This paper was presented as part of the Mini-Conference at IEEE INFOCOM 2011

Consensus Extraction from Heterogeneous Detectors to Improve Performance over Network Traffic Anomaly Detection

Jing Gao[†] Wei Fan[‡] Deepak Turaga[‡] Olivier Verscheure[‡] Xiaoqiao Meng[‡] Lu Su[†] Jiawei Han[†]
[†]University of Illinois at Urbana-Champaign
[‡]IBM T. J. Watson Research Center
[†]{jinggao3,lusu2,hanj}@illinois.edu, [‡]{weifan,turaga,ov1,xmeng}@us.ibm.com

***Abstract*—Network operators are continuously confronted with malicious events, such as port scans, denial-of-service attacks, and spreading of worms. Due to the detrimental effects caused by these anomalies, it is critical to detect them promptly and effectively. There have been numerous softwares, algorithms, or rules developed to conduct anomaly detection over traffic data. However, each of them only has limited descriptions of the anomalies, and thus suffers from high false positive/false negative rates. In contrast, the combination of multiple atomic detectors can provide a more powerful anomaly capturing capability when the base detectors complement each other. In this paper, we propose to infer a discriminative model by reaching consensus among multiple atomic anomaly detectors in an unsupervised manner when there are very few or even no known anomalous events for training. The proposed algorithm produces a per-event based non-trivial weighted combination of the atomic detectors by iteratively maximizing the probabilistic consensus among the output of the base detectors applied to different traffic records. The resulting model is different and not obtainable using Bayesian model averaging or weighted voting. Through experimental results on three network anomaly detection datasets, we show that the combined detector improves over the base detectors by 10% to 20% in accuracy.**

I. Introduction

In today's large-scale computer networks, an immense amount of flow data are observed each day, among which there are some records that do not conform to the normal network behavior. Some of them are malicious and can cause serious damage to the network. Therefore, it is important to sift through traffic data and detect anomalous events as they occur to ensure timely corrective actions. Although network anomaly detection has been widely studied [1], it remains a challenging task due to the following factors: 1) in large and complicated networks, the normal behavior can be multi-modal, and the boundary between normal and anomalous events is often not precise; 2) usually, the network attacks adapt themselves continuously to cheat the firewalls and security filters, making the anomaly detection problem more difficult; 3) previous known anomalies would soon be out of date and labeling current anomalies is expensive and slow, therefore, very few or even no labeled data are available for training or validation of anomaly detection techniques; and 4) network traffic data often contain noise which tends to be similar to the true anomalies, and it is difficult to remove them.

All in all, it's very hard for a single model to capture the network-wide traffic behavior. However, due to the complexity of network systems, each detector may only be partially effective. To address this problem, we propose to *combine anomaly detectors* generated by various definitions, formations and rules applied on each event to *gain better detection accuracy*. The benefits of combining anomaly detectors include: 1) it can reduce the chance of misclassifying noise and false alarms as anomalies by averaging out uncorrelated errors; 2) although the atomic detectors can be very simple and not quite effective, the combination can make them strong; 3) it can capture the diversified behavior of both normal and anomalous traffic when each of the detectors is good at detecting certain kinds of anomalies; and 4) the combination of multiple detectors is usually more robust than single detectors in a dynamic system when the traffic behavior continuously evolves.

In a dynamic network, it is impossible to manually label the anomalies continuously. Without labeled data, it is difficult to evaluate the performance of base detectors. In this case, a consensus solution would represent the best we can get from the base detectors. The strength of one detector usually complements the weakness of the other, and thus maximizing the agreement among them can significantly boost the detection accuracy. As an example, Figure I shows the results of six atomic detectors[1]. The first six rows show their predictions on a batch of 1000 traffic records and the last row provides the label, where 1 indicates an anomaly and 0 represents a normal point. Clearly, most of the base detectors have a high false positive rate, but the true anomalies usually occur at the points where the detectors agree with each other. From this example, we can see that the combined detector can identify the truth from different sources by consolidating their differences.

