Anomaly Detection In Cellular Network Data Using Big Data Analytics

Ilyas Alper Karatepe, Engin Zeydan

Abstract—**Anomaly detection is a key component in which perturbations from a normal behavior suggests a misconfigured/mismatched data in related systems. In this paper, we present a call detail record based anomaly detection method (CADM) that analyzes the users's calling activities and detects the abnormal behavior of user movements in a real cellular network. CADM is capable of detecting the location of the site that an anomaly has occurred. We evaluate the proposed CADM by performing experiments over the call-detail records of AVEA, a mobile service provider in Turkey.**

Index Terms—**anomaly detection, big data, cellular networks, network management, Hadoop**

I. Introduction

Mobile Service Providers (MSPs) are forced to process huge amounts of user generated call records (sms, voice, data, etc.) on a daily basis. Analyzing this big data can help in solving some of the most common problems in MSPs. In addition, mobile operator services are diverse and have been offered with a multiple number of functionalities, but the current service structure faces several challenges: first of all, the possibility of failure/error prone manual configurations and management to deliver new networking services (e.g. wide area network-intensive applications such as video, pictures, etc.) presents a constant challenge for mobile operators as well as service providers particularly in areas where correctness of the supplied information is critical (e.g. billing and charging, policy and regulation requirements, etc.). Every new roll-out and deployments in a mobile operator infrastructure introduces new modifications on every system and each related entity (both operational and planning departments) should adapt into that. These frequent changes in a cellular network environment sometimes lead to transient states caused by the network having to manage all changes from multiple components in IT departments, resulting in misconfiguration and instability. Second, the deployment life-cycle of new changes in an enterprise network can be long. During long lifecycle operations, an excessive time is wasted in order to settle for an undisrupted service delivery after addition of each new functionalities into their services. Moreover, the problem does not necessarily decrease with each additional collection of devices. This requires MSPs to either redesign their enterprise architecture for each changes that occur or to limit their services to a lower level.

Third, the forecasting and planning of additional services can be imperfect, hence the location, timing and severity of flaws are sometimes very difficult to predict. Several times MSPs are faced with the challenges of not being able to diagnose the solution to a problem that may occur. Answering simple questions such as ”What are the misconfigured attributes?” or ”What is the frequency of anomalous activities?” in a mobile operator network is particulary difficult today and remains a very challenging issue. Forth, major problem that many network/IT system administrators face is the problem of detecting a defect in user activity from a pool with many users with millions of transactions. In large datasets, there occurs small percentage of anomalies which may be common and difficult to observe. This anomaly may indicate a bad data, a random variation or of utmost interest for operation. In all cases, additional actions should be required.

One solution to these network/IT management problems is to apply knowledge based anomaly detection methods and set rule policies depending on network behavior. In this paper, we present one variant of knowledge based technique, a rule-based technique, for detecting network anomalies for users traveling from one city to another. The method is robust and flexible for detecting anomalies, which may also be adjusted based on the needs of the operations. The contributions of the paper can be summarized as follows:

- A novel approach for anomaly detection by analyzing call-detail records attributes in combination with recent big data analytical tools (Hadoop (HDFS, Map-Reduce), Hive, etc). This approach reduces the cost of data processing compared to traditional data warehouse approaches since big data analytics provides computational ease which is provided by cloud services. Moreover, keeping and processing the data in conventional enterprise systems may be expensive, depending on the licence agreements between the operator and database providers.
- The method is based on knowledge (or rule) based approach rather than signatures and patterns that need training data sets to detect any type of anomaly that falls out of normal activity.
- Reduced IT service lifecycle for error/misconfiguration detection: The proposed approach can easily discover anomalies especially that occur during roll-out or change periods in a reasonably short period of time compared to traditional long IT service lifecycle.

A. Related Work

Anomaly or outlier detection is the method of searching for data items that do not match in an expected behavior or a pattern in a given data set. They often provide critical and

I. A. Karatepe and E. Zeydan are with AveaLabs at AVEA İletişim Hizmetleri A.Ş. İstanbul, Turkey. (e-mail: alper.karatepe@avea.com.tr, engin.zeydan@avea.com.tr)

actionable information. There are many extensive surveys of anomaly detection techniques developed in several domains of machine learning and statistics (e.g. see [1] [2]) and reference therein). Network anomaly detection methods in the literature can be classified into five categories: (i) Statistical-based (ii) Classification-based (iii) Clustering and Outlier-based (iv) Soft computing-based (v) Knowledge-based.

