Big Data Analytics and Anomaly Detection in Wireless Cellular Networks

Rashmi I Chamakeri[1], Roshnee Krishnamurthy[2], T Roja[3], Vaishnavi M[4], Dr. Jitendranath Mungara, HOD and guide, Department of Information Science and Engineering, New Horizon College of Engineering.

ABSTRACT

Big Data is a large data set which can be analysed to identify necessary information in the form of patterns that usually relate to human behaviour and their interactions. This large amount of data is generated from various sources such as mobile networks. The evolution of technologies, like 5G, is expected to generate an enormous amount of data. The information generated in the mobile network can be managed and optimized with the help of technologies such as Big Data approach to Self Organized