may benefit by first exchanging information about their local observations. The following subsections describe a specific probabilistic setup and analyze alternative schemes by which the nodes collectively process their observations to generate both communication-related and state-related decisions.

2.1. Problem setup

We first specify the distribution $p(x, y) = p(x)p(y|x)$ for the model in Fig. 3, jointly defining the random variables across both nodes i.e., state process $X = (X_A, X_B)$ and observation process $Y = (Y_A, Y_B)$. Given parameters $q_A \in (0, \frac{1}{2}]$, $q_B \in (0, q_A]$ and $r \in [-\frac{q_A q_B}{Z(q)}, \frac{(1-q_A)q_B}{Z(q)}]$ with $Z(q) = \sqrt{q_A(1-q_A)q_B(1-q_B)}$, let

$$p(x) = \begin{cases} (1-q_A)(1-q_B) + rZ(q), & x = (-1,-1) \\ (1-q_A)q_B - rZ(q) \quad , & x = (-1,+1) \\ q_A(1-q_B) - rZ(q) \quad , & x = (+1,-1) \\ q_A q_B + rZ(q) \quad , & x = (+1,+1) \end{cases}.$$

It is straightforward [1] to verify that q_i is the (marginal) probability of binary variable X_i taking its positive value i.e., $q_i = p_{X_i}(+1)$ for $i \in \{A, B\}$, and that r is the correlation coefficient between X_A and X_B. Here, probabilities $q = (q_A, q_B)$ capture assumptions about how commonly intrusion attempts occur on a per-node basis, while the magnitude of r captures how commonly the nodes are targeted simultaneously (when $r > 0$) or individually (when $r < 0$). Let the global likelihood function be given by the product $p(y|x) = p(y_A|x_A)p(y_B|x_B)$, each node's local likelihood function a Gaussian distribution with state-dependent mean,

$$p(y_i|x_i) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(y_i - \frac{s_i x_i}{2}\right)^2\right], \quad i \in \{A, B\}.$$

Here, parameter $s_i \in (0, \infty)$ is the signal-to-noise ratio for the real-valued observation process Y_i e.g., if $s_A > s_B$, then Y_A is more informative about X_A than Y_B is about X_B.

It remains to specify the cost function $c(\hat{x}, x)$, quantifying the relative undesirability of all decision-state pairs across both nodes. We take this cost to be the sum of two local costs i.e., let $c(\hat{x}, x) = c(\hat{x}_A, x_A) + c(\hat{x}_B, x_B)$ with

$$c(\hat{x}_i, x_i) = \begin{cases} t_i & , \quad \hat{x}_i = +1 \text{ and } x_i = -1 \\ 1 & , \quad \hat{x}_i = -1 \text{ and } x_i = +1 \\ 0 & , \quad \text{otherwise} \end{cases},$$

where parameter $t_i \in (0, 1]$ represents the cost of a false positive relative to that of a false negative local to each node $i \in \{A, B\}$. Here, a false positive is never more costly than a false negative, and $t_i = 1$ is the special case in which node i considers either type of error to be equally costly.

As implied by Fig. 3, subject to design is the *strategy*, or function $\gamma : \mathcal{Y} \to \mathcal{X}$, by which any particular *composite* observation $y = (y_A, y_B)$ maps to both nodes' state-related decisions $\hat{x} = (\hat{x}_A, \hat{x}_B)$. The associated risk is defined by

$$J(\gamma) = \mathbf{E}\left[c(\hat{X}, X)\right] = \mathbf{E}\left[\mathbf{E}\left[c(\gamma(Y), X)|\, Y\right]\right], \qquad (2)$$

which in correspondence with the cost function specified above can be expressed as the sum $J_A(\gamma) + J_B(\gamma)$ with

$$J_i(\gamma) = \sum_{x_i \in \{-1,+1\}} p(x_i) \sum_{\hat{x}_i \in \{-1,+1\}} c(\hat{x}_i, x_i) p^\gamma(\hat{x}_i|x_i)$$

for each node $i \in \{A, B\}$. Here, the (strategy-dependent) conditional distribution $p^\gamma(\hat{x}_i|x_i)$ encapsulates the false-positive and false-negative probabilities at each node i, so in terms of the specific problem setup above we have

$$J_i(\gamma) = (1 - q_i) t_i p^\gamma_{\hat{X}_i|X_i}(+1|-1) + q_i p^\gamma_{\hat{X}_i|X_i}(-1|+1).$$