Statistical methods based anomaly detection methods are used with applying statistical models for anomaly detection which can be categorized based on parametric and non-parametric techniques. This is the approach that has been followed by [3] [4]. Statistical approaches have the advantage of requiring no previous knowledge and reporting abrupt changes after a long monitoring time. However, they have some drawbacks such as long training period for statistical inferences and difficult task of adjusting or tuning different parameters.

There are several classification-based methods such as k-nearest neighbor, support vector machines and decision trees that have been applied to network anomaly detection problem [5] [6] [7]. Classification-based techniques are based on categorization of new observations into different classes depending on the relevant model that was constructed based on the training data set. Although classification-based techniques are popular due to their considerable flexibility and high detection rates, they require relevant training information and is highly dependent on assumptions of the classifiers. Another anomaly detection technique is to use both clustering and outlier detection techniques. Clustering techniques such as k-means, hierarchical clustering and single-link clustering are also frequently used in anomaly detection [8] [9]. Although clustering has some advantages such as reducing computational complexity for large data sets, stable performance compared to statistical methods, it also has some drawbacks such as time consuming dynamic updates and inappropriate proximity measures that may lead into lower detection rates. In contrast outlier based detection methods are useful for detecting bursty and isolated events but they also suffer from high parameter dependencies.

Soft computing-based anomaly detection methods and systems (such as genetic algorithms, artificial neural network approaches and fuzzy set approaches [10] [11] [12]) have the advantage of learning without any feedback from the environment and adaptability. On the other hand, scalability and over-fitting may be some of the disadvantages of these methods. Knowledge-based anomaly detection methods (such as rule-based, ontology-based and logic-based) are checking against the predefined rules or patterns in order to detect anomalies. Rule-based and expert system approaches are the most widely used knowledge-based methods. Based on the knowledge base, matching rules against the current state of the system ensures anomaly detection in the system. Some of the approaches are: processing rules on packet level signatures (e.g. Adaboost [13] and Snort [14]), an expert system that makes a decision that is reasonable to a common human sense [15]. Another technique called *prudence* ensures that when a new value that is outside the range of recorded list is monitored by the system, the expert system raises a warning [16].

Although all the network anomaly-detection techniques mentioned above have their own strengths and weaknesses, due to the nature of our problem we are using a rule-based approach in this paper. One of the advantages of using this approach is that the knowledge-based techniques are robust and flexible which may be adjusted based on the needs of the operations. Moreover, the detection rate is high compared to other methodologies such as statistical or unsupervised learning based methods (e.g. clustering based methods) where expanding the anomaly detection methods with may give different results than expected. The problem that we are investigating in this paper is related to traveling of a user from one location to another in a physically impossible time duration. There may be two reasons for this anomaly: misconfiguration of location information from the network side or the incorrect mappings of related location attributes in IT systems (especially in mediation section).

This paper is organized as follows: Section II provides system model for Call Detail Record, a.k.a. Call Data Record (CDR) generation, proposed architecture and concepts. Section III presents the proposed CDR based anomaly detection method (CADM). Section IV gives the experimental results of the proposed solution. Finally, Section V provides conclusions and future work.

II. System Model, Concepts and the Proposed Architecture

In this subsection, we investigate the system model, concepts as well as the CDR generation lifecycle architecture with the proposed anomaly detection method. The proposed architecture leverages big-data analytic techniques (Hadoop (HDFS – Hadoop Distributed File System and MapReduce) and Hive) for cost effective computation to detect network anomalies by using the proposed CDR based anomaly detection method (CADM). Among the available big data platforms, Hadoop which allows the distributed processing of big amount of data sets across clusters of computers, stands out as the most notable one as it is an open source solution [17]. Its main components include MapReduce, HDFS (Hadoop Distributed File System), HBase, ZooKeeper, etc.

