Decentralized Detector Generation in Cooperative Intrusion Detection Systems

Rainer Bye[1], Katja Luther[1], Seyit Ahmet Çamtepe[1],
Tansu Alpcan[2], Şahin Albayrak[1], and Bülent Yener[3]

[1] DAI-Labor, Technische Universität Berlin
[2] Deutsche Telekom Laboratories, Berlin
[3] Department of Computer Science, Rensselaer Polytechnic Institute, NY

Abstract. We consider *Cooperative Intrusion Detection System* (CIDS) which is a distributed AIS-based (Artificial Immune System) IDS where nodes collaborate over a peer-to-peer overlay network. The AIS uses the negative selection algorithm for the selection of detectors (e.g., vectors of features such as CPU utilization, memory usage and network activity). For better detection performance, selection of all possible detectors for a node is desirable but it may not be feasible due to storage and computational overheads. Limiting the number of detectors on the other hand comes with the danger of missing attacks. We present a scheme for the controlled and decentralized division of detector sets where each IDS is assigned to a region of the feature space. We investigate the trade-off between scalability and robustness of detector sets. We address the problem of self-organization in CIDS so that each node generates a distinct set of the detectors to maximize the coverage of the feature space while pairs of nodes exchange their detector sets to provide a controlled level of redundancy. Our contribution is twofold. First, we use *Symmetric Balanced Incomplete Block Design*, *Generalized Quadrangles* and *Ramanujan Expander Graph* based deterministic techniques from combinatorial design theory and graph theory to decide how many and which detectors are exchanged between which pair of IDS nodes. Second, we use a classical epidemic model (SIR model) to show how properties from deterministic techniques can help us to reduce the attack spread rate.

1 Introduction

In this paper, we introduce self-organizing, self-adaptive and self-healing capabilities to the *Cooperative Intrusion Detection System* (CIDS) which is a distributed Artificial Immune System (AIS) based Intrusion Detection System (IDS) where nodes collaborate over a peer-to-peer overlay network to improve the detection capability and to decrease the false alarm rate. Hence, CIDS becomes a promising step towards the realization of *Autonomous Security* (AS) framework. We define Autonomous Security as an environment which provides smart and usable security mechanisms, distributed monitoring-detection-prevention with self-* properties and a security simulation/evaluation tool.

T. Masuzawa and S. Tixeuil (Eds.): SSS 2007, LNCS 4838, pp. 37–51, 2007.

The Artificial Immune System (AIS), like the Biological Immune System, is based on the distinction between self and non-self. Initially, an n-dimensional feature space is covered by detectors (i.e., n-dimensional vectors of features such as CPU utilization, memory usage, ..., number of tcp connections). During the training phase, these detectors are presented to feature vectors describing the self. Matching detectors are eliminated and the remaining are used for the detection of anomalies. Selecting all possible detectors for an IDS node would certainly provide best attack detection capability. But, it may not be feasible to store large amount of detectors and process them while trying to detect attacks in timely manner. Limiting the number of detectors on the other hand comes with the danger of missing attacks. Therefore, creating a detector set which covers the whole feature space may not be feasible, and random selections can create holes which means some critical regions on the feature space may not be covered. Kim *et al.* [1] use evolutionary algorithms and Gonzalez *et al.* [2] use monte carlo based approaches to face the coverage problem. In this paper, we assume that each IDS is assigned to a mutually exclusive region in the feature space due to approaches in [1] and [2]. Each IDS node can independently generate a detector set for the region of the feature space it is responsible for. In this scalable and decentralized approach, at most one IDS can have the proper detector for an attack and the attack may spread faster than the case where each node stores the global detector set. Thus, IDS nodes should have overlapping detector sets; in other words, detectors should be redundantly distributed to IDS nodes.

Goel *et al.* propose to distribute the detector generation [3] where each IDS is responsible for a non-overlapping region in the feature space, and it generates a mutually exclusive subset of the detector space. When an attack is detected, the corresponding detector set is broadcasted to other vulnerable nodes in the network. In this approach, processing and memory overhead of the detector generation is reduced by the expense of increased communication. Moreover, an attack may spread quickly since only one IDS has the proper detector to detect the attack, and vulnerable nodes are updated only when this specific IDS is attacked. Thus, Goel *et al.* [3] consider to create an overlap of detector sets stored on IDS nodes by using random graph $G(N,p)$[1] approach due to Erdős-Rényi [4] where each IDS exchanges its detector set with $(\log N)$ other IDS nodes.

In this paper, we assume a homogeneous network environment where each node has similar capabilities and configuration. Nodes have limited resources so that they can not store and process all possible detectors. Wireless sensor networks, networks of pico-satellites, smartphones, wireless ad-hoc networks of mobile devices can be the examples of such networks. We assume that attacks are equally probable and they are uniformly distributed over the feature space.

1.1 Our Contribution

We investigate the trade-off between scalability and redundancy of detector sets. Our contribution is twofold. First, we use Symmetric Balanced Incomplete

[1] $G(N,p)$ is a graph with N nodes where each pair of nodes have a link in between with probability p.

Block Design (SBIBD), Generalized Quadrangles (GQ) and Ramanujan Expander Graph (REG) techniques from combinatorial design theory and graph theory to decide how many and which detectors are exchanged between which pair of IDS nodes. Each IDS node uses the proposed deterministic techniques to independently decide which nodes to contact with and get their detector sets.

Communication and detector set exchange is done through a peer-to-peer overlay network. Unlike probabilistic approaches (i.e., $G(N, p)$), our deterministic approaches provide regular logical graphs for detector set exchange. In the SBIBD-based approach, every pair of IDS nodes have exactly one detector set in common. This approach provides the highest level of overlap. In GQ, not every pair of IDS nodes share a detector set but if two IDS nodes do not share a detector set, there are other IDS nodes sharing detector set with both. The GQ-based approach decreases the level of overlap in a controlled way. REG-based deterministic approach is comparable to $G(N, p)$-based probabilistic approach due to Goel *et al.* [3]. REG provides better defense against attack spread since REG are the best known expanders meaning that any subset of nodes are connected to a larger subset of nodes for a fixed node degree; equivalently, any subset of IDS nodes share detectors with the largest possible subset of IDS nodes. This property of REG help them to provide a better immunity against attacks.