Moreover, both nodes' error probabilities can depend on parameters q_A, q_B, r, s_A and s_B as implied by the identities

$$p^\gamma(\hat{x}|y) = \begin{cases} 1 & , \quad \text{if } \hat{x} = \gamma(y) \\ 0 & , \quad \text{otherwise} \end{cases}, \qquad (3)$$

$$p^\gamma(\hat{x}|x) = \int_{\mathcal{Y}} p^\gamma(\hat{x}|y) p(y|x)\, dy$$

and $p^\gamma(\hat{x}_i|x_i) = \frac{p^\gamma(\hat{x}_i, x_i)}{p(x_i)}$ where, at node A for example,

$$p^\gamma(\hat{x}_A, x_A) = \sum_{x_B \in \{-1,+1\}} p(x) \sum_{\hat{x}_B \in \{-1,+1\}} p^\gamma(\hat{x}|x).$$

2.2. Performance Analysis

We now consider several decision strategies for this two-node network, differing in the assumed communication overhead, and quantify their performances as measured by (2). The *myopic* strategy requires zero communication overhead, viewing the two-node problem as two isolated single-node problems. We then describe a *heuristic* strategy that requires two bits of communication overhead, each node transmitting the minimum-error estimate of its local state and, in turn, interpreting its received symbol as the value of the other node's local state. These two strategies are compared with the *team* strategy, to be described in Section 3, which similarly requires only two bits of communication overhead. The results show that the heuristic strategy fails catastrophically, performing even worse than the myopic strategy in most cases, whereas the team strategy consistently improves upon myopic performance. We also provide results for the team strategy when allotted additional communication bits, showing its performance rapidly approaches that of the optimal *centralized* strategy in which the link is unconstrained (i.e., supports infinite precision).

The myopic strategy is easiest to describe, each node i employing its instance of the single-node solution in (1),

$$\underbrace{\exp(s_i y_i) = L_i(y_i)}_{\text{likelihood-ratio local to node } i} \begin{array}{c} \hat{x}_i = +1 \\ > \\ < \\ \hat{x}_i = -1 \end{array} \underbrace{\eta_i = \frac{(1 - q_i)}{q_i} t_i}_{\text{threshold at node } i}. \qquad (4)$$

That is, the decision strategy amounts to the pair of local rules $\gamma = (\delta_A, \delta_B)$, each node's rule $\delta_i : \mathcal{Y}_i \to \{-1, +1\}$ defined by threshold η_i. The associated error probabilities then specialize to $p^\gamma(\hat{x}|x) = p^\gamma(\hat{x}_A|x_A)p^\gamma(\hat{x}_B|x_B)$ with

$$p^\gamma_{\hat{X}_i|X_i}(-1|x_i) = \mathbf{\Phi}\left(\frac{\log(\eta_i)}{s_i} - \frac{s_i x_i}{2}\right), \quad x_i \in \{-1, +1\}$$

and $p^\gamma_{\hat{X}_i|X_i}(+1|x_i) = 1 - p^\gamma_{\hat{X}_i|X_i}(-1|x_i)$ for each node i. Here, $\mathbf{\Phi}(z)$ denotes the cumulative distribution function [1] of a zero-mean, unit-variance Gaussian random variable Z.

The heuristic strategy unfolds in two distinct stages, each node i exchanging a binary-valued communication decision $u_i \in \{-1, +1\}$ with its neighbor before making its final state-related decision $\hat{x}_i$. More specifically, from the perspective of node A's state-related decision, first node B

communicates the minimum-error estimate of X_B based on local observation y_B i.e., take $t_B = 1$ in (4), or

$$L_B(y_B) \begin{array}{c} u_B = +1 \\ > \\ < \\ u_B = -1 \end{array} \theta_B = \frac{p_{X_B}(-1)}{p_{X_B}(+1)};$$

second, node A interprets the received symbol u_B as the correct value of peripheral state X_B and proceeds to generate its local minimum-risk estimate $\hat{x}_A$ based on y_A i.e.,

$$L_A(y_A) \begin{array}{c} \hat{x}_A = +1 \\ > \\ < \\ \hat{x}_A = -1 \end{array} \eta_A[u_B] = \frac{p_{X_A,X_B}(-1,u_B)}{p_{X_A,X_B}(+1,u_B)} t_A.$$