Typically, an atomic detector can be a simple rule, (e.g., number of port scans > 50), or a complicated learning algorithm, (e.g. PCA algorithm applied on traffic data). As long as multiple atomic detectors provide "complementary" expertise, we will gain from their consensus combination. Suppose we have a set of traffic traces $X = \{x_1, x_2, \ldots, x_n\}$ and k atomic detectors that predict each record in X either as an anomalous or a normal event. Our goal is to obtain a consolidated anomaly detection solution on X which agrees with the base detectors as much as possible. We refer to this problem as *consensus combination*. Different from existing alert fusion methods [2, 3], we propose to fuse the decisions of multiple anomaly detectors *without* guidance of any labels. To achieve this, a naive solution is to conduct a majority voting among the detectors, but it is far from being accurate because: 1) majority voting only considers consensus "locally" whereas we seek a

[1] This example is one of the data sets we conduct experiments on, i.e., the IDN data set. Please refer to Section IV for more details.

978-1-4244-9921-2/11/$26.00 ©2011 IEEE

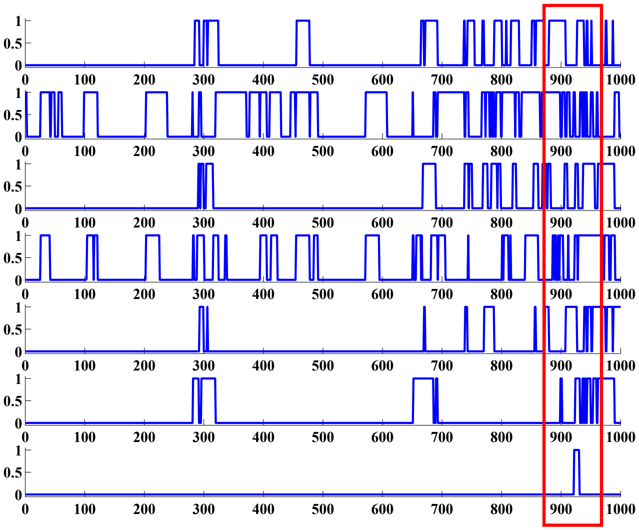

Fig. 1. Example of Atomic Anomaly Detectors

global optimal assignment of anomalous/normal labels for all the records in the collection; and 2) majority voting treats each base detector equally, but in fact the detection performance varies among detectors and a weighted combination of the base detectors is more precise.

In this paper, we formulate the problem as an optimization problem over a graph and solve it by iteratively propagating the detection information among neighboring nodes. Both the detectors' weights and the labels of records are automatically derived through the propagation. We show that the expected error incurred by randomly choosing a base detector is greater than that of combining them, and discuss how to select a good set of atomic anomaly detectors. We extend the algorithm to the online scenario where continuously arriving traffic data is processed. Moreover, when a few labeled anomalous and normal traffic traces are available, we can use such information to bias the process towards better joint predictions of the whole collection. We validate the performance of the proposed consensus combination algorithm on three real network traffic datasets. The theoretical analysis and encouraging experimental results suggest that consensus combination of multiple diversified detectors can reach a more accurate final decision for the task of network traffic anomaly detection.

II. Consensus Combination

Suppose we have the output of k detectors applied on the data set X with n records. We seek to find the solution that is the most consistent with the decisions of base detectors. Intuitively, the detector that agrees with the other detectors more often should be weighted higher in the voting. To this end, we propose to automatically derive the importance of anomaly detectors as well as the probability of each record being an anomaly through an iterative propagation process.

We first split the records into two clusters based on each detector A_r's predictions on X, and thus there are totally $s = 2k$ clusters. The cluster with index $2r-1$ $(r = 1, \ldots, k)$ contains the records that are predicted to be anomalies by A_r, whereas the cluster with index $2r$ $(r = 1, \ldots, k)$ has the records that are normal in the decisions of A_r. Consider an example where $X = \{x_1, \ldots, x_7\}$. For the sake of simplicity, we only consider the output of two detectors: $A_1 = \{1,0,1,1,0,0,0\}$ and $A_2 = \{0,1,0,1,0,0,0\}$ where 1 denotes a predicted anomaly and 0 represents a normal record. Four clusters are formed, where c_1 and c_3 contain the records that are predicted to have label 1 by A_1 and A_2 respectively. On the other hand, records in c_2 and c_4 are predicted to have label 0. Then we represent the records and the clusters in a graph as shown in Figure 2, where a cluster links to all the records it contains. As can be seen, the decisions of all the base detectors are summarized losslessly in this graph, or equivalently, the affinity matrix W of the graph:

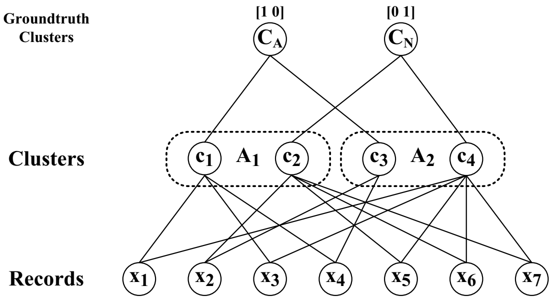

Fig. 2. Graph Representation

$$w_{ij} = 1 \; x_i \text{ assigned to cluster } c_j \text{ by a detector;} \quad 0 \text{ otherwise.}$$

Let Y denote the variable of true label where $Y = 1$ for an anomaly and $Y = 0$ otherwise. We aim at estimating the probability of each record x_i being an anomaly $\hat{P}(Y = 1|x_i)$. As a nuisance parameter, the probability of each record x_i being normal $\hat{P}(Y = 0|x_i)$ is also estimated. Therefore, each record x_i is associated with a two-dimensional probability vector $\vec{u}_{i\cdot} = \Big(\hat{P}(Y = 1|x_i), \hat{P}(Y = 0|x_i)\Big)$. On the other hand, atomic detectors could make mistakes, and thus each cluster is actually a mixture of normal and anomalous events. Therefore, we also associate each cluster c_j with a probability vector: $\vec{v}_{j\cdot} = \Big(\hat{P}(Y = 1|c_j), \hat{P}(Y = 0|c_j)\Big)$.

In Figure 2, we introduce a "groundtruth" detector with two clusters having probability vectors $(1, 0)$ (anomalous cluster C_A) and $(0, 1)$ (normal cluster C_N) respectively. We connect each cluster of the base detectors to the corresponding "truth" cluster by an edge with weight α, where α represents our confidence in the base detectors' predictions. In the aforementioned example, c_1 and c_3 are connected to C_A as anomalous clusters, whereas c_2 and c_4 are linked to C_N. Although the base detectors make mistakes, we assume that at least their predictions should not be flipped. For example, the probability vector of c_1 could be (0.8,0.2), but is unlikely to be (0.1,0.9). Such constraints are encoded in the groundtruth probability matrix $F_{s\times 2} = (\vec{f}_{1\cdot}, \ldots, \vec{f}_{s\cdot})^T$, where

$$\vec{f}_{j\cdot} = \begin{cases} (1,0) & j = 2r-1 \ (r = 1, \ldots, k) \\ (0,1) & j = 2r \ (r = 1, \ldots, k) \end{cases}$$

Actually, the probability vector of each cluster in F is the probability vector of the groundtruth cluster it links to. Based on the graph, reaching consensus among the base detectors is defined as the following optimization problem:

$$\min_{\vec{u}_{i\cdot},\vec{v}_{j\cdot}} \quad \Big(\sum_{i=1}^{n}\sum_{j=1}^{s} w_{ij}||\vec{u}_{i\cdot}-\vec{v}_{j\cdot}||^2 + \alpha\sum_{j=1}^{s}||\vec{v}_{j\cdot}-\vec{f}_{j\cdot}||^2\Big) \quad (1)$$

$$\text{s.t.} \quad \vec{u}_{i\cdot} \geq \vec{0}, \quad |\vec{u}_{i\cdot}| = 1 \quad \text{for} \quad i = 1:n \quad (2)$$

$$\vec{v}_{j\cdot} \geq \vec{0}, \quad |\vec{v}_{j\cdot}| = 1 \quad \text{for} \quad j = 1:s \quad (3)$$

where $||.||$ and $|.|$ denote a vector's L2 and L1 norm respectively. In the graph, C_A and C_N have fixed probability vectors, which only act as constraints. The first term in the objective function (Eq. (1)) ensures that a record x_i has similar probability vector as the cluster to which it belongs. The second term puts the constraint that a cluster c_j's probability vector should not deviate much from the corresponding groundtruth assignment. For example, c_1's conditional probability needs to be close to that of the groundtruth cluster C_A, as well as that of the records it links to: x_1, x_3 and x_4. $\vec{u}_{i\cdot}$ and $\vec{v}_{j\cdot}$ are probability vectors, so each component must be greater than or equal to 0 and the sum equals 1.