In general, our approaches can be classified as decentralized, self-organizing, self-adaptive and self-healing. They are *decentralized* because detector set generation task is distributed to networked IDS nodes. Next, they are *self-organizing* because once assigned to a feature subspace, each IDS node independently generates its detector set, and independently decides nodes with which to exchange its detector set to create a controlled level of overlap. Then, they are *self-adaptive* because the network size can be increased by inserting new IDS nodes in which case IDS nodes shall re-organize. Any faulty IDS can be replaced with a new one and new IDS may independently decide which other nodes to contact to recover the detector sets. Finally, they are *self-healing* because when an attack reaches to an IDS node with a proper detector, the detector is sent to other IDS nodes to stop the attack spread.

As the second contribution, we enhance the classical epidemic model (SIR model [5]). We model the spread of both attacks and defenses (detector updates). We show how properties from deterministic techniques can help us to create a regular and controlled level overlap between detector sets of individual IDS nodes to reduce the attack spread rate.

1.2 Organization

Organization of the paper is as follows: in Section 2 we provide background information on Cooperative IDS, classical epidemic model, Balanced Incomplete Block Designs (BIBD), Generalized Quadrangles (GQ) and Ramanujan Expander Graphs (REG). In Section 3 we introduce our combinatorial design and expander graph based approaches. In Section 4 we analyze our approaches by using an epidemic model. Finally, in Section 5 we conclude.

2 Background

2.1 Cooperative IDS

Cooperative Intrusion Detection System (CIDS) utilizes the AIS (Artificial Immune System) introduced by Forrest *et al.* [6,7,8] and a peer-to-peer (P2P) overlay system [9]. The AIS principle bases on the *self/non-self* discrimination. It uses the negative selection algorithm described by Hofmeyr *et al.* [8] as the algorithm for the selection of detectors for a given set of feature vectors. A *feature vector* is an n-dimensional vector of numerical values. A detector denotes a point in the feature space, like a feature vector, with the additional information such as age.

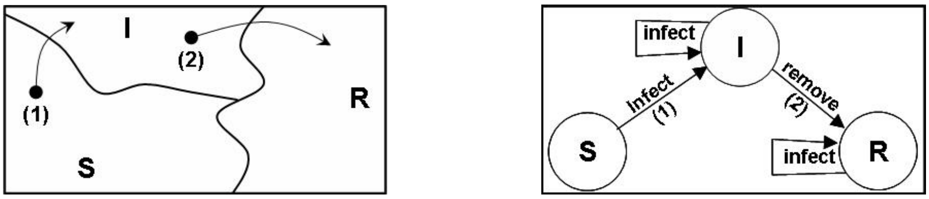
The main idea is to produce detectors randomly and compare them to the normal patterns obtained during the training. Every detector that matches to these normal patterns is eliminated, and hence, the remaining detectors recognize only abnormal patterns. In the case of the Cooperative IDS, the feature vectors might be composed of network traffic statistics measured at an end device in a specific time interval. A sample feature vector might be $<$ *timestamp, number of TCP connections, number of TCP packets, number of UDP packets, number of used ports, number of port scans* $>$. After the learning period, new measurements are presented to the remaining detectors, and the distances between the feature vectors and the detectors are computed. If the distance between a detector and a feature vector crosses a specific threshold, the matching detector is distributed.

The used P2P infrastructure is based on a hybrid decentralized architecture. The nodes are equal in their abilities but one node acts always as the *super node* for the purpose of the initial look-up of other peers containing AIS-systems. Apart from this, the nodes communicate with all other peers autonomously. The P2P overlay enables the collaborating nodes to exchange status information or detectors in the scenario of this paper. A general classification of P2P systems can be found in [10]. For further details regarding the CIDS, we refer to [9].

2.2 Classical Epidemic Model

Classical epidemic model for the spread of Internet worms [5] considers a group of homogeneously mixed susceptible (S), infected (I) and removed (R) nodes. A node is susceptible if it is not infected and if it has no proper protection for the attack. A node which is attacked by a worm (a.k.a., virus or any malicious code) becomes infected, and immediately starts to spread the infection to other susceptible nodes. An infected node can be recovered and immunized when proper protection mechanisms are enabled in which case the node becomes removed. Removed nodes can not be infected again. Suppose that at time t_i there are $S(t_i)$ susceptible, $I(t_i)$ infected and $R(t_i)$ removed nodes where $S(t_i) + I(t_i) + R(t_i) = N$. Let

$$s(t_i) = \frac{S(t_i)}{N},\ i(t_i) = \frac{I(t_i)}{N},\ r(t_i) = \frac{R(t_i)}{N}$$



(a) (b)

Fig. 1. Classical epidemic model for spread of Internet worms in a group of homogeneously mixed Susceptible (S), Infected (I) and Removed (R) nodes where $S(t_i) \ + \ I(t_i) \ + \ R(t_i) \ = \ N$ at time t_i. (1) Susceptible nodes get infected at the rate given by Equation 3 as they are contacted by the infected nodes. (2) Infected nodes are immunized and they become removed at the rate given by Equation 4. A removed node can not be infected again.