This heuristic decision strategy amounts to a collection of four rules $\gamma = (\mu_A, \mu_B, \delta_A, \delta_B)$, each node's communication rule $\mu_i : \mathcal{Y}_i \to \{-1,+1\}$ defined by threshold θ_i and detection rule $\delta_i : \mathcal{Y}_i \times \{-1,+1\} \to \{-1,+1\}$ defined by symbol-dependent thresholds $\eta_i[-1]$ and $\eta_i[+1]$. The error probabilities, again from node A's perspective, specialize to

$$p^{\gamma}(\hat{x}_A|x_A) = \sum_{u_B \in \{-1,+1\}} p^{\mu_B}(u_B|x_A) p^{\delta_A}(\hat{x}_A|x_A,u_B),$$

$$p^{\mu_B}(u_B|x_A) = \sum_{x_B \in \{-1,+1\}} p(x_B|x_A) p^{\mu_B}(u_B|x_B)$$

with $p(x_B|x_A) = p(x_A,x_B)/p(x_A)$ and

$$p^{\mu_B}_{U_B|X_B}(-1|x_B) = \mathbf{\Phi}\left(\frac{\log(\theta_B)}{s_B} - \frac{s_B x_B}{2}\right),$$

$$p^{\delta_A}_{\hat{X}_A|X_A,U_B}(-1|x_A,u_B) = \mathbf{\Phi}\left(\frac{\log\left(\eta_A[u_B]\right)}{s_A} - \frac{s_A x_A}{2}\right).$$

Fig. 4 compares the risk $J(\gamma)$ in (2) achieved by the myopic and heuristic strategies to that achieved by the team strategy (described in the next section) as a function of all model parameters (though holding those local to node A constant throughout). Except when the two hidden states are weakly correlated (i.e., when $r \approx 0$, also occurring when $q_B \to 0$) or when the observation is much more informative at node B than at node A (i.e., when $s_B \gg s_A = 1$), the heuristic strategy performs worse than the myopic strategy! That is, despite the honest intention of each node to transmit the best estimate of its local state, assuming that it's always correct at the receiver generally causes more harm than good. In contrast, the team strategy uses the two bits of communication more aptly, always performing at least as well as (and often notably better than) the myopic strategy.

Fig. 5 explores the performance comparison of Fig. 4 in more depth, focusing on the upper-left case when varying correlation coefficient r with all other parameters fixed.

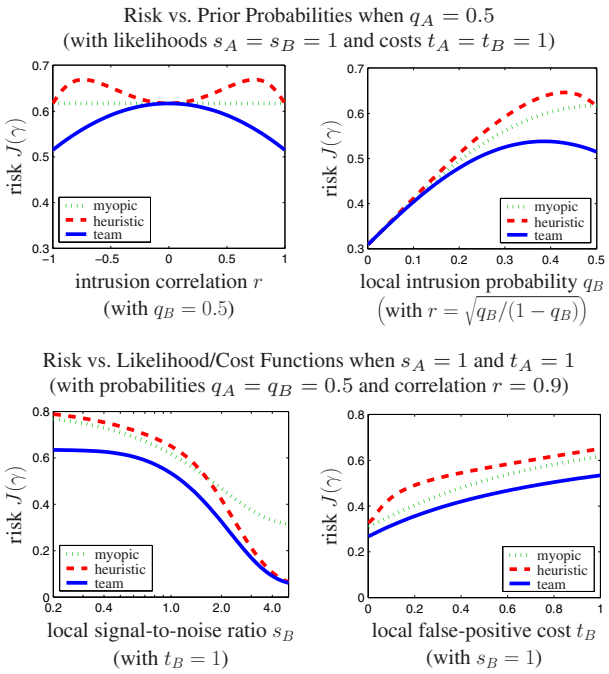

Figure 4. Two-node performance analysis

In the standard minimum-risk solution [1], which neglects the network constraints, each node's decision threshold depends on the other node's likelihood-ratio: specifically for node A to decide $\hat{x}_A$, this threshold is $\frac{1+\alpha L_B(y_B)}{\alpha + L_B(y_B)}$ with $\alpha = \frac{1-r}{1+r}$. Fig. 5(a) contrasts the decision regions implied by these centralized thresholds against those of the network-constrained thresholds: in essence, the team strategy strives to best mimic these centralized regions subject to the shape that respects the allotted two-bits of communication overhead. Fig. 5(b) shows team performance improving monotonically with additional communication overhead, in this example already close to optimal centralized performance when each node i conveys a two-bit symbol, or a communication decision $u_i \in \{1,2,\ldots,d\}$ with $d = 4$, to the other before either then makes its local state-related decision $\hat{x}_i$.