We propose to solve this problem using the block coordinate descent method. At the t-th iteration, we set the partial derivatives to 0 and obtain the unique global minimum of the cost function with respect to $\vec{u}_{i\cdot}$ and $\vec{v}_{j\cdot}$ respectively:

$$\vec{u}^{\,(t)}_{i\cdot} = \frac{\sum_{j=1}^{s} w_{ij}\vec{v}^{\,(t)}_{j\cdot}}{\sum_{j=1}^{s} w_{ij}} \quad \vec{v}^{\,(t+1)}_{j\cdot} = \frac{\sum_{i=1}^{n} w_{ij}\vec{u}^{\,(t)}_{i\cdot} + \alpha\vec{f}_{j\cdot}}{\sum_{i=1}^{n} w_{ij} + \alpha} \quad (4)$$

We start by initializing the probability of clusters (i.e.,$\vec{v}_{j\cdot}$) using the groundtruth probability $\vec{f}_{j\cdot}$. Then the clusters first propagate the information to their neighboring records when updating $\vec{u}_{i\cdot}$. In turn, to update $\vec{v}_{j\cdot}$, the clusters receive the information from neighboring records and try to retain initial values. It is straightforward to prove that the objective function is convex and thus $\vec{u}_{i\cdot}, \vec{v}_{j\cdot}$ converges to the global minimum [4]. Finally, the consensus probability of each record being an anomaly can be found in $\vec{u}_{i1}$. In some sense, $\vec{v}_{j\cdot}$ represents the weights assigned to the base detector. For a detector A_r, if the conditional probabilities of its two clusters c_{2r-1} and c_{2r} are rather skewed, then A_r's predictions will cast important votes in determining the probabilities of the records they link to. On the other hand, if the two clusters of A_r have uniform probabilities, A_r is not informative and has a lower weight. Note that although the example discussed in this section only has binary output, the proposed framework can be applied on probabilistic output by utilizing the probability as edge weight of the bipartite graph in the optimization framework.

Performance Analysis. Suppose each detector A outputs an estimated probability $P(y|x,A)$ ($y = 0$ or 1) for a record x, and the chance of picking A is $P(A)$. The true conditional and joint probability of a record is $P(y|x)$ and $P(x,y)$ respectively. The expected error incurred by randomly choosing an atomic detector is the error integrated over all detectors and all records: $Err^S = \sum_A P(A)\big[\sum_{(x,y)} P(x,y)(P(y|x) - P(y|x,A))^2\big]$. On the other hand, consensus combination takes the expectation of the predicted probabilities over all the base detectors as the final output, and its expected error is the error integrated over all the records: $Err^C = \sum_{(x,y)} P(x,y)(P(y|x) - \sum_A P(A)P(y|x,A))^2$. Here $P(A)$ can be regarded as the consensus weight of detector A. By comparing them, we find that the expected error of a combined detector is always less than or equal to the expected error of randomly selecting an atomic detector, i.e., $Err^C \leq Err^S$. The proof is omitted due to space limit. Note that we are not claiming that the consensus detector is more accurate than any single detector all the time. It is possible that there exists a very good atomic detector which has much better performance than the others, and thus is also better than the combined detector. However, in dynamic network systems, the traffic behavior always evolves over time, without being known by the operators. It is nearly impossible for any atomic detector to perform consistently well and stand out as the best one. Therefore, we need to utilize the combination of detectors to reduce the risk because it can achieve the best *expected* performance. The base detectors can be simple rules, or anomaly detection algorithms. Many studies in ensemble learning have shown that the diversity among base detectors can greatly improve the combination performance [5, 6]. Intuitively, when we use some rules or algorithms that capture different aspects of the traffic data, the base detectors would be diversified. In fact, we can use some correlation measures to explore the diversity degree of detectors based on the predictions they made [6].

III. Extensions

In this section, we propose an online component for the anomaly detection system as well as extend the algorithm to semi-supervised learning scenarios.

Incremental Consensus Combination. Suppose we have a batch of traffic data $X = \{x_1, \ldots, x_n\}$, a collection of s clusters obtained from $s/2$ anomaly detectors $C = \{c_1, \ldots, c_s\}$ and the affinity matrix W which contains their links. Suppose we have estimated the probabilities of each object (or cluster) being anomalous and normal by the consensus combination algorithm. Now given some continuously arriving traffic records $x_{n+1}, x_{n+2}, \ldots$, we aim at calculating the anomalous probability of each new record as well as updating the weights of detectors based on the new information.

