www.ioirp.com
International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 3, Issue – 3. May 2018

# A Novel Graph-based Descriptor for the Detection of Billing-related Anomalies in Cellular Mobile Networks

P.Tharanya[1]
Department of CSE,
Greentech college of Engineering for women,
Attur, Salem.
Tharan2397@gmail.com

G.Gowri priya[2]
Department of CSE,
Greentech college of Engineering for women,
Attur, Salem
Bharathig278@gmail.com

V.Nandhini[3]
Department of CSE,
Greentech college of Engineering for women,
Attur, Salem
Nandhinisv1610@gmail.com

**Abstract—**Mobile devices are evolving and becoming increasingly popular over the last few years. This growth, however, has exposed mobile devices to a large number of security threats. Malware installed in smartphones can be used for a variety of malicious purposes, including stealing personal data, sending spam SMSs, and launching Denial of Service (DoS) attacks against core network components. Authentication and access- control-based techniques, employed by network operators fail to provide integral protection against malware threats. In order to solve this issue, the activity of each mobile device in the network must be taken into account, and combined with the activities of all the other devices. The communication activity in the mobile network has a source, a destination, and possibly communication weights (e.g. the number of calls between two mobile devices). This relational nature of the communication activity is naturally represented with graphs. This indicates that graphs can be utilized in order to provide better representations of the entire network activity, and lead to better detection results when compared to methods that consider the activity of each mobile device individually. Towards this end, this paper proposes a novel graph-based descriptor for the detection of anomalies in mobile networks, using billing-related information. The graph-based descriptor represents the total activity in the network. Smaller graphs are afterwards extracted from the graph-based descriptor, each one representing the activity of one mobile device (e.g. Calls or SMSs), while multiple features are calculated for each such graph. These features are subsequently used for the supervised classification on network events, and the identification of anomalous mobile devices. Experimental results and comparison of the proposed anomaly detection method to the existing work, show that the graph-based descriptor has superior performance in a variety of scenarios.

## I INTRODUCTION

The wide adoption of smartphones encompassing personal data such as, contacts' list, financial data, and credentials for online banking, has augmented the interest of cyber-criminals, not only due to possible financial gains, but also because of the possibility to utilize these devices for launching attacks against the mobile core network. . Attacks against the core network can have a negative impact on the network performance. As demonstrated by Traynor et al. a botnet composed of as few as 11,750 compromised mobile phones, is able to degrade service to area-code sized regions by 93%. As a countermeasure to Denial of Service (DoS) attacks against core network components, mobile network operators employ authentication-based techniques to prevent illegitimate users from attaching to the network. These techniques, however, do not offer sufficient protection against attacks, since malicious individuals can still infiltrate the network by utilizing compromised mobile devices of legitimate subscribers by using malware. DoS attacks against the core network can be detected from either the signaling (control) or the billing (data) planes, depending on the type of attack. The signaling plane in the mobile network is comprised of all the signals that control or are needed for the network services. Previous research has demonstrated the effect of several attack scenarios against the core network using the signaling plane.

## II ANOMALY DETECTION METHODOLOGY

The technique uses graph-based descriptors (Section IV-A) to capture the network billing activity for a specific time period, where nodes in the graph represent users and/or servers, and edges represent communication events. Section IV-B presents the method followed to create multiple graphs in each vertex neighborhood of predefined size. Section IV-C presents the features that are extracted from the neighborhood graphs for different neighborhood sizes k, and used in order to train a supervised classification algorithm, namely, the Random Forest (RF) classifier [27], to recognize anomalous graphs.

A. **Graph-based Descriptor:** A graph G(V,E) is comprised of a set of vertices V = {v1,v2,...,v|V|} and a set of edges E = {e1,e2,...,e|E|}, where E ⊂ V ×V . A graph descriptor is a weighted directed graph in which W : E →R is a function that takes as input a specific edge and returns its corresponding edge weight. The vertices of the graph descriptor can thus, represent network entities (e.g. users, servers, etc.), the edges correspond to communication events between them, and the edge weights capture the attributes of these communications.

TABLE I EXAMPLE OF RECORD DATA REPRESENTING THE ORIGIN AND DESTINATION OF EACH COMMUNICATION EVENT.

| Record ID | Origin | Destination |
|---|---|---|
| 1 | ID-1 | ID-2 |
| 2 | ID-1 | ID-2 |
| 3 | ID-2 | ID-1 |
| 4 | ID-3 | ID-2 |

In an example of record data representing the origin and destination of each communication event for three entities: ID-1, ID-2, and ID-3.