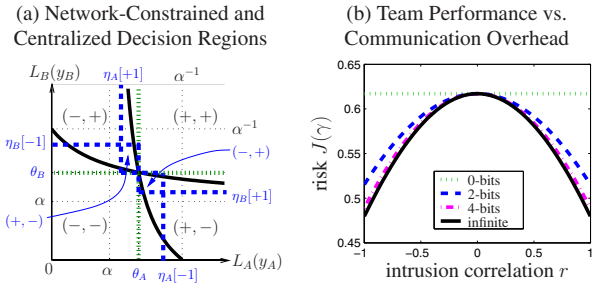

Figure 5. Insight into two-node team strategy

3. General formulation and its team solution

The n-node generalization of the problem analyzed in the preceding section is a special case of the problems studied in [2]. The treatment here must be brief: the reader interested in derivations and other technical background may consult [2]. Throughout, given an index set $\mathcal{I} \subseteq \{1, \ldots, n\}$ and index $i \in \mathcal{I}$, the notation $\mathcal{I} \backslash i$ is shorthand for the set difference $\mathcal{I} - \{i\}$. Given a multi-variate distribution $p(v_1, v_2, \ldots, v_n)$ of n distinct random variables, each V_i taking values in a set $\mathcal{V}_i$, the associated length-n random vector $V = (V_1, \ldots, V_n)$ takes its values in the product set $\mathcal{V}_1 \times \cdots \times \mathcal{V}_n$ with the same distribution $p(v)$. The length-$|\mathcal{I}|$ random vector $V_{\mathcal{I}} = \{V_i \mid i \in \mathcal{I}\}$ takes its values in the product set $\prod_{i \in \mathcal{I}} \mathcal{V}_i$ with distribution $p(v_{\mathcal{I}})$ obtained by marginalizing $p(v)$ over all variables V_j, $j \notin \mathcal{I}$.

A given undirected graph $\mathcal{G}$ defines the n-node network, each edge (i, j) representing a $(\log_2 d)$-bit bidirectional communication link between node i and node j. The subset of nodes $ne(i) = \{j \mid \text{edge } (i, j) \text{ in } \mathcal{G}\}$ are then the neighbors of each node i. The state and observation processes X and Y each generalize to length-n random vectors, having components in correspondence with the n nodes, as do the (strategy-dependent) communication-related and state-related decision processes $(U, \hat{X}) = \gamma(Y)$.

The problem is otherwise the same as that posed in Section 2: given joint distribution $p(x, y)$ and cost function $c(\hat{x}, x)$, minimize (over all decision strategies γ) the risk $J(\gamma)$ in (2) subject to the communication constraints implied by the network. Specifically, a network-constrained strategy consists of a collection of $2n$ local rules $\gamma = (\mu_1, \ldots, \mu_n, \delta_1, \ldots, \delta_n)$, the ith such communication rule and detection rule, respectively, of the functional form $\mu_i : \mathcal{Y}_i \to \{1, \ldots, d\}$ and $\delta_i : \mathcal{Y}_i \times \{1, \ldots, d\}^{|ne(i)|} \to \{-1, +1\}$. Note that, for each node i, fixing the rules by which $U_i = \mu_i(Y_i)$ and $\hat{X}_i = \delta_i(Y_i, U_{ne(i)})$ define distributions

$$p^{\mu_i}(u_i|y_i) = \begin{cases} 1 & , \quad \text{if } u_i = \mu_i(y_i) \\ 0 & , \quad \text{otherwise} \end{cases}$$

and

$$p^{\delta_i}(\hat{x}_i|y_i, u_{ne(i)}) = \begin{cases} 1 & , \quad \text{if } \hat{x}_i = \delta_i(y_i, u_{ne(i)}) \\ 0 & , \quad \text{otherwise} \end{cases} ,$$

respectively. In turn, fixing the network-consrained strategy $\gamma = (\mu_1, \ldots, \mu_n, \delta_1, \ldots, \delta_n)$ defines the distribution

$$p^{\gamma}(u, \hat{x}|y) = \prod_{i=1}^{n} p^{\mu_i}(u_i|y_i) p^{\delta_i}(\hat{x}_i|y_i, u_{ne(i)}),$$

which is the network generalization of (3).

The following assumption enables the team solution to be reduced to an iterative numerical algorithm that retains correctness, convergence and tractability in large networks.

