Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers

D I Parfenov¹, I P Bolodurina¹, M A Lapina²

¹Orenburg State University, 13, Pobedy Ave., Orenburg, 460018, Russia

²North-Caucasus Federal University, 1, Pushkin St., Stavropol, 355017, Russia

E-mail parfenovdi@mail.ru

Abstract. In the framework of this study, a technical solution was developed that makes it possible to detect network security incidents with a high probability using data arrays about device statuses, network events, and information stored in system logs. A model for identifying attacks on a network has been developed, using behavioral analysis and allowing the identification of suspicious network activity. An algorithmic solution has also been built that allows aggregating data in a single store based on Cassandra and correlating events from specified sources using gradient boosting of decision trees in the CatBoost implementation. During the computational experiment, the study of the proposed hybrid solution for the accuracy of identification of individual types of attacks was conducted. It is proved that the proposed approach can effectively detect and repel attacks by reducing the response time to security incidents.

Keywords: Security incidents; intrusion detection; network of telecommunication service providers; network monitoring

1. Introduction

The volume of services provided by providers using publicly available data networks is increasing annually. The modern borderless is the main network architecture for the organization of the technology of the Internet of things (IoT). In turn, the borderless network architecture allows you to associate each object or device with network capabilities. Today, devices included in such a network can have completely different purposes, from household items and video cameras to automation systems, sensors, medical equipment, and other vital real-time systems [1]. Ensuring network security in such difficult conditions is one of the main tasks of telecommunication service providers. At the same time, I take into account the flexible and fairly dynamic topology, as well as the wide range and convergent nature of the services provided, the issue of trust between devices when sharing communication channels is no less important. It is important for users to be sure that the data network to which their network devices are connected has a fairly high degree of security. A feature of modern communication networks is the
presence of many heterogeneous devices that use a variety of data transfer protocols. Each protocol has its own mechanisms and security measures.

Traditionally, to protect the network perimeter, telecom operators use approaches based on the use of intrusion detection systems (IDS) and intrusion prevention systems (IPS), as well as firewalls [2]. Such methods are effective if the attack on the network comes from the external perimeter. However, as noted earlier, in the modern conditions in which telecom operators work, the boundaries of networks are blurred. In such circumstances, each client network device is a potential source of cyber attacks. In such circumstances, traditional control systems are not effective enough.

Generally, there are three basic approaches to detecting attacks in the networks.

The most common are signature-based approaches. Signature intrusion detection systems detect attacks when the behavior of a system or network matches the attack profile stored in their own IDS databases. If any activity of the system or network matches the saved templates, a warning will be issued. Signature-based intrusion detection systems are accurate in detecting known threats, and their technique is easy to understand. However, this approach is ineffective for detecting new attacks and other variants of known attacks, since a matching signature for these attacks has not yet been established.

Anomaly-based approaches are also quite common. They allow you to almost instantly compare the current traffic profile with the profile of normal behavior. This approach is effective in detecting new attacks, in particular those attacks that involve resource abuse. A significant drawback of this behavior is that any changes in traffic profiles, including legitimate ones, are considered an intrusion. Therefore, using exclusively this approach to detect attacks has a high level of false positives. To build a normal behavior profile, statistical methods or machine learning algorithms are usually used.

Specification-based approaches typically include groups of rules and thresholds that determine the expected behavior for network elements, such as nodes, protocols, and routing tables. Specification-based approaches detect intrusions when network behavior deviates from specification definitions. Therefore, specification-based detection has the same purpose as the anomaly-based approach: detecting deviations from normal behavior. The key difference between these two approaches is the following: in the approaches based on specifications, a human specialist must manually determine the rules for each specification. Manual specifications typically provide lower false positives than anomaly-based detection. In addition, specification-based detection systems do not require a training phase, as they can start working immediately after setting up the specification. However, manually defined specifications may not adapt to different environments and may be time consuming and error prone.

Thus, it has been established that existing trips have a number of significant drawbacks, and today there is no comprehensive solution providing a mechanism for organizing the protection of more devices from a single point. Therefore, in the framework of this study, it is proposed to use a hybrid solution, which includes: the analysis of network traffic and the search for abnormal activity on network devices, as well as the analysis of data logs generated on network equipment. Such an approach will allow using ideas of detection based on signatures, specifications and anomalies, which will maximize the advantages of these approaches and minimize their disadvantages.

2. Related work

Scientists around the world offer various options for approaches to detecting and identifying attacks in networks.