Fig. 1 represents the CDR generation lifecycle architecture in a typical mobile operator in combination with the proposed anomaly detection method. Some of the CDR attributes and their definitions that are used in this paper are as follows: $MSISDN$ represents user's phone number, $CELL\ ID$ represents user's cell-ID information where the user has initiated or received a call, $LAC\ ID$ represents the location area code information, $CITY\ ID$ represents user's city-ID information, $CALL\ DATE$ represents user's call date as year, month and day, $CALL\ TIME$ represents user's call time and $CDR\ TYPE$ represents call detail record type (e.g. sms, voice and data). These attributes are call-activity related measures that are used to construct the anomaly detection method.

Basically, the CDR activities of an observed user can be described as follows: When the user initiates or receives a call i

(e.g. voice, sms or data) in a given location, the system records the t_i (the time when the i^{th} call happened, i.e. $CALL\ DATE$ and $CALL\ TIME$ pairs) and L_i (the location where the i^{th} call happened, i.e. $LAC\ ID$ and $CELL\ ID$ pairs.). In Fig. 1, the mediation first collects the CDR data via ftp from network nodes such as GGSN, SGSN, PCEF, etc. Then, it distributes the collected CDR data via ftp into relevant departments such as datawarehouse, billing and charging departments, etc for a wide range of purposes. During this process, mediation obtains $CITY\ ID$ of the location from $LAC\ ID$ for each i by an internal query and adds $CITY\ ID$ as an additional attribute to each i before distribution of CDR to relevant departments. We also define l_i as the location where the i^{th} call happened, i.e. $CITY\ ID$ and $CELL\ ID$ pairs. The above measures can be defined as attributes to anomaly detection method. Note also that during CDR generation process, managing the same configuration during the service lifecycle can be complex and error-prone due to involvement of multiple heterogeneous components and different mappings with different systems.

In Fig. 1, CADM collects the processed data from one of the relevant department (such as datawarehouse in this example) and runs the proposed method on this data. Then, the detected anomaly output of CADM is feedbacked into mediation department for taking necessary actions.

The proposed architecture paves the way for a flexible and robust way of anomaly detection for mobile operators by better coordinating their configurations throughout the service lifecycle to meet their policy level needs. Moreover, the anomaly detections are implemented by big data analytics and rule based approaches in CADM which will be detailed in next section.

III. CDR based Anomaly Detection Method (CADM)

In this section, we present our anomaly detection scheme based on user's call-detail record (CDR) activities that is collected in a big data platform. The main idea of the proposed algorithm is to detect suspicious $CITY\ ID$ and $CELL\ ID$ pairs which are anomalies by monitoring the user's call detail activities. The categorization is either normal or anomalous.

A user j's city-arrival time ($a^j_{t_k}$) and city-departure time ($d^j_{t_k}$) for a given city k can be identified from the collected data. The advantage of using city-arrival and city-departure times is that they can easily be obtained with the help of existing CDR attributes t_i and l_i. This approach can be used to characterize the calling activities of each user. Based on the above considerations, we define $R_j = \{P_1, P_2, P_3, ..., P_N\}$ as the sequence of traveled cities by a user j. Hence, any route of a user can be represented as a sequence of N elements that depends on the number of traveled distinct cities. In R_j, we determine a quintuple definition to represent the travel activities of a mobile user j for a specific city k: P_k = (ID_j, $l_k(a^j_{t_k})$, $l_k(d^j_{t_k})$, $a^j_{t_k}$, $d^j_{t_k}$) where ID_j is the $MSISDN$ of user j, $l_k(a^j_{t_k})$ and $l_k(d^j_{t_k})$ define the ($CITY\ ID$, $CELL\ ID$) pairs for arrival and departure location time information respectively. For instance, for a four hops route, i.e. $N = 4$, the route R_j can be represented as $R_j = \{P_1, P_2, P_3, P_4\}$. Definition of these attributes simplifies the tracking of users' inter-city traveling activities.