### B. Graph neighborhoods

Let $N_k\ (v_i)$denote the set of k-neighbors of vertex $v_i \in$ V . This set is comprised of all the nodes that have graph geodesic distance smaller than or equal to k, where the geodesic distance GD($v_{i,}v_j$) between two vertices vi and vj is the length of the shortest path connecting them. Hence, the k- neighbors of vertex $v_i \in$ V are defined as

$$N_k\ (v_i)\ = \{v_j \in \text{V}|\forall \text{j such that GD } (v_i, v_j) \leq \text{k}\} \quad (1)$$

Graph neighborhood extraction then consists in creating a new graph for each vertex vi, denoted as $G_i(V_i, E_i)$, where $V_i$ = N($v_i$) ∪ $v_i$, $E_i$= $\{e_j = \{v_k,v_h\}||\forall e_j \in$ E, and $v_k,v_h \in V_i\}$, and E is the set of edges of the initial graph descriptor G.

Since the graph is directed, we denote by $e_j^s$ the source vertex of edge $e_j \in E_i$, and by $e_j^d$its destination vertex. For each k-neighbors of vertex vi ∈ V , the set of outward directed edges represent all the edges that have their source in the set $N_k(v_i)$ and their destination outside of this set. More precisely, the set of outgoing edges for $N_k(v_i)$ is given by:

$$E_i^{out} = \{e_j \in E_i\ |\forall e_j^s \in N_k(v_i) \text{ and } e_j^d \neq N_k(v_i)\}$$

(2)

The set of ingoing edges $E_i^{in}$ for the set $N_k(v_i)$ is defined in a similar manner.

### C. Graph-based Feature Extraction

The features proposed in this paper focus specifically on anomaly detection, and can be applied on arbitrary neighborhood sizes.

These features are extracted for multiple neighborhood sizes k ∈ {0,...,N} and are all fed into the supervised classifier. N represents the maximum neighborhood size to be taken into account, and is equal to the graph diameter3. The definition range for the values of neighborhood sizes k for each feature.

## III AN OVERVIEW OF THE GRAPH-BASED FEATURES UTILIZED FOR ANOMALY DETECTION.

| S.No | Feature | Short Description | Definition Range |
|---|---|---|---|
| 1 | Volume | The number of edges of the graph | k ≥ 1 |
| 2 | Edge Entropy | The entropy of the weights of the edges in the graph | k ≥ 1 |
| 3 | Graph Entropy | Captures the structural characteristics of the graph | k ≥ 1 |
| 4 | Edge Weight Ratio | The ratio of the sum of weight of outward to inward edges | k ≥ 0 |
| 5 | Average Outward/Inward Edge Weight | The average weights of the Outward/ Inward Edges | k ≥ 0 |

1) **Volume:** The volume feature is defined as follows:

$$f_{vol}^{G_i} = \sum_{e_j \in E_i} g(W(e_j)) \quad (3)$$

where: $$g(x) = \begin{cases} 1, & for\ |x| \neq 0 \\ 0, & for |x| = 0 \end{cases}$$

and $G_i$ denotes the neighborhood graph of vertex $V_i$, $E_i$ its set of edges, and W($e_j$) the weight of edge $e_j \in E_j$.

2) **Edge Entropy**: The edge entropy feature is used to capture behavioral changes within a specific static graph neighborhood. The edge entropy of a graph $G_i$ is defined as:

$$f_{ee}^{G_i} = -\sum_{j=1}^{y^i} \frac{y_j^i}{y_{total}^i} \log \left(\frac{y_j^i}{y_{total}^i}\right)$$

Where $Y^i$ is the number of different edge weight instances of graph $G_i$, $y_j^i$ is the number of occurrences of the j-th weight, and $y_{total=}^i \sum_{j=1}^{Y^i} y_j^i$ is the total number of weight occurrences.

3) **Graph Entropy**: More formally, Korner's entropy is defined

$$f_{ge}^{G_i} = \min_{X,Y} I(X \wedge Y)$$

where I(X∧Y ) is the mutual information of the variables X and Y.

4) **Edge Weight Ratio:** The edge weight ratio feature $f_{wr}$ captures the total difference in the values of the weights between the edges of a graph $G_i$ directed outward from vi and inward to vi.

$$f_{wr}^{Gi} = \frac{\sum_{e_{j \in E_i^{out}}} W(e_j)}{\sum_{e_{j \in E_i^{in}}} W(e_j)}$$

where $E_i^{out}$ and $E_i^{in}$ are the set of outward and inward directed edges for the k-neighbors of graph $G_i$.