As noted earlier, existing signature mechanisms are not enough to identify zero-day attacks. Researchers [3] offer an approach to combat this type of attack, based on the use of a modified support vector system (SVM) method. To increase the efficiency of Enhanced SVM, it preliminarily forms a profile of legitimate traffic, and for training SVM, a self-organizing characteristic map (SOFM) is used. The authors also apply a packet filtering scheme based on a passive TCP / IP fingerprint to cut off packets of the same type. In addition, the authors use a feature selection method using a genetic algorithm to obtain optimized information from raw traffic.

Study authors [4] propose the construction of two-tier models for classifying traffic anomalies based on Naïve Bayes in versions of KNN classifiers with the determination of the confidence coefficient, as
well as linear discriminant analysis to reduce the dimension. Experimental results with the NSL-KDD dataset have shown significant gains in the speed of anomaly detection by reducing computational time.

Modern networks change dynamically over time and to adapt security systems to work in such conditions, a number of researchers use deep learning methods. In article [5] researchers proposed models for detecting anomalies based on various structures of deep neural networks. Training samples from the public datasets NSLKDD, NSLKDDTest + and NSLKDDTest21 were used as the main data set. A feature of the computational experiment is the use of a GPU, which significantly accelerates the learning process of the proposed model. A comprehensive review of deep learning methodologies for identifying anomalies in networks was conducted by the authors of the study [6]. They conducted extensive comparative computing experiments to compare the efficiency and accuracy of network traffic analysis.

As part of the study [7] the authors propose a new approach to identifying anomalies based on the center of the cluster and the nearest neighbor (CANN). The experimental results of the authors show that the CANN classifier works better in terms of classification accuracy, detection frequency and false positives.

As part of the study [8] the authors propose a model of the support vector machine (SVM), combining the analysis of the kernel principal component analysis (KPCA) with the genetic algorithm (GA). The authors propose using the multilayer SVM classifier as the basic classifier for attack / non-attack. At the same time, KPCA is used as a preprocessor to reduce the dimension of feature vectors and reduce training time. The genetic algorithm is used to optimize penalty functions. Experimental results show that the proposed model provides a higher accuracy of forecasting, a higher rate of convergence.

Modern network intrusion detection systems must be able to handle large and rapidly changing data. Therefore, the authors of the study [9] propose the use of ensemble solutions and a distributed approach and implementation of intrusion detection systems. For such systems, data exchange between distributed nodes is critical. The authors review the studies related to the performance of such solutions.

To search for attacks, the authors of the study [10] propose a distributed environment for extracting functions and preparing data for extracting functions from raw network traffic. The results are used to calculate during clustering, and anomalies are detected based on the outlier search.

In view of the high dynamics of changes in modern networks, the prediction of the development of attacks in priority. The authors of the study [11] propose a mechanism for detecting anomalies based on the statistical procedure, analysis of the main components, and metaheuristics of optimization of the ant colony. Based on the data obtained, the main traffic characteristics are obtained, from which they subsequently form a profile. The resulting typical profile is used in the future to compare forecast traffic values, due to this, anomalous activities are detected.

A review of existing approaches to solving the problem of searching for attacks in networks showed that most of the approaches considered to one degree or another use anomaly search, a signature approach or specifications, to identify attacks. The main significant drawback of the approaches considered is the high computational complexity, which is unacceptable for real-time systems. An equally important drawback of research is an experimental study that affects no more than 5-6 varieties of attacks. Nevertheless, it is worth noting that most of the solutions presented show a fairly high accuracy in determining attacks, on average from 50 to 90%.

3. Materials and Methods

Intrusion detection can be considered as one of the following machine learning tasks.

The task of forecasting. This task is aimed at predicting future opportunities and trends. Having at its disposal a prior information about the nature of upcoming events and upcoming changes, it is possible to prepare the necessary solution in advance. One of the main forecasting methods is regression analysis, with which you can predict the possibility of an attack. Regression analysis allows predicting similar behavior from collected attack logs.

The task of classification. Similar tasks are aimed at dividing into classes of objects (network behavior) according to certain and predefined signs. Classification allows you to highlight everything
that does not fit into standard classes, i.e., to identify an intrusion. Thus, the classification helps the security administrator determine the direction of protection and analysis. To solve the classification problem, teaching methods with a teacher are used.

The task of finding associative rules. This task is aimed at identifying hidden relationships between data. A relation from a time stream is called a sequence rule. If patterns are found in such sequences, it is possible with some probability to predict the occurrence of events in the future, which makes it possible to make more correct decisions. This analysis method can determine abnormal network behavior by analyzing user behavior or the operation of network devices.