Assumption 1 *The global likelihood function $p(y|x)$ and global cost function $c(\hat{x}, x)$ satisfy, respectively,*

$$p(y|x) = \prod_{i=1}^{n} p(y_i|x_i) \quad \textit{and} \quad c(\hat{x}, x) = \sum_{i=1}^{n} c(\hat{x}_i, x_i).$$

In words, (i) each node's observation process Y_i, conditioned on the local state process X_i, is mutually independent of all other (state and observation) processes in the network and (ii) the global risk is the sum of all local risks.

In light of our focus on large and sparsely-connected networks, the model restrictions imposed by Assumption 1 are arguably less limiting in practice than in theory. This is because a general representation of $p(y|x)$ or $c(\hat{x}, x)$ requires a number of parameters that scales exponentially with n (and, in turn, the learning of $p(y|x)$ presents daunting data collection requirements). However, representations that satisfy Assumption 1 require only the per-node quantities $p(y_i|x_i)$ and $c(\hat{x}_i, x_i)$, implying a number of parameters that scales linearly in n (and presenting no additional learning challenges than those of a single-node intrusion detection problem). Moreover, the factored form of $p(y|x)$ supports per-node compartmentalization of all observables (e.g., sets $\mathcal{Y}_i$ and $\mathcal{Y}_j$ may be disjoint for every pair of nodes i and j), which is inherently the case in network intrusion detection problems. As such, Assumption 1 alleviates the challenge of timely disseminating a (typically voluminous) set of actual observables throughout the network. Of course, there remains the challenge of selecting all local rules so that communicated summaries are "maximally-informative" for the nodes' final state-related decisions.

Definition 1 *Given a communication rule μ_i^k at node i, let*

$$P_i^k(u_i|x_i) = \int_{\mathcal{Y}_i} p(y_i|x_i) p^{\mu_i^k}(u_i|y_i)\, dy_i;$$

similarly, given a detection rule δ_i^k at node i, let

$$Q_i^k(\hat{x}_i|x_i, u_{ne(i)}) = \int_{\mathcal{Y}_i} p(y_i|x_i) p^{\delta_i^k}(\hat{x}_i|y_i, u_{ne(i)})\, dy_i.$$

Definition 2 *Given a communication rule μ_m^k for every neighbor $m \in ne(i)$ of node i, let*

$$R_i^k(u_{ne(i)}, x_i) = \sum_{x_{ne(i)}} p(x_i, x_{ne(i)}) \prod_{m \in ne(i)} P_m^k(u_m|x_m);$$

given also a detection rule δ_i^{k+1} at node i, then for each individual neighbor $j \in ne(i)$ let

$$S_{i \to j}^{k+1}(\hat{x}_i, x_i|x_j, u_j) = c(\hat{x}_i, x_i) \sum_{x_{ne(i) \backslash j}} \left(p(x_i, x_{ne(i)}|x_j) \right.$$
$$\left. \sum_{u_{ne(i) \backslash j}} Q_i^{k+1}(\hat{x}_i|x_i, u_{ne(i)}) \prod_{m \in ne(i) \backslash j} P_m^k(u_m|x_m) \right).$$

Notice that the quantities P_i^k and Q_i^k in Definition 1 are themselves (conditional) probability mass functions (PMFs), statistically relating each node's rule-dependent decision processes (i.e., U_i and $\hat{X}_i$, respectively) to its local hidden state process X_i and the information $U_{ne(i)}$ received from its neighbors. In Definition 2, the PMF R_i^k statistically relates the received information $U_{ne(i)}$ to the local state process X_i (as a function of all neighbors' communication rules), whereas the quantity $S_{i\to j}^k$ statistically relates the costs at node i to the communication process of neighbor j (as a function of node i's detection rule and its other neighbors' communication rules). Also notice that the above definitions make no assumptions on the prior probabilities $p(x)$ except that, for every node i, the *neighborhood marginals* $p(x_i, x_{ne(i)})$ are available. In general, computing such marginals is itself a hard problem when n grows large; in practice, however, it is common for the distribution $p(x)$ to be represented in a compact form, which typically facilitates marginalization. Indeed, in our network setup, it is natural to assume statistical correlations between neighboring states are much stronger than between states multiple hops away. In combination with the assumed network sparsity (as in Fig. 2), standard methods in the field of graphical models [3] can be employed.