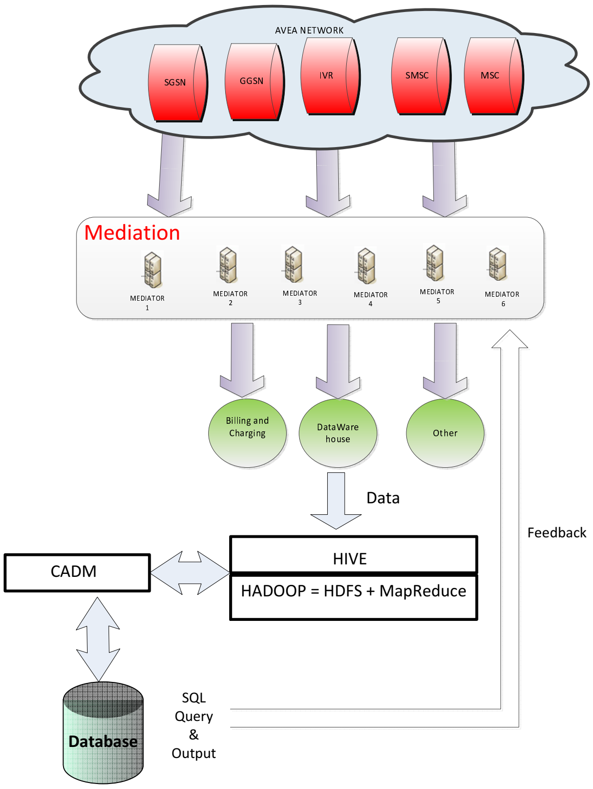

Fig. 1. CDR generation in a cellular network architecture and the proposed anomaly detection method

- Let TD represents the inter-city travel duration that a user has taken between P_k and P_{k+1} in his traveling route R_j. The k^{th} TD is defined as $TD_k = (a^j_{t_{k+1}} - d^j_{t_k})$.
- Let CD_k represents the average distance between cities of l_k and l_{k+1}.
- Let TV represents the inter-city travel velocity that a user has taken between P_k and P_{k+1} in his traveling route R_j. TV is defined as $TV_k = CD_k/TD_k$
- Let $MPTV_k$ represents maximum possible travel velocity between two consecutive cities l_k and l_{k+1} in a given route R_j

Algorithm 1 is the pseudo code of the proposed CDR based Anomaly Detection Method (CADM), originated from a rule-based approach which is used to detect anomalous $CITY\ ID$, $CELL\ ID$ pairs through the TV measurements. The *first step* in CADM is to extract the relevant attributes that can be used. Note that relevant attributes should be selected that can reflect the exact activities of a user. In CADM, attributes such as $MSISDN$, $CITY\ ID$, $LAC\ ID$, $CELL\ ID$, $CALL\ DATE$, $CALL\ TIME$, $CDR\ TYPE$ are extracted to form the $CDR\ ANALYSIS\ TABLE$ that reflects user's CDR activities. We eliminated the non-traveling users (i.e. users where $R_j = P_1$) during the analysis duration and obtained a total of M traveling users. This helps in decreasing the analysis space, hence the computational overhead of the system. Note also that during this step, we exploit distributed processing framework provided by map-reduce, which has advantage on computational overhead compared to non-Hadoop solutions.

Then, we formulate the problem as a knowledge based anomaly detection problem. In $step$ 2 of the algorithm 1, a rule-based technique is proposed to flag anomalous activities of users traveling between two different cities by calculating the travel velocities and comparing this velocity with $MPTV_k$. We use the following example to show how rule based approach is applied in the proposed CADM. Anomalies resulting from misconfigurations may cause deviations in user's tracks. By the usage of MPTV, one can ascertain how "reasonable" the calculated TV is for a given route R_j. Suppose $v_1 = 500km/h$ and $v_2 = 1500km/h$ are TVs between two given locations ($P_1- > P_2$ and $P_2- > P_3$) in $R_j = \{P_1, P_2, P_3\}$ respectively and $1000km/h$ is selected as MPTV for both of travels. We could then decide whether there exists any anomaly by just comparing these two velocities with MPTV which implies the existence of an anomaly for v_2 for the given $\{P_2- > P_3\}$ travel. Hence for each route of j^{th} user, if a TV exceeds a reasonable travel velocity, MPTV, we can use this "jump" information to extract the anomalous ($P_k- > P_{k+1}$) whereby the $CITYID$'s labeling of one of them is misconfigured in the system. Then, this table is stored in a relational database such as MySQL. Note that at this step, one of the stored (P_m)s where $m \in k, k+1$ is anomalous for each (P_k,P_{k+1}) pair. Based on this, another SQL query is run at the last step of Algorithm 1 in order to obtain the most probable anomalous $\{CITY\ ID, CELL\ ID\}$ pairs. Finally, CADM results can be sent into the relevant departments for taking necessary actions.