be the ratio of susceptible, infected and removed nodes respectively. Each contact in between susceptible and infected nodes will result in an infection. Therefore, in an interval $\triangle t$, there will be $\beta \ I(t) \ \frac{S(t)}{N} \ \triangle t$ contacts resulting in infection where β is the average number of contacts per infected node. Furthermore, infected nodes are removed from the system at a rate γ due to recovery or immunization. The number of removal at time interval $\triangle t$ is $\gamma \ I(t) \ \triangle t$. That is:

$$\triangle I(t) \ = \ \beta \ I(t) \ \frac{S(t)}{N} \ \triangle t \ - \ \gamma \ I(t) \ \triangle t, \tag{1}$$

$$\frac{di(t)}{dt} \ = \ \beta \ i(t) \ s(t) \ - \ \gamma \ i(t), \tag{2}$$

$$\frac{ds(t)}{dt} \ = \ -\beta \ i(t) \ s(t), \tag{3}$$

$$\frac{dr(t)}{dt} \ = \ \gamma \ i(t). \tag{4}$$

Epidemic can not build up if $[di(t)/dt]_{t_0} \ \leq 0$. That means, $\beta \ i(t) \ s(t) \ - \ \gamma \ i(t) \ \leq \ 0$ and $s(t) \ \leq \ \gamma/\beta$. The ratio γ/β is called *relative removal-rate* or *threshold density for susceptible nodes* and represented by $\rho \ = \ \gamma/\beta$. Figure 1 summarizes the overall model.

2.3 Balanced Incomplete Block Designs (BIBD)

A *BIBD* is an arrangement of v distinct objects into b blocks such that: (i) each object is in exactly r distinct blocks, (ii) each block contains exactly k distinct objects, (iii) every pair of distinct objects is in exactly λ blocks. The design is expressed as (v, b, r, k, λ) (a.k.a., (v, k, λ)) where: $b \cdot k \ = \ v \cdot r$ and $\lambda \cdot (v - 1) \ = \ r \cdot (k - 1)$. It is called *Symmetric BIBD* (a.k.a., *Symmetric Design* or SBIBD) when $b \ = \ v$ and $r \ = \ k$ [11] meaning that not only every pair of objects occurs in λ blocks but also every pair of blocks intersects on λ objects.

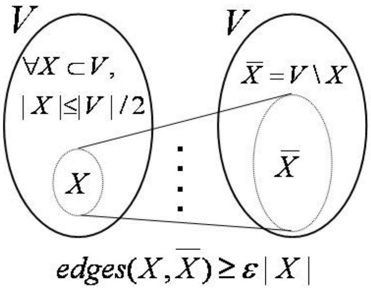
In this paper, we are interested in the *Finite Projective Plane* which is a subset of *Symmetric BIBD*. *Finite Projective Plane* consists of points (a finite set P of points) and lines (a set of subsets of P) of the *projective space* $PG(2, q)$ of dimension 2 and order q. For each prime power q where $q \geq 2$, there exists a *Finite Projective Plane* of order q [12, Theorem 2.10] with following four properties: (i) every line contains exactly $k = q + 1$ points, (ii) every point occurs on exactly $r = q + 1$ lines, (iii) there are exactly $v = q^2 + q + 1$ points, and (iv) there are exactly $b = q^2 + q + 1$ lines. Thus, a *Finite Projective Plane* of order q is a *SBIBD* with parameters $(q^2 + q + 1, q + 1, 1)$ [11].

We consider two methods to construct *SBIBD* of the form $(q^2 + q + 1, q + 1, 1)$. First method is *Difference Set Method* where the construction is done by simple modular addition operations on a *cyclic difference set.* A cyclic (v, k, λ) *difference set* $(mod\ v)$ is a set $B = \{b_1, b_2, \ldots, b_k\}$ of distinct elements in Z_v such that each one of the $(v-1)$ elements, say b, can be expressed in the form of difference $b = b_i - b_j\ (mod\ v)$ in λ different ways where $b_i, b_j \in B$ [11, Definition 2.1.1]. SBIBD blocks can be constructed by $B, B + 1, B + 2, \ldots, B + (v - 1)$ $(mod\ v)$ [11, Theorem 2.1.3] where $B + i = \{(b_1 + i)\ mod\ v, \ldots,\ (b_k + i)\ mod\ v\}$. For example, *difference set* $\{1, 2, 4\}$ can be used to generate $(7, 3, 1)$ SBIBD with blocks $\{1, 2, 4\}$, $\{2, 3, 5\}$, $\{3, 4, 6\}$, $\{4, 5, 7\}$, $\{5, 6, 1\}$, $\{6, 7, 2\}$, $\{7, 1, 3\}$ (0 is replaced with 7). Difference method provides a very efficient construction which can be used on low-resource devices if a cyclic different set for the target design is known. In fact, *cyclic difference sets* for small designs are listed in [11]. But, generating a cyclic difference set for a large design is not trivial [11, Theorem 2.5.2]. Second method is used for large designs where *complete set* of $(q - 1)$ *Mutually Orthogonal Latin Squares (MOLS)* are used to first construct an *affine plane* of order q which is a $(q^2, q, 1)$ design. The *affine plane* of order q is then converted into a *projective plane* of order q which is a $(q^2+q+1, q+1, 1)$ *SBIBD*. Construction can be done in $O(v^{3/2})$ time as described in [13] and references there in.

2.4 Finite Generalized Quadrangle (GQ)

A *Finite Generalized Quadrangle* $GQ(s, t)$ is a point-line incidence relation with following properties: (i) each point is incident with $t + 1$ lines $(t \geq 1)$ and two distinct points are incident with at most one line, (ii) each line is incident with $s + 1$ points $(s \geq 1)$ and two distinct lines are incident with (a.k.a., intersect on) at most one point, and (iii) if x is a point and L is a line not incident (I) with x, then there is a unique pair $(y, M) \in Points \times Lines$ for which $x\ I\ M\ I\ y\ I\ L$. In a $GQ(s, t)$, there are $v = (s+1)(st+1)$ points and $b = (t+1)(st+1)$ lines where each line includes $s + 1$ points and each point is incident with $t + 1$ lines. In this work, we are interested in $GQ(q, q)$ from projective space $PG(4, q)$. Probability that two lines intersect in $GQ(q, q)$ is given by the Equation 5.