4. Informal statement of the problem
An urgent task is to develop models and mechanisms for ensuring network security. To do this, it is necessary to develop a model for identifying attacks on the network using behavioral analysis based on monitoring the characteristics (attributes) of network behavior. Under the behavior of the network we understand the totality of the characteristics (attributes) of the network at a certain point in time.

Identification involves matching current network behavior with known behavior (normal behavior, behavior under attack). Thus, the task of identifying attacks reduces to the task of classifying network behavior.

In the framework of this study, the classification problem can be formulated as follows.

Let there be given many behavior classes of the network \( C \) (normal behavior, behavior during an attack); many signs of the behavior of the network \( K \); many investigated network behaviors \( X_c \).

Of the many signs of the behavior of the network \( K \) we single out a subset \( K_i \subseteq K \) – the set of the most informative signs. That is, each behavior is represented by a feature vector \( K_i \) of dimension \( k \).

Assuming that, \( X_i \) – the set of studied objects represented by a feature vector \( K_i \), \( |K_i| = k \).

Let, with the help of some rule, \( R \) be constructed a training set - a set of labeled vector-signs of network behavior, i.e. \( \Psi_{X_i} = (x_i, \bar{c}) \) \( \cup \bigcup_{c \in C} \cup \bigcup_{s \in C_c \cap X_c^t} (x, \lambda(C)), \) where \( X_c^t \subseteq X_c \) is plurality of training vectors, and \( x_i, \bar{c} \) is attitude of belonging.

It follows that to solve the problem of identifying attacks on a network, it is necessary to solve the following optimization problem based on a training sample \( \Psi_{X_i} \).

For this, we additionally introduce the set \( G = \{K_{i_1}, ..., K_{i_k}\} \), where \( K_{i_j} \subseteq K_i \), \( |K_{i_j}| = j \), as well as an auxiliary function \( \psi : G \rightarrow \bigcup_{c \in C} \bigcup_{s \in C_c \cap X_c^t} (X_{C_{s_{i_j}}}, ..., X_{C_{s_{i_j}}}) \), where \( X_{C_{s_{i_j}}} \) is investigated network devices represented by a vector of length \( i \).

It is required at a given level of identification efficiency \( \sigma \) to minimize the number of used signs \( k \in K_i \), i.e.

\[
\Lambda(a, X_c, \sigma) = \min_{i \leq k, i \in K_i} \left| K_{i_j} \right|, \quad \Omega(a, \psi(K_{i_j})) \geq \sigma \rightarrow \min_{i \leq k, i \in K_i},
\]

where \( \Lambda(a, X_c, \sigma) \) is function to evaluate the minimum set of behavioral signs; \( \sigma \) is some given positive number.

To implement the identification and classification of attacks on the network in the framework of this study, an appropriate model is constructed.

5. Formalization and analysis of the attack identification model
The developed model of node attack identification is based on the definition of a typical attack profile. An attack profile usually includes the formation of a whole chain of nodes connected by the same attack. Such a process requires significant computing resources. However, resource monitoring systems already
have the necessary profiles of legitimate traffic behavior, as well as historical records of recorded attacks, on the basis of which it is possible to build profiles of illegitimate traffic behavior in the network. We will identify attacks based on tracking changes in the main characteristics of network nodes. Suppose that at time $t$ in the network of telecommunication service providers on some node 2, an attack was carried out previously carried out on node 1 (Figure 1).

![Figure 1. Telecommunications Service Provider Experimental Network Segment Diagram](image)

In addition, it is known that on each of the attacked nodes $\{0, 1, 2\}$ a chain of changes in the states of the nodes was found, which characterizes the profile of the recorded attack:

$$S_i \xrightarrow{t^1} S_j \xrightarrow{t^2} S_k \xrightarrow{t^3} S_m \xrightarrow{t^4} S_r \xrightarrow{t^5} \ldots \xrightarrow{t^q} S_p,$$

It is clear that each transition from one state to another is accompanied by a change in characteristics:

1. The amount of traffic $K_1$ passing through the node at time $t$;
2. Indicator CPU $K_2$%;
3. $K_3$ node response time, ms;
4. The number of lost packets $K_4$%;
5. The state of RAM $K_5$%;
6. Throughput $K_6$%;
7. Statistics of violations of inputs $K_7$%.

To describe a typical attack profile based on a change in the characteristics of a node, we pass from discrete values to a description format of the form:

-1 if $K_i$ increased during the transition from one state to another;
-1 if $K_i$ decreased during the transition from one state to another;
0 if $K_i$ not changed.