Algorithm: For every node i, choose an initial communication rule μ_i^0 and, in each iteration $k = 1, 2, \ldots$, choose

$$\delta_i^k(y_i, u_{ne(i)}) = \arg\min_{\hat{x}_i} \sum_{x_i} b_i^k(\hat{x}_i, x_i, u_{ne(i)}) p(y_i|x_i) \tag{5}$$

with $b_i^k(\hat{x}_i, x_i, u_{ne(i)}) = c(\hat{x}_i, x_i) R_i^{k-1}(x_i, u_{ne(i)})$ and

$$\mu_i^k(y_i) = \arg\min_{u_i} \sum_{x_i} a_i^k(u_i, x_i) p(y_i|x_i) \tag{6}$$

with $a_i^k(u_i, x_i) = \sum_{m\in ne(i)} \sum_{\hat{x}_m} \sum_{x_m} S_{m\to i}^k(\hat{x}_m, x_m|x_i, u_i)$.

It is instructive to relate the general form of the rules in (5) and (6), which remain in the class of likelihood-ratio tests (and parameters a_i and b_i map directly to the respective thresholds), to the different two-node strategies analyzed in Section 2. As a first example, the myopic strategy is recovered by setting $b_i(\hat{x}_i, x_i; u_{ne(i)}) = c(\hat{x}_i, x_i)p(x_i)$ for all i, corresponding to every node foregoing the opportunity to communicate and simply employing its local single-sensor solution. As another example, in the case that $d = 2$ in Section 2, the team strategy is of the same form discussed for the heuristic strategy but with thresholds

$$\theta_i = \frac{a_i(+1,-1) - a_i(-1,-1)}{a_i(-1,+1) - a_i(+1,+1)}$$

and

$$\eta_i[u_{ne(i)}] = \frac{b_i(+1,-1;u_{ne(i)}) - b_i(-1,-1;u_{ne(i)})}{b_i(-1,+1;u_{ne(i)}) - b_i(+1,+1;u_{ne(i)})}.$$

Note that, in contrast to the heuristic strategy, in the team strategy each node i generates its communication decision u_i cognizant of how the symbol will be interpreted by its neighbors (captured within parameters a_i in an expected cost sense by summing the quantities $S_{j\to i}$ over $j \in ne(i)$); in addition, each node i generates its state-related decision $\hat{x}_i$ cognizant of how much uncertainty the neighbors face while generating their communication decisions (captured within parameters b_i in a probabilistic sense by the quantity R_i). Our main result establishes that this (nonlinear) coupling between the transmitters' perspectives and the receivers' perspectives of all network communication is refined with successive iterations of the optimization algorithm, avoiding the catastrophic failures in network-wide performance exhibited by the heuristic strategy.

Proposition 1 *If Assumption 1 holds, the risk in (2) at each iteration k of the algorithm is $J(\gamma^k) = \sum_{i=1}^n J_i(\gamma^k)$ with*

$$J_i(\gamma^k) = \sum_{\hat{x}_i} \sum_{x_i} c(\hat{x}_i, x_i) \sum_{u_{ne(i)}} \left[R_i^k(u_{ne(i)}, x_i) \, Q_i^k(\hat{x}_i|x_i, u_{ne(i)})\right],$$

and the sequence $\{J(\gamma^k)\}$ is non-increasing and converges.

4. Conclusion

We have proposed a simplest probabilistic model for network intrusion detection, focusing on the problem of optimal threshold selection assuming (i) each node hosts its own tunable detector and (ii) inter-node communication overhead must be kept to a minimum. We provided examples exposing the difficulty of finding satisfactory heuristic solutions, as well as a numerical algorithm for optimizing the thresholds as a function of the uncertain environment, the network constraints and the relative tolerances of false positives/negatives. While our solution is scalable to large (and sparsely-connected) networks, its practical value depends in large part on how meaningfully network attacks and local intrusion detectors can be modeled probabilistically, an important question that requires much additional research.

References

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, Belmont, MA, 2002.

[2] O. P. Kreidl. *Graphical Models and Message-Passing Algorithms for Network-Constrained Decision Problems*. PhD thesis, MIT EECS Department, Cambridge, MA, 2008.

[3] S. L. Lauritzen. *Graphical Models*. Oxford University Press, New York, NY, 1996.

[4] J. Li et. al. "Dependency-based distributed intrusion detection," in *USENIX Security '07 (DETER Workshop)*, 2007.

[5] T. Peng et. al. "Information sharing for distributed intrusion detection systems," *Journal of Network and Computer Applications*, 30(3):877-899, 2007.