$$P_{GQ} = \frac{t(s+1)}{(t+1)(st+1)} = \frac{q(q+1)}{(q+1)(q^2+1)} = \frac{q^2+q}{q^3+q^2+q+1} \approx \frac{1}{q}. \quad (5)$$


V
$\forall X \subset V,$
$|X| \leq |V|/2$
X
V
$\overline{X} = V \setminus X$
$\overline{X}$
$edges(X, \overline{X}) \geq \varepsilon |X|$

0	0	2	8	9	15	0
1	6	14	1	1	11	3
2	12	10	0	14	0	6
3	1	16	12	15	5	9
4	7	4	5	10	4	12
5	13	3	14	4	8	15
6	2	7	10	8	16	1
7	8	15	11	13	6	4
8	14	5	6	0	0	7
9	3	0	0	11	12	10
10	9	11	4	6	2	13
11	15	1	9	7	10	16
12	4	9	13	3	14	2
13	10	13	7	12	13	5
14	16	12	2	5	1	8
15	5	0	3	0	7	11
16	11	6	16	16	3	14

Fig. 2. Expander graphs are regular multi-graphs with expansion coefficient ϵ. Adjacency matrix for a sample Ramanujan Expander Graph (REG) $X^{5,17}$.

In $GQ(s,t) = GQ(q,q)$, there are $v = b = q^3 + q^2 + q + 1$ lines and points. Each line contains $s + 1 = q + 1$ points, and each point is incident with $t + 1 = q + 1$ lines. Consider $GQ(s,t) = GQ(2,2)$ for $q = 2$ as an example. There are 15 points $S = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\}$ and 15 lines $\{1,8,9\}$ $\{1,12,13\}$ $\{1,4,5\}$ $\{3,12,15\}$ $\{2,8,10\}$ $\{2,12,14\}$ $\{2,4,6\}$ $\{5,11,14\}$ $\{3,4,7\}$ $\{6,11,13\}$ $\{5,10,15\}$ $\{3,8,11\}$ $\{7,9,14\}$ $\{7,10,13\}$ and $\{6,9,15\}$ where each line contains $s + 1 = 3$ points and each point is incident with $t + 1 = 3$ lines. Note that lines $\{1,8,9\}$ and $\{3,12,15\}$ do not intersect but GQ provides three other lines intersecting with both: $\{1,12,13\}$, $\{3,8,11\}$ and $\{6,9,15\}$. $GQ(q,q)$ can be constructed from projective space $PG(4,q)$ with the canonical equation $Q(\boldsymbol{X}) = x_0^2 + x_1x_2 + x_3x_4 = 0$. Each point is a vector of the form $\boldsymbol{X} =< x_0, x_1, x_2, x_3, x_4 >$ in $GF(q)$, and each line contains $q{+}1$ bilinear points. $GQ(q,q)$ can be constructed in $O(v^2)$ time as described in [13] and references there in.

2.5 Ramanujan Expander Graphs (REG)

An *expander* is a *regular* multi-graph in which any subset of vertices has a large number of neighbors. It is highly connected, it has small diameter, small degree and many alternate disjoint paths between vertices. Formally, a graph $G = (V,E)$ is said to be *ϵ-edge-expander* if for every partition of the vertex set V into X and $\overline{X} = V \backslash X$, where $|X| \leq |V|/2$, the number of *cross edges* is $e(X,\overline{X}) \geq \epsilon|X|$. ϵ is called *expansion coefficient* ([14], [15], [16]). Almost all random bipartite graphs are expanders but it is possible to deterministically construct the best expanders with maximum possible *expansion coefficient* ϵ and with bounded vertex degrees. Figure 2 illustrates the expansion property.

There is a relation between expansion of a graph and eigenvalues of its adjacency matrix. A graph $G(V,E)$ with n vertices can be represented as an $(n \times n)$ *adjacency matrix* $A(G)$. The eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$ of $A(G)$ are the *spectrum* of graph G. Spectrum of a k-regular graph has the property that $\lambda_0 = k \geq \lambda_1 \geq \ldots \geq \lambda_{n-1}$. The relation between spectral gap (i.e., $\lambda_0 - \lambda_1$) and expansion coefficient ϵ is given in Equation 6 where larger spectral gap $(\lambda_0 - \lambda_1)$ implies higher expansion. *Ramanujan Expander Graphs* (REG) are asymptotically optimal and best known explicit expanders [17] where $|\lambda_i| \leq 2\sqrt{k-1}$ for $(1 \leq i \leq n-1)$. A REG $X^{s,t}$ is a k-regular graph with $N = t+1$ nodes for $k = s+1$ where s and t are primes congruent to 1 (mod 4). REG can be constructed in $O(st)$ time as described in [18].

$$\frac{k-\lambda_1}{2} \leq \epsilon \leq \sqrt{2k(k-\lambda_1)}, \tag{6}$$

3 Distributed Detector Set Generation

We consider the *Cooperative Intrusion Detection System* (CIDS) where N AIS-based IDS nodes collaborate over a peer-to-peer overlay network. Each IDS is assigned to a subspace in the feature space due to approaches in [1] and [2], and generates a set of detectors. Therefore, IDS nodes have *mutually exclusive sets* of detectors as proposed in [3]. Advantage of this approach is that for a fixed capacity of IDS nodes (i.e., χ detectors per nodes), it is possible to have an aggregated global detector database of $N\chi$ detectors meaning that the feature space coverage is maximized. However, when an attack starts spreading, only one IDS node will have the proper detector and this IDS will not be updating others until the attack reaches itself. Thus, detector sets on IDS nodes should have a level of overlap to be able to stop an attack spread in its earliest stage possible. In the following sections, we describe a probabilistic approach [3] and our three novel deterministic approaches which enable IDS nodes to create controlled level of detector set overlap by using techniques from combinatorial design theory and graph theory.