From Eq. (4), we can see that the proposed algorithm can be adapted to the stream scenario. We can rewrite the update equation for $\vec{v}_{j\cdot}$ as follows:

$$\vec{v}_{j\cdot} = \frac{(\sum_{i=1}^{n-1} w_{ij}\vec{u}_{i\cdot} + \alpha\vec{f}_{j\cdot}) + w_{nj}\vec{u}_{n\cdot}}{(\sum_{i=1}^{n-1} w_{ij} + \alpha) + w_{nj}} \quad (5)$$

where both the numerator and the denominator can be split into two parts: the summation over the $n-1$ records, and the information carried by x_n. Therefore, the update can be done incrementally by storing the summation over the existing records. When each new record x_m arrives, we first let the base detectors make predictions, and the links from x_m to the clusters C are represented by the vector $\vec{w}_{m\cdot} = (w_{m1}, \ldots, w_{ms})^T$. The probability vector of x_m, i.e., $\vec{u}_{m\cdot}$, is computed as the average of the probability vectors of the clusters it links to.

TABLE I
AREA UNDER ROC CURVE

	IDN			DARPA			LBNL
	1	2	3	1	2	3	1
WB	0.5269	0.2832	0.3745	0.5804	0.5930	0.5851	0.5005
BB	0.6671	**0.8059**	0.8266	0.6068	0.6137	0.6150	0.8230
AB	0.5904	0.5731	0.6654	0.5981	0.6021	0.6022	0.7101
MV	0.7089	0.6854	0.8871	0.7765	0.7865	0.7739	0.8165
UC	0.7255	0.7711	0.9076	0.7812	0.7938	0.7796	0.8180
SC	0.7204	0.8048	0.9089	**0.8005**	**0.8173**	**0.7985**	**0.8324**
IC	**0.7270**	0.7552	**0.9090**	0.7730	0.7836	0.7727	0.8160

Then we add $w_{mj}\vec{u}_{m\cdot}$ and w_{mj} to the numerator and the denominator respectively in the updating equation of $\vec{v}_{j\cdot}$ (Eq. (5)). After that, we update the probabilities of clusters, and in turn, update the probabilities of the previous records based on the new probabilities of the clusters according to Eq. (4).

Semi-supervised Consensus Combination. Suppose we now know the labels of a few traffic records, i.e., whether they are normal or anomalous. We can incorporate such labeled information into the consensus combination process to gain better performance. Suppose the labeled information is encoded in an $n \times 2$ matrix Z:

$$\vec{z}_{i\cdot} = \begin{cases} (1,0) & x_i \text{ is observed as an anomaly,} \\ (0,1) & x_i \text{ is observed as a normal record,} \\ (0,0) & x_i \text{ is unlabeled.} \end{cases}$$

We add an additional term $\beta \sum_{i=1}^{n} q_i ||\vec{u}_{i\cdot} - \vec{z}_{i\cdot}||^2$ to the objective function in Eq. (1) to penalize the deviation of the probabilities from the observed labels where $q_i = z_{i1} + z_{i2}$. When x_i is labeled, $q_i = 1$, so we impose the constraint that a record x_i's estimated label should be close to its observed label. The cost paid for violating the constraint is β, which represents our confidence in the correctness of known labels. To update the conditional probabilities of each record, we incorporate its prior labeled information, which in turn affect the probabilities of clusters:

$$\vec{u}_{i\cdot}^{(t)} = \frac{\sum_{j=1}^{s} w_{ij}\vec{v}_{j\cdot}^{(t)} + \beta q_i \vec{z}_{i\cdot}}{\sum_{j=1}^{s} w_{ij} + \beta q_i} \qquad (6)$$

In this way, the labeled information will be propagated over the graph iteratively.

IV. EXPERIMENTS

We evaluate the performance of the proposed algorithm on three network anomaly detection data sets.