IV. Experimental validation

The CDR records were gathered for an $ANALYSIS\_DURATION = 10$ days during a holiday season in October 2013 where the amount of travels in Turkey are at its peak levels during the year. Moreover, during this period, there was routine new roll-outs and deployments in IT/Network systems.

The cumulative data is taken from the datawarehouse in the form of CDRs with total of 3800 million of rows, each row representing a certain call transaction (voice and sms) of a certain user. In our analysis, we have ignored $CDR\ TYPE = GPRS$ related CDRs and foreign MSISDNs. We selected $MPTV = 1000km/h$ for all travels which is faster than any transportation vehicle nowadays. In our analysis, we excluded all anomalous jumps between neighboring cities due to usage of average distance definition CD.

Our Hadoop cluster used in this experiment consists of 5 nodes with four slaves and one master. In our analysis, $162,506$ of $19,717,327$ city changes which corresponds to % 0.824 are detected as anomalous travel. Table I shows the top anomalous $\{CITY\ ID\ ,CELL\ ID\}$ pairs with their corresponding percentages among all anomaly occurrences in the collected data over the course of the observed day interval. According to results in Table I, necessary configuration checks and required fixes are done by network and mediation system operator.

Algorithm 1 CDR based Anomaly Detection Method (CADM)

```
Inputs:
$ANALYSIS\_DURATION$ : The duration of the observed time interval for analysis.
$CDR(Call\ Detail\ Record)$ : All calling activities of all users stored in operator's database for a certain time duration.
Outputs: List of anomalous $CELL\ ID$ and $CITY\ ID$ pair records
Method:
  1) Obtain the relevant attributes from call detail records stored in HDFS. The attributes are: $MSISDN$, $CELL\ ID$, $CITY\ ID$, $CALL\ DATE$, $CALL\ TIME$, $CDR\ TYPE$. Include users only who have done call record activities in at least two distinct $CITY\_IDs$ during $ANALYSIS\_DURATION$. Group the records by $MSISDN$ and sort them by $CALL\ DATE$, $CALL\ TIME$, then store the result in $CDR\_ANALYSIS\_TABLE$.
  2) for all j= 1,....,M do
       Obtain the route $R_j$ for each user $j$ in $CDR\_ANALYSIS\_TABLE$ where $R_j = \{P_1, P_2, P_3, ..., P_N\}$ and $P_k = (ID_j, l_k(a^j_{t_k}), l_k(d^j_{t_k}), a^j_{t_k}, d^j_{t_k})$ by traversing over each row record of $CDR\_ANALYSIS\_TABLE$. Note that the condition for adding a new $P_k$ into $R_j$ is the change of $CITY\ ID$ between two sequential records.
       for all k= 2,....,N do
         If $TV_{k-1} > MPTV_{k-1}$, then flag the $l_{k-1}$ and $l_k$ pair of $P_{k-1}$ and $P_k$ as anomaly and store the output in $CDR\_ANOMALY\_TABLE$ in a relational database such as MySQL.
       end for
  3) end for
  4) Execute a query on $CDR\ ANOMALY\ TABLE$ that finds and outputs the number of anomalous $CITY\ ID$ and $CELL\ ID$ pairs.
```

CITY ID	CELL ID	percentage
101	773	4.5721
126	1932	3.7216
115	19074	3.1195
104	59287	3.0691
106	105	2.9841
101	13521	2.5643
116	12381	2.4350
115	28116	2.3799
143	42632	2.0567
116	12386	2.0159
...	...	...