Algorithm 1 summarizes the decentralized detector generation scheme for a homogeneous network of N AIS-based (Artificial Immune System) IDS enabled nodes. Initially, each IDS node receives a unique ID (i.e., IDS_i) and gets assigned to a mutually exclusive subspace S_i in the feature space. In Algorithm 1 step 1, each IDS_i generates a set of detectors D_i of size d for the subspace S_i based on the approaches described in [9]. Next, through steps 2 to 14, each IDS_i decides on a list of node ID B_i for $|B_i| = k$ by using either one of the Random Graph (RG), Symmetric BIBD (SBIBD), Generalized Quadrangles (GQ) or Ramanujan Expander Graph (REG) based approaches. IDS_i contacts to each node $j \in B_i$ to get the detectors D_j for the subspace S_j through steps 15 to 18. IDS_i aggregates the received detectors $AD_i = \{\bigcup D_j | \forall j \in B_i\}$ where $|AD_i| = k \cdot d$ in step 19.

We assume that there exist a detector for the attack and initially only one node is infected (i.e., $I(t=0)=1$). Let β be the number of nodes an infected node can attack and γ be the number of nodes that the detector for the attack is sent per unit time. *Probability P(t) that infected nodes will not attack any node which has a proper detector at time t* is an important metric for spread of attack. Because, once an IDS with a proper detector is attacked, it will start distributing the detector at a rate γ as modeled in Section 4. *Feature space coverage* is another metric where we want to maximize coverage while minimizing probability $P(t)$. *Communication overhead* is our third metric to evaluate our approaches. It is the number and size of messages exchanged until every node has its target aggregated detector set.

Algorithm 1. Decentralized Detector Generation

```
Require:
    N {Total number of IDS enabled nodes},
    $IDS_i$ {ID of the IDS node running this algorithm},
    $S_i \subset Feature\ Space$ {Subspace assigned to $IDS_i$},
    d {Number of detectors generated from each subspace $S_i$},
    Algorithm {SBIBD, GQ, REG or RG}.
 1: $IDS_i$ generates the detector set $D_i$ ($|D_i| = d$) from the subspace $S_i$
 2: if Algorithm = Symmetric BIBD then
 3:     Generate $(v, k, \lambda)$-Design where $v = N = q^2 + q + 1$ and $k = q + 1$
 4:     $IDS_i$ selects the block $B_i$ ($|B_i| = k$)
 5: else if Algorithm = Generalized Quadrangles then
 6:     Generate $GQ(q, q)$-Design where $N = q^3 + q^2 + q + 1$ and $k = q + 1$
 7:     $IDS_i$ selects the block $B_i$ ($|B_i| = k$)
 8: else if Algorithm = Ramanujan Expander Graph then
 9:     Generate $X(s, t)$ expander where $N = t + 1$ and $k = s + 1$
10:     $B_i$ for $IDS_i$ is the neighbor list of $i^{th}$ node in $X^{s,t}$
11: else if Algorithm = Random Graph then
12:     $IDS_i$ selects $k = \log N$ IDS nodes
13:     $B_i$ for $IDS_i$ is the list of selected nodes
14: end if
15: for $i = 1$ to $|B_i|$ do
16:     $IDS_i$ contacts $IDS_j$ where $j \in B_i$
17:     $IDS_i$ receives the detector set $D_j$ from $IDS_j$
18: end for
19: Aggregated detector set for $IDS_i$ is $AD_i = \{\bigcup D_j | \forall j \in B_i\}$ where $|AD_i| = k \cdot d$
```

3.1 Random Graph $G(N, p)$ Based Approach

Mapping: Random graph approach is proposed by Goel *et al.* [3]. This approach employs graphs $G(N, p)$ due to Erdős-Rényi [4] where there are N nodes and each pair of nodes have a link in between with probability p. It is shown that each node should have $\log N$ neighbors for the graph $G(N, p)$ to be connected [19]. We define following mapping from random graphs to decentralized detector generation: nodes in graph $G(N, p)$ are mapped to IDS nodes and two nodes share a common detector set if there is an edge in between in the graph.

Construction: IDS_i (for $1 \le i \le N$) randomly selects $\log N$ IDS nodes. Selected nodes form the set B_i where it is possible that IDS_i selects itself (i.e., $i \in B_i$). IDS_i contacts each IDS_j for $j \in B_i$ and requests its detector set D_j. IDS_i ends up with the aggregated detector set $AD_i = \{\bigcup D_j | \forall j \in B_i\}$. Overall process is given through steps 11 to 14 of Algorithm 1. Graph $G(N, p)$ should be connected to make sure that a detector set D_j appears in at least two IDS nodes. A new coming node can randomly select $\log N$ nodes to receive their detector sets and $\log N$ nodes to send its detector set. A replacement node should only randomly select $\log N$ nodes to receive their detector sets.

Analysis: Each detector appears in average of $\log N$ nodes because each IDS node is contacted by average of $\log N$ nodes. Probability $P_{RG}(t)$ that infected nodes will not attack any node which has a proper detector at time t:

$$P_{RG}(t) = \left(1 - \frac{\log N}{N}\right)^{\beta I(t)} \tag{7}$$

For a fixed IDS node capacity of χ detectors, each node generates $\frac{\chi}{\log N}$ detectors and there are total of $\frac{N\chi}{\log N}$ distinct detectors. Each IDS contacts $\log N$ nodes and get contacted by $\log N$ nodes at average, *communication overhead* for each nodes is $2 \log N$ messages where each message includes $\frac{\chi}{\log N}$ detectors.

3.2 Ramanujan Expander Graph Based Approach

Problem with the random graph is its lack of regularity. Each node has average of $\log N$ neighbors (a.k.a., a detector appears in average of $\log N$ nodes) meaning also that some nodes may have less (a.k.a., some detectors may appear on less nodes) while some nodes may have more neighbors (a.k.a., some detectors may appear on more nodes). For the same node degree as $G(N, p)$, Ramanujan Expander Graph (REG) based deterministic approach provides best known expansion which means any subset of IDS nodes share detectors with the largest possible subset of IDS nodes. This property of REGs help REG-based approaches to provide better immunity against attacks when compared to RG-based approaches.