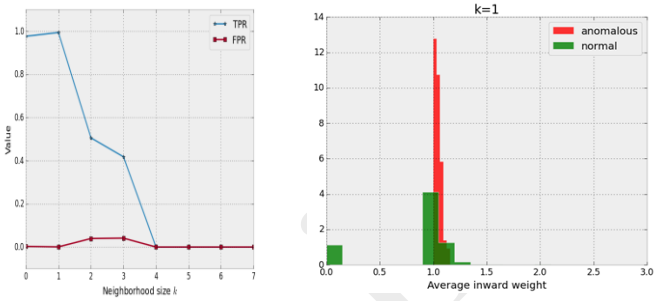
5) **Average Outward/Inward Edge Weight**: The average outward edge weight $f_{a\,vout}$ represents the ratio of the volume of traffic generated by a node to the number of its destinations:

$$f_{a\,vout}^{G_i} = \frac{\Sigma_{e_{j \epsilon E_i^{out}}} W(e_j)}{|E_i^{out}|}$$

## IV EXPERIMENTAL RESULTS

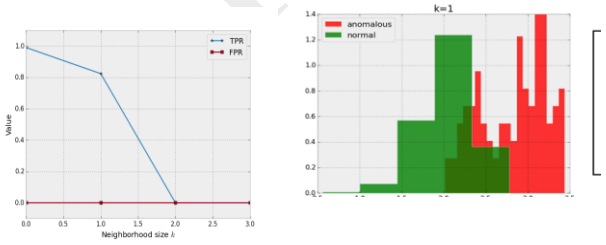
SMS flood Results: Since the important factor in this dataset is SMS activity, the graph descriptor is created to represent this behavior. The weight of an edge represents the number of messages exchanged between the corresponding devices. This results in a dynamic graph (similar to [20]), comprised of sequence of 24∗7 = 168 static graphs.

Fig: Evaluation of the graph descriptor approach on the SMS flood dataset, by taking into account only the features extracted from different values of neighborhood sizes k, and not all together (as done in the default method).




Spam SMS: There are in total 10,000 mobile devices which are divided into three distinct groups according to behavior: (1) 4,000 users exhibit low levels of SMS activity, (ii) 3,000 users exhibit medium levels of SMS activity, and(iii) 4,000 users exhibit high levels of SMS activity. In the simulations, 102 mobile devices, uniformly distributed across the three distinct groups, are assumed to be infected with spam malware

.




Evaluation of the graph descriptor approach on the Spam SMS dataset, by taking into account only the features extracted from different values of neighborhood sizes k, and not all together (as done in the default method).

## V CONCLUSION

This paper presented an anomaly detection approach using billing-related Information. A novel graph-based descriptor is introduced, which represents the activity of each mobile subscriber. Multiple features extracted from the graphs are able to characterize different aspects of their activity, and discriminate between normal and abnormal behaviors. The proposed features are able to capture changes in the communication destination (e.g. Graph entropy) the communication volume (e.g. volume, edge entropy) and the communication directions (e.g. edge weight ratio, average outward/inward edge weight.

## REFERENCES

[1] I. Murynets and R. P. Jover, "Anomaly detection in cellular Machine- to-Machine communications," in Communications (ICC), 2013 IEEE International Conference on, pp. 2138–2143, IEEE, 2013.

[2] Q. Xu, E. W. Xiang, Q. Yang, J. Du, and J. Zhong, "Sms spam detection using noncontent features," IEEE Intelligent Systems, vol. 27, no. 6, pp. 44–51, 2012.

[3] P. Laskov, P. D¨ussel, C. Sch¨afer, and K. Rieck, "Learning intru- sion detection: supervised or unsupervised?," in Image Analysis and Processing–ICIAP 2005, pp. 50–57, Springer, 2005.

[4] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," Data Mining and Knowledge Discovery, pp. 1– 63, 2014.

[5] E. K. Kim, "A detection mechanism for SMS flooding attacks in cellular networks," in Security and Privacy in Communication Networks, pp. 76–93, Springer, 2013.

[6] O. H. Abdelrahman and E. Gelenbe, "Signalling storms in 3G mobile networks," in Communications (ICC), 2014 IEEE International Confer- ence on, pp. 1017–1022, 2014.

[7] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5– 32, 2001.

[8] G. Simonyi, "Perfect graphs and graph entropy. An updated survey," Perfect graphs, pp. 293–328, 2001.

[9] N. Gobbo, "A Denial of Service Attack to GSM Networks via Attach Procedure," in Security Engineering and Intelligence Informatics, pp. 361–376, Springer, 2013.

[10] G. Yan and S. Eidenbenz, "Sim-Watchdog: Leveraging Temporal Similarity for Anomaly Detection in Dynamic Graphs," in Distributed Computing Systems (ICDCS), 2014 IEEE 34th International Conference on, pp. 154–165, IEEE, 2014.