TABLE I
Top anomaly $CITY\ ID$ and $CELL\ ID$ pairs with their corresponding percentages among all anomaly occurrences

V. Conclusions

MSPs are operating over large IT and network infrastructure, including multiple access network and IT domains (carrying user as well as internal traffic). In this paper, we have proposed a rule based anomaly detection technique that can help MSPs improve their system consistency and reduce the time to detect location based anomalies. A rule based approach using big data analytics technique is promising for both detecting those anomaly problems with high accuracy and presenting flexible and robust ways to solve these problems according to operators' needs and domains. The advantages of the proposed anomaly detection method, CADM, is fast lifecycle in order to reach target results, better ways to extract relevant rules without needing a training phase and the ability to use modern cost-effective big data analytics technique. Note that for future work in order to obtain more accurate results, location update information can be included in the analysis.

VI. Acknowledgement

The present work was carried out within the framework of Celtic-Plus SHARING project. We also thank AVEA's network interface solutions department for providing the CDR data for analysis and recommendations.

References

[1] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *accepted to IEEE Communications surveys and Tutorials*, 2014.

[2] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[3] C. Manikopoulus and S. Papavassiliou, "Network intrusion and fault detection: A statistical anomaly approach," *IEEE Communications Magazine*, vol. 40, pp. 76–82, October 2002.

[4] Z. Zhang, J. Li, C. N. Manikopoulus, J. Jorgenson, and J. Ucles, "HIDE: a hierarchical network intrusion detection system using statistical prepocessing and neural network classification," in *Proc. IEEE Man. Systems and Cybernetics Information Assurance Workshop*, 2001.

[5] T. Abbes, A. Bouhoula, and M. Rusinowitch, "Efficient decision tree for protocol analysis in intrusion detection,," *Internation Journal Security and Networks*, vol. 5, pp. 220–235, December 2010.

[6] S. R. Gaddam, V. V. Phoha, and K. S. Balagni, "K-means + ID3: A novel method for supervised anomaly detection by cascading k-means clustering and ID3 decision tree learning methods,," *IEEE Trans. Knowl. Data Eng.*, vol. 19, pp. 345–354, March 2007.

[7] C. Wagner, J. Francois, R. State, and T. Engel, "Machine learning approach for IP-flow record anomaly detection," in *Proc. 10th International IFIP TC 6 Conference on Networking - Volume Part I,*, pp. 28–39, 2011.

[8] L. Ertoz, E. Eilertson, A. Lazarevic, P. Tan, V. Kumar, and J. Srivastava, "Data mining - next generation challenges and future directions," *ch. MINDS- Minnesota Intrusion Detection System*, 2004.

[9] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "An effective unsupervised network anomaly detection method," in *Proc. Internation Conference on Advances in Computing, Communications and Informatics,*, pp. 533–539, 2012.

[10] M. S. A. Khan, "Rule based network intrusion detection using genetic algorithm," *International Journal Computer Applications*, vol. 18, pp. 26–29, March 2011.

[11] G. Liu, Z. Yi, and S. Yang, "A hierarchical intrusion detection model based on the PCA neural networks," *Neurocomputing*, vol. 70, no. 7-9, pp. 1561–1568, 2007.

[12] S. Mabu, C. Chen, N. Lu, K. Shimada, and K. Hirasawa, "An intrusion detection model based on fuzzy class-association-rule mining using genetic network programming," *IEEE Trans. Syst. Man. Cybern. Part C Appl. Rev.*, vol. 41, no. 1, pp. 130–139, 2011.

[13] R. E. Schapire, "A brief introduction to boosting," in *Proc. 16th Internation joint conference on Artificial Intelligence*, 1999.

[14] M. Roesch, "Snort- lightweight intrusion detection for networks," in *Proc. 13th USENIX Conference on System Administration*, pp. 229 – 238, 1999.

[15] A. Prayote and P. Compton, "Detecting anomalies and intruders," in *AI 2006: Advances in Artificial Intelligence* (A. Sattar and B.-h. Kang, eds.), vol. 4304 of *Lecture Notes in Computer Science*, pp. 1084–1088, Springer Berlin Heidelberg, 2006.

[16] G. Edwards, B. Kang, P. Preston, and P. Compton, "Prudent expert systems with credentials: Managing the expertise of decision support systems," *International Journal of Biomedical Computing*, vol. 40, no. 2, pp. 125–132, 1995.

[17] *Apache Hadoop, http://hadoop.apache.org/.*