Thus, the profile of the attack corresponding to the parameters will look like:

$$K_0 \begin{bmatrix} k_{01} & k_{02} & \ldots & k_{0p} \\ k_{11} & k_{12} & \ldots & k_{1p} \\ k_{21} & k_{22} & \ldots & k_{2p} \\ k_{31} & k_{32} & \ldots & k_{3p} \\ k_{41} & k_{42} & \ldots & k_{4p} \\ k_{51} & k_{52} & \ldots & k_{5p} \\ k_{61} & k_{62} & \ldots & k_{6p} \end{bmatrix},$$

where $k_{ij} \in \{-1; 0; +1\}$. 

5
This approach allows you to quickly identify low-intensity and high-intensity attacks of the same type to determine regardless of the values of the parameters themselves, focusing on the nature of the processes.

Then the task of detecting attacks on related nodes \{3, 4, 5, 6, etc.\} can be determined much earlier, before it damages network performance. The solution scheme can be represented as the following sequence of steps.

To identify the nodes associated with the identified, in which the values of the characteristics as a whole differ from the norm. This stage will reduce the dimension of the problem.

Among the nodes that have a deviation from the norm in the characteristics, it is necessary to analyze the coincidence of the chain of states (partial). Due to the fact that we consider the attack on these nodes to be incomplete. For example, nodes 4 and 5 have at state time $t$ chains of states $S_i \rightarrow S_j$ and $S_i \rightarrow S_j \rightarrow S_k$, respectively.

To directly confirm the conduct of the same type of attack, it is sufficient to classify the attack according to the profile of the change in characteristics. However, to improve the accuracy of determining attacks, in the framework of this study, an algorithm for collecting and analyzing logs generated on network equipment was developed.

6. Algorithm for collecting and analyzing logs for system logs of network devices

The algorithm for collecting and analyzing logs of network devices is an additional mechanism for detecting suspicious activities in the network. At the same time, to obtain complete and reliable information, it is necessary to carry out not only collection, but also filtering of aggregated data. This is necessary in order to generate incidents only for extremely dangerous events. At the same time, such a solution should work in real time with minimal delays. The solution to this problem is a combination of these factors and a large number of devices becomes difficult to implement. To reduce the computational complexity and increase the performance of the algorithmic solution, it is proposed to use the Cassandra distributed database, which allows collecting logs of network devices in a single repository. The algorithm is also based on a previously developed mathematical decision-making model based on the identification of related sets of events by generating sets of metadata at the preliminary stages of information receipt (in the signature analysis system, quick classifier, anomaly identifier).

Thus, the algorithm for collecting and analyzing logs of network devices can be represented as the following sequence of steps.

Step 1. Information is collected from security elements and network nodes in the syslog (log of system, services and applications) and SNMP (node status, performance counters) formats through a server that saves raw data for long-term storage in a relational distributed database Cassandra. The database stores information about system logs of applications and services operating in the network of a telecommunications service provider which do not write to syslog. In this case, a separate, private collection client is installed on the node, which publishes the data in the Apache Kafka sections. The collection server then reads the data from Kafka and sends it to Cassandra. NetFlow / sFlow network traffic data is also placed in Cassandra.

Step 2. Raw data is extracted from the Cassandra distributed database and events are generated from them using a set of structuring rules. Events have a hierarchical structure, each new node in the hierarchy adds additional fields. A tree of nodes is created for security elements of a particular type, which take into account all possible security events. The resulting structured set of events is saved back to Cassandra. To work in streaming mode, this step uses the Apache Spark framework.

Step 3. The structured data must be processed and the events in the network correlated with the state of the devices from which the data was received. Based on the data obtained, it is necessary to generate incidents with an intrusion detection system. For these purposes, it is proposed to use Spark Streaming in combination with the use of gradient boosting of decision trees in the CatBoost implementation. The advantage of this approach is the ability to identify new types of attacks that are a modification of existing ones, which existing signature methods cannot cope with.
It is worth noting that the originality of the developed algorithm is the possibility of its operation in two modes - in the mode of immediate processing as new events occur and in the execution mode on demand (for a retrospective analysis of events and investigation of incidents). For data analysis according to the presented algorithm using Python-based tools, a prototype module for data mining of system logs of network devices logs is implemented.