IDN. We have employed an intrusion detection network (IDN) at IBM to empirically test the performance of the proposed approach. IDN provides security services to the network infrastructure by analyzing raw traffic packets and outputting a large number of events, such as DOS flooding, SYN flooding and port scanning, etc. We use two high level measures to describe the probability of observing events during each interval. An atomic detector can be defined based on each of the measures by setting a threshold. In the experiments, we collect three data sets, each of which has 1000 intervals, and combine six anomaly detectors.

DARPA. MIT Lincoln Labs set up an environment to acquire nine weeks of raw TCP dump data for a local-area network. The raw data were processed into about several million connection records, where each contains a sequence of TCP packets transmitted during a time interval. Each connection is labeled as either a normal record, or an attack. The original data set is from DARPA 1998 intrusion detection program and some high level features are defined to help distinguish normal connections from attacks[2]. We randomly extract three subsets of executive records, each with size 1832. To generate diversified detectors, we randomly select 2-5 features each time, apply LOF anomaly detection algorithm [7] on the selected subset of features, and repeat 20 times.

LBNL. Another dataset is an enterprise traffic dataset[3] collected at the edge routers of the Lawrence Berkeley National Lab (LBNL). The traffic in this dataset comprises background traffic (normal) and scanner traffic (malicious). The labels of the packets are provided by the data collector [8]. We aggregate the packet traces by intervals spanning 1 minute (3704 intervals in total), and the anomalous score of each interval is computed as the percentage of anomalous packets among all the packets of the interval. We calculate the following metrics for each interval: 1) number of TCP SYN packets, 2) number of distinct IPs in the source, 3) number of distinct IPs in the destination, 4) maximum number of distinct IPs an IP in the source has contacted, 5) maximum number of distinct IPs an IP in the destination has contacted, and 6) maximum pairwise distance between distinct IPs an IP has contacted. An atomic detector is applied on each metric to predict the records with highest or lowest values (around 20%) as anomalies. We then combine the six detectors and compare the performance.

Evaluation Measures and Baselines. The ROC curve, which represents the trade-off between detection rate and false alarm rate, is widely used for anomaly detection evaluation. The area under ROC curve (AUC), which is in the range [0,1], is a good evaluation metric. A good algorithm would produce an ROC curve as close to the left-top corner as possible, which has an AUC value close to 1. We show the following baseline performance: 1) the worst, best and average AUC values of the base detectors, denoted as **WB**, **BB**, and **AB** respectively; and 2) the AUC obtained by majority voting among detectors (**MV**). Accordingly, we evaluate the unsupervised (**UC**), semi-supervised (**SC**) and incremental (**IC**) consensus combination algorithms. On each data set, we run the experiments 100 times and compute the *average* AUC of all the baseline methods and the proposed algorithms. At each run, 2% records are selected to be labeled records, which work as additional input to the semi-supervised algorithm. The remaining records are used as the test bed for all the methods. For incremental consensus combination algorithm, we select 30% of the unlabeled records as batch data for initialization, and the remaining 70% of the unlabeled records are processed incrementally.

Detection Accuracy As shown in Table I, the proposed consensus combination algorithm improves the anomaly detection performance. It can be observed that the base detectors usually have large variabilities in their abilities of detecting

[2] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[3] http://www.icir.org/enterprise-tracing/

anomalies. Majority voting, which leverages the decisions of the atomic detectors, beats the individual detectors most of the time. However, more improvement can be achieved by using the probabilistic consensus combination among all the atomic detectors. Instead of considering each base detector to be equally capable of detecting anomalies as in majority voting, the consensus combination approach computes the weighted combination of the detectors' decisions, where a detector has a higher weight if it agrees with the other detectors more often. Using the proposed algorithm, consistent improvements can be obtained. For example, on one of the DARPA datasets, the AUC of the best base detector is only 0.6137, whereas the consensus combination approach improves it to 0.7938. By incorporating only 2% labeled records, the semi-supervised consensus combination further improves the anomaly detection accuracy. Due to lack of global information, the incremental algorithm usually is not as good as the batch consensus combination, but is still much better than the atomic detectors.