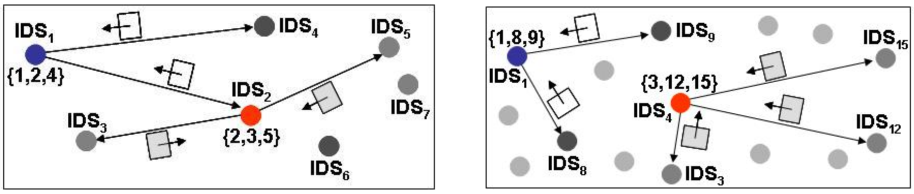
Mapping: A Ramanujan Expander Graph (REG) $X^{s,t}$ is a regular multi-graph where there are $N = t + 1$ nodes and where each node has $s + 1$ neighbors. We define following mapping from REGs to decentralized detector generation: nodes of the $X^{s,t}$ are mapped to IDS nodes and two nodes share a common detector set if there is an edge in between in $X^{s,t}$.

Construction: Each IDS_i deterministically generates the list of its neighbors B_i. IDS_i contacts each IDS_j for $j \in B_i$ and requests its detector set D_j. IDS_i ends up with the aggregated detector set $AD_i = \{\bigcup D_j | \forall j \in B_i\}$. Overall process is given through steps 8 to 10 of Algorithm 1. An additional node or a replacement node can generate the design and contact with its neighbors to exchange the detector sets.

Analysis: If we let the REG to have same degree as RG (i.e., $s + 1 = \log N$), it will have same probability $P_{REG}(t) = P_{RG}(t)$, feature space coverage and communication overhead. Advantage of the REG is its regularity and expansion property which provides better immunity against attacks when compared to RG-based approaches as we discuss in Section 4.

3.3 Combinatorial Design Based Approach

Mapping: We propose two novel approaches, Generalized Quadrangles (GQ) and Symmetric Balanced Incomplete Block Designs (SBIBD) from combinatorial



(a) (b)

Fig. 3. Decentralized detector generation. (a) Symmetric BIBD (SBIBD) based approach. (b) Generalized Quadrangle (GQ) based approach. Each block is assigned to an IDS node (IDS_1 uses block {1,2,4} in (a)) where each object is a distinct IDS node ID. Each IDS contacts the nodes in its block and collects the detector sets they generated.

design theory. Both SBIBD and GQ design assign v objects into b blocks ($v = b = q^2 + q + 1$ in SBIBD and $v = b = q^3 + q^2 + q + 1$ in $GQ(q, q)$) so that every pair of blocks have exactly one common object in SBIBD, and at most one common object in GQ design (probability of object share is given in Equation 5). We define following mapping from GQ and SBIBD to decentralized detector generation: each object is mapped to a distinct IDS node ID and each block is assigned to an IDS node (IDS_i uses the block B_i). Each IDS contacts the nodes in its block and collects the detector sets they generated. Thus, two IDS have common detectors if they have the same node ID in their blocks which is true for every pair of nodes in SBIBD based approach. Overall process is illustrated in Figure 3.

Construction: After the first step of Algorithm 1, each IDS_i generates the blocks in SBIBD or GQ. Assume that difference set method is used in SBIBD. Given a cyclic difference set B, node IDS_i can construct block B_i by simply performing $O(\sqrt{N})$ modular addition operations. Larger SBIBD can be constructed with MOLS in $O(N^{3/2})$ and GQ designs can be constructed in $O(N^2)$ as described in Section 2. Finally, IDS_i contacts each IDS_j for $j \in B_i$ and requests its detector set D_j. Aggregated detector set for IDS_i becomes $AD_i = \{\bigcup D_j | \forall j \in B_i\}$. Overall process is given through steps 2 to 7 of Algorithm 1. An additional node or a replacement node can generate the design and contact with its neighbors to exchange the detector sets.

Analysis: In SBIBD-based approach, each detector appears in exactly $q + 1$ nodes where $N = q^2 + q + 1$. Thus, probability $P_{SBIBD}(t)$ that infected nodes will not attack any node which has a proper detector at time t:

$$P_{SBIBD}(t) = \left(1 - \frac{q+1}{q^2+q+1}\right)^{\beta I(t)} \approx \left(1 - \frac{1}{\sqrt{N}}\right)^{\beta I(t)} \qquad (8)$$

For fixed IDS node capacity, say χ detectors, each node generates $\frac{\chi}{q+1} \approx \frac{\chi}{\sqrt{N}}$ detectors and there are $\sqrt{N}\chi$ distinct detectors. During detector exchange process, each IDS contacts $q+1 \approx \sqrt{N}$ nodes and gets contacted by $\sqrt{N}$ nodes. *Communication overhead* for each node is $2\sqrt{N}$ messages where each message includes $\frac{\chi}{\sqrt{N}}$ detectors. In the Generalized Quadrangle (GQ) based approach, each detector appears in exactly $q+1$ nodes where $N = q^3 + q^2 + q + 1$. Thus, probability $P_{GQ}(t)$ is:

$$P_{GQ}(t) = \left(1 - \frac{q+1}{q^3+q^2+q+1}\right)^{\beta I(t)} \approx \left(1 - \frac{1}{\sqrt[3]{N^2}}\right)^{\beta I(t)} \tag{9}$$

For fixed IDS node capacity of χ detectors, each node generates $\frac{\chi}{q+1} \approx \frac{\chi}{\sqrt[3]{N}}$ detectors and there are $\sqrt[3]{N^2}\chi$ distinct detectors. During detector exchange process, each IDS contacts $q+1 \approx \sqrt[3]{N}$ nodes and gets contacted by $\sqrt[3]{N}$ nodes. *Communication overhead* for each node is $2\sqrt[3]{N}$ messages where each message includes $\frac{\chi}{\sqrt[3]{N}}$ detectors.

4 Analysis and Comparisons

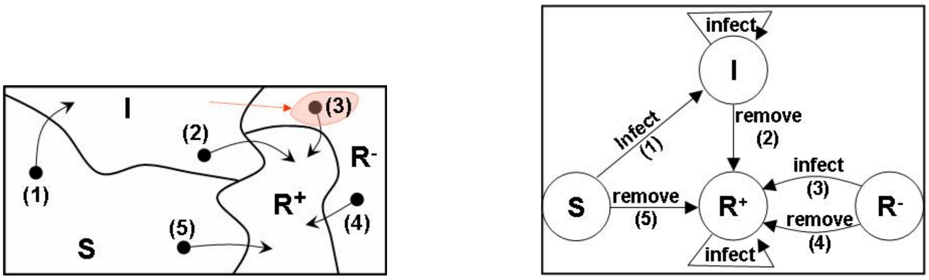
We enhance the classical epidemic model to analyze our decentralized detector distribution approaches in cooperative intrusion detection systems. We consider a homogeneous network of susceptible (S), infected (I), active removed (R^+) and passive removed (R^-) nodes where each node stores a subset of detectors and has the CIDS capability. A node is susceptible for a specific attack if it is not infected and if it does not have the detector for the attack. A susceptible node which is the target of the attack becomes infected, and it immediately starts to spread the attack to other susceptible nodes. A node is removed if it has the detector for the attack. Initially all removed nodes are passive removed nodes. A node becomes an active removed node and spreads the detector for the attack under following conditions: (i) when a passive removed node is attacked by an infected node, (ii) when an infected or susceptible node receives the detector from an active removed node, and (iii) when a passive node is contacted by an active removed node.

Suppose that at time t_i there are $S(t_i)$ susceptible nodes, $I(t_i)$ infected nodes, $R^+(t_i)$ active removed nodes and $R^-(t_i)$ passive removed nodes where $S(t_i) + I(t_i) + R^+(t_i) + R^-(t_i) = N$ and where $R^+(t_0) = 0$. Let

$$s(t_i) = \frac{S(t_i)}{N},\ i(t_i) = \frac{I(t_i)}{N},\ r^+(t_i) = \frac{R^+(t_i)}{N},\ r^-(t_i) = \frac{R^-(t_i)}{N}$$

be the ratio of susceptible, infected, active removed and passive removed nodes respectively. Each contact in between susceptible and infected nodes will result in an infection. Let β be the average number of contacts per infected node and γ be the average number of contacts per active removed node, in an interval $\triangle t$:

1. $\beta\ I(t)\ \frac{S(t)}{N}\ \triangle t$ susceptible nodes will be infected due to the attacks from infected nodes {flow (1) in Figure 4},
2. $\gamma\ R^+(t)\ \frac{I(t)}{N}\ \triangle t$ infected nodes will be removed due to the detector updates from active removed nodes {flow (2) in Figure 4},



(a) (b)

Fig. 4. Epidemic model for Cooperative Intrusion Detection Systems (CIDS) for group of homogeneously mixed susceptible (S), infected (I), active removed (R^+) and passive removed (R^-) nodes where $S(t_i) + I(t_i) + R^+(t_i) + R^-(t_i) = N$ at time t_i. (1) Susceptible nodes become infected when attacked by infected nodes, (2) infected nodes become removed nodes when updated by active removed nodes, (3) passive removed nodes become active removed nodes when attacked by infected nodes, (4) passive removed nodes become active removed nodes when updated by active removed nodes, and (5) susceptible nodes become active removed nodes when updated by active removed nodes.

3. $\beta\ I(t)\ \frac{R^-(t)}{N}\ \triangle t$ passive removed nodes will be active removed due to the attacks from infected nodes {flow (3) in Figure 4},
4. $\gamma\ R^+(t)\ \frac{R^-(t)}{N}\ \triangle t$ passive removed nodes will be active removed due to the detector updates from active removed nodes {flow (4) in Figure 4},
5. $\gamma\ R^+(t)\ \frac{S(t)}{N}\ \triangle t$ susceptible nodes will be active removed due to the detector updates from active removed nodes {flow (5) in Figure 4}.

Thus, infection and removal rates can be formulated as follows:

$$\triangle I(t) = \beta\ I(t)\ \frac{S(t)}{N}\ \triangle t - \gamma\ R^+(t)\ \frac{I(t)}{N}\ \triangle t, \tag{10}$$

$$\frac{di(t)}{dt} = \beta\ i(t)\ s(t) - \gamma\ r^+(t)\ i(t), \tag{11}$$

$$\frac{dr^+(t)}{dt} = \beta\ i(t)\ r^-(t) + \gamma\ r^+(t)\ \big(s(t) + i(t) + r^-(t)\big)\,. \tag{12}$$

Epidemic can not build up if $[di(t)/dt]_{t_i} \leq 0$ when $r^+(t_i) > 0$. The ratio $\frac{\gamma}{\beta}\ r^+(t)$ in Equation 13 is *relative removal-rate* or *threshold density for susceptible nodes.*

$$\beta\ i(t)\ s(t) - \gamma\ r^+(t)\ i(t) \leq 0, \quad s(t) \leq \frac{\gamma}{\beta}\ r^+(t). \tag{13}$$