7. Experimental results
To assess the effectiveness of the sets when detecting suspicious network activity, real data is needed to describe the operation of the network of telecommunication service providers in which various attacks took place. Therefore, in the study, the CICIDS2017 data set developed at the University of New Brunswick was used to evaluate intrusion detection in the framework of this study. To study the effectiveness of these systems in this work, the following types of attacks were studied: DOS Slowloris, DOS Hulk, DOS Slowhttptest, DOS Goldeneye, Heartbleed, Web Brute Force, Web XSS, Web Sql inject, Infiltration - Dropbox download, Meta exploit Win Vista. On the proposed data set, a study of the accuracy of the proposed solution was conducted. The results of analysis are presented in Table 1.

| Attack Type                      | fp   | tp   | fpf   | tpf         |
|----------------------------------|------|------|-------|-------------|
| DOS Slowloris                    | 988  | 4012 | 0.01643 | 0.100635    |
| DOS Hulk                         | 5124 | 5430 | 0.101641 | 0.236837    |
| DOS Slowhttptest                 | 4929 | 5625 | 0.18361 | 0.377932    |
| DOS Goldeneye                    | 6100 | 4454 | 0.285051 | 0.489653    |
| Heartbleed                       | 4599 | 5955 | 0.361532 | 0.639025    |
| Web Brute Force                  | 7601 | 2953 | 0.487935 | 0.713096    |
| Web XSS                          | 7748 | 2806 | 0.616783 | 0.78348     |
| Web Sql inject                   | 7776 | 2778 | 0.746096 | 0.853162    |
| Infiltration – Dropbox download  | 7327 | 3227 | 0.867943 | 0.934106    |
| Meta exploit Win Vista           | 7941 | 2627 | 1      | 1           |

The result of applying the proposed solution to the developed solution for identifying attacks in the network of a telecommunications service provider, taking into account the monitoring of system logs to recognize suspicious activities on network devices, is the correct identification of critical situations in 90% of the cases considered, which confirms its sufficient effectiveness.

8. Conclusion
In the framework of this study, the problem of ensuring network security has been solved; it is reduced to the problem of identifying attacks carried out on a telecommunications service provider network using an analysis based on monitoring the characteristics of network behavior and the information from system logs of network devices.

During the study, existing approaches to the search for anomalies were examined and a hybrid solution was proposed that allows identifying individual attacks in the general flow of events, as well as comparing attacks with events occurring in the system logs of network devices.
Acknowledgment
The study was carried out with the financial support of the RFBR in the framework of scientific projects No. 18-07-01446, No. 20-07-01065, as well as a grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation (NSh-2502.2020.9) and a grant from the President of the Russian Federation for state support of young Russian scientists - candidates of sciences (MK-860.2019.9).

References
[1] Zabrodina L, Parfenov D, Bolodurina I, Torchin V and Zhigalov A 2019 Development of a model of cyberattacks identification based on the analysis of device states in the network of a telecommunications service provider Proceedings of the 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (IEEE) pp 0675-0680.
[2] Bolodurina I, Parfenov D and Torchin V 2018 Development and investigation of adaptive firewall algorithm to protect the software-defined infrastructure of multi-cloud platforms Proceedings of the 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT) (IEEE) pp 1–5.
[3] Shon T and Moon J 2007 A hybrid machine learning approach to network anomaly detection Information Sciences 177(18) (Elsevier) pp 3799-3821.
[4] Pajouh H H, Dastghaiybafard G and Hashemi S 2017 Two-tier network anomaly detection model: a machine learning approach J Intell Inf Syst 48 (Cham: Springer) pp 61-74.
[5] Naseer S, Saleem Y, Khalid S, Bashir M K, Han J, Iqbal M M and Han K 2018 Enhanced network anomaly detection based on deep neural networks IEEE Access 6 (IEEE) pp 48231-48246.
[6] Kwon D, Kim H, Kim J, Suh S C, Kim I And Kim K J 2019 A survey of deep learning-based network anomaly detection Cluster Comput 22 (Cham: Springer) pp 949-961.
[7] Lin W C, Ke S W and Tsai C F 2015 CANN: An intrusion detection system based on combining cluster centers and nearest neighbours Knowledge-based systems 78 (Elsevier) pp 13-21.
[8] Kuang F, Xu W and Zhang S 2014 A novel hybrid KPCA and SVM with GA model for intrusion detection Applied Soft Computing 18 (Elsevier) pp 178-184.
[9] Ashfaq R A R, Wang X Z, Huang J Z, Abbas H and He Y L 2017 Fuzziness based semi-supervised learning approach for intrusion detection system Information Sciences 378 (Elsevier) pp 484-497.
[10] Bhuyan M H, Bhattacharyya D K and Kalita J K 2016 A multi-step outlier-based anomaly detection approach to network-wide traffic Information Sciences 348 (Elsevier) pp 243-271.
[11] Fernandes Jr G, Carvalho L F, Rodrigues J J and Proença Jr M L 2016 Network anomaly detection using IP flows with principal component analysis and ant colony optimization Journal of Network and Computer Applications 64 (Elsevier) pp 1-11.