V. Related Work

Anomaly detection [1] has received considerable attention in the field of network traffic analysis. Existing methods include statistical modeling [9, 10], Principle Component Analysis (PCA) [11], information-theoretic measures [12], wavelet methods [13], Kalman filtering [10], and data mining approaches [14, 15]. Ensemble methods have emerged as a powerful tool for improving robustness and accuracy of both supervised and unsupervised solutions [5, 16, 17]. The basic idea is to combine multiple competing models into a committee to reduce uncorrelated errors. Researchers have studied how to combine multiple intrusion detection systems into a general system. The developed methods mainly fall into two categories: alert correlation and alert fusion. First, alert correlation has a different objective, which aims at creating a general view of the attacks, e.g., attack scenarios or graphs, based on the various types of low-level alerts generated from multiple intrusion detection systems [18, 19]. Second, alert fusion methods deal with combination of alerts, each of which represents independent detection of the same attack occurrence. Specifically, earlier approaches adopt the idea of multi-sensor data fusion [20] and later, alert fusion is mainly solved through ensemble of classifiers [2, 3]. Note that in the machine learning and statistical inference community, many algorithms, including bagging, boosting, Bayesian averaging and random forests, have been developed to build ensemble classifiers [5, 21]. However, these approaches require labeled data for training, and it is usually unrealistic to obtain plenty of training data for network anomaly detection tasks. In such scenarios, combining classifiers is infeasible, whereas the proposed consensus combination method can combine the detectors in an *unsupervised* way.

VI. Conclusions

Network anomaly detection is a challenging task due to heterogeneous nature of the network. Automatically extracting and analyzing network anomalies from immense amount of traffic data is difficult, and thus atomic anomaly detectors could have very low detection rate. This paper has demonstrated the advantages of combining various anomaly detectors through consensus optimization. We summarize the decisions of base detectors in a graph, and maximize the consensus by promoting smoothness of label assignment over the graph. The problem is solved by propagating information between cluster and record nodes iteratively. We also extend the algorithm to semi-supervised learning and incremental learning cases. On three real network anomaly detection datasets, the proposed algorithm improves the detection accuracy of the base detectors by 10% to 20%.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," University of Minnesota, Tech. Rep. 07-017, 2007.

[2] G. Gu, A. A. Cárdenas, and W. Lee, "Principled reasoning and practical applications of alert fusion in intrusion detection systems," in *Proc. ACM ASIACCS*, 2008, pp. 136–147.

[3] I. Corona, G. Giacinto, and F. Roli, *Intrusion Detection in Computer Systems Using Multiple Classifier Systems*. Springer, 2008, pp. 91–113.

[4] D. P. Bertsekas, *Non-Linear Programming*, 2nd ed. Athena Scientific, 1999.

[5] G. Seni and F. Elder, *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool, 2010.

[6] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proc. ACM SIGMOD*, 2000, pp. 93–104.

[8] R. Pang, M. Allman, V. Paxson, and J. Lee, "The devil and packet trace anonymization," *SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 29–38, 2006.

[9] H. Wang, D. Zhang, and K. G. Shin, "Detecting syn flooding attacks," in *Proc. IEEE INFOCOM*, 2002.

[10] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proc. ACM IMC*, 2005, pp. 331–344.

[11] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *Proc. ACM SIGCOMM*, 2005, pp. 217–228.

[12] Y. Gu, A. McCallum, and D. Towsley, "Detecting anomalies in network traffic using maximum entropy estimation," in *Proc. ACM IMC*, 2005, pp. 345–350.

[13] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in *Proc. ACM IMW*, 2002, pp. 71–82.

[14] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, "Using artificial anomalies to detect unknown and known network intrusions," *Knowl. Inf. Syst.*, vol. 6, pp. 507–527, 2004.

[15] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," in *Proc. Mobihoc*, 2007, pp. 219–228.

[16] A. Strehl and J. Ghosh, "Cluster ensembles – a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[17] J. Gao, F. Liang, W. Fan, Y. Sun, and J. Han, "Graph-based consensus maximization among multiple supervised and unsupervised models," in *Proc. NIPS*, 2009.

[18] F. Cuppens and A. Miège, "Alert correlation in a cooperative intrusion detection framework," in *Proc. IEEE S & P*, 2002, pp. 202–215.

[19] H. Debar and A. Wespi, "Aggregation and correlation of intrusion-detection alerts," in *Proc. RAID*, 2001, pp. 85–103.

[20] T. Bass, "Intrusion detection systems and multisensor data fusion," *Communications of ACM*, vol. 43, no. 4, pp. 99–105, 2000.

[21] T. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Systems*, 2000, pp. 1–15.