In all approaches, SBIBD-based approach provides the largest overlap between detector sets and minimum probability $P(t)$ (Equations 7, 8 and 9) at the expense of increased communication overhead and decreased coverage of feature space. In addition to these, deterministic approaches have two advantages. First, unlike $G(N, p)$, deterministic approaches are based on the regular graphs due to SBIBD, GQ and REG. That means, each detector is replicated in equal number of IDS initially. Second, each node can generate the SBIBD, GQ and REG at very low cost. That means, all nodes know exactly which IDS has which detectors; moreover, they know $R^-(t_0)$. We can safely assume that, when an attack reaches to a node in $R^-(t_0)$ all others in $R^-(t_0)$ can be informed because such a broadcast update should have much lower overhead compared to attack and detector spread. Thus, a node in $R^+(t)$ does not need to update the nodes in $R^-(t_0)$ preventing the update collisions. Equation 10 then becomes:

$$\triangle I(t) \;=\; \beta\; I(t)\; \frac{S(t)}{N}\; \triangle t \;-\; \gamma_{det}\; R^+(t)\; \frac{I(t)}{N}\; \triangle t$$

where $\gamma_{det} = \gamma \frac{N}{N - R^-(t_0)}$ is the improved detector update rate due to the deterministic approaches. Problem can also be formulated as a spread of a detector on the logical graph due to underlying SBIBD, GQ and REG techniques. It is shown in [20] and [21] that epidemic spread threshold is related to underlying logical graph properties (i.e., node degree, diameter, spectral radius and largest eigenvalue of the adjacency matrix). More specifically, larger degree and spectral radius mean faster spread. SBIBD, GQ and REG provides regular graphs; in fact, SBIBD provides a complete graph which has the largest spectral radius (n where $N = n^2 + n + 1$).

5 Conclusions

We address the problem of self-organization and decentralized detector generation in Cooperative Intrusion Detection Systems (CIDS). We consider a set of AIS-based IDS nodes each of which is assigned to a distinct subspace of the feature space. Each IDS node generates a subset of the global detector set so that the coverage of the future space is maximized. Then, pairs of IDS nodes exchange detector sets to create controlled level redundancy so that the spread rate of the attack is limited. More specifically, our contribution is twofold. First, we use Symmetric Balanced Incomplete Block Design (SBIBD), Generalized Quadrangles (GQ) and Ramanujan Expander Graph (REG) techniques from combinatorial design theory and graph theory so that each node can independently decide how many and which detectors to exchange with which IDS nodes. Second, we applied classical epidemic model on both spread of attack and spread of detector, and showed that regular structures in deterministic techniques provides better immunity in a self-organized, self-adaptive and self-healing cooperative intrusion detection system when compared to probabilistic approaches.

References

1. Kim, J., Bentley, P.: The artificial immune model for network intrusion detection. In: EUFIT. 7th European Conference on Intelligent Techniques and Soft Computing (1999)

2. Gonzalez, F., Dasgupta, D.: Anomaly detection using using real-valued negative selection. In: Genetic Programming and Evolvable Machines (2003)
3. Goel, S., Bush, S.F.: Kolmogorov complexity estimates for detection of viruses in biologically inspired security systems: a comparison with traditional approaches. Complexity 9(2), 54–73 (2003)
4. Erdős, P., Rényi, A.: On random graphs. Publ. Math. Debrecen 6, 290–297 (1959)
5. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Review 42(4), 599–653 (2000)
6. Forrest, S., Perelson, A.S., Allen, L., Cherukuri, R.: Self-nonself Discrimination in a Computer. In: Proceedings of the IEEE Symposium on Research in Security and Privacy, pp. 202–212. IEEE Computer Society Press, Los Alamitos (1994)
7. Forrest, S., Hofmeyr, S.A., Somayaji, A., Longstaff, T.A.: A Sense of Self for Unix Processes. In: Proceedings of the 1996 IEEE Symposium on Security and Privacy, pp. 120–128. IEEE Computer Society Press, Los Alamitos (1996)
8. Hofmeyr, S., Forrest, S.: Architecture for an Artificial Immune System. Evolutionary Computation Journal 8(4), 443–473 (2000)
9. Luther, K., Bye, R., Alpcan, T., Muller, A., Albayrak, S.: A cooperative ais framework for intrusion detection. In: IEEE International Conference on Communications, IEEE Computer Society Press, Los Alamitos (2007)
10. Androutsellis-Theotokis, S., Spinellis, D.: A survey of peer-to-peer content distribution technologies. ACM Computing Surveys 36(4), 335–371 (2004)
11. Anderson, I.: Combinatorial designs: construction methods. Ellis Horwood Limited (1990)
12. Stinson, D.R.: Combinatorial designs: construction and analysis. Springer, Heidelberg (2004)
13. Camtepe, S.A., Yener, B.: Combinatorial design of key distribution mechanisms for wireless sensor networks. IEEE/ACM Transactions on Networking 15(2), 346–358 (2007)
14. Linial, N., Wigderson, A.: Expander graphs and their applications. Lecture Notes, Hebrew University, Israel (January 2003)
15. Linial, N.: Expanders, eigenvalues and all that. In: NIPS 2004 Talk (2004)
16. Govindaraju, R.: Design of Scalable Expander Interconnection Networks. PhD thesis, Rensselaer Polytechnic Institute, Troy, New York 12180, USA (1994)
17. Lubotzky, A., Phillips, R., Sarnak, P.: Ramanujan graphs. Combinatorica 8(3), 261–277 (1988)
18. Camtepe, S.A., Yener, B., Yung, M.: Expander graph based key distribution mechanisms in wireless sensor networks. In: IEEE International Conference on Communications, IEEE Computer Society Press, Los Alamitos (2006)
19. Xue, F., Kumar, P.R.: The number of neighbors needed for connectivity of wireless networks. Wireless Networks 10, 169–181 (2004)
20. Draief, M., Ganesh, A., Massoulié, L.: Thresholds for virus spread on networks. In: 1st International Conference on Performance Evaluation Methodolgies and Tools, p. 51 (2006)
21. Wang, Y., Chakrabarti, D., Wang, C., Faloutsos, C.: Epidemic spreading in real networks: An eigenvalue viewpoint. In: 22nd Symposium on Reliable Distributed Computing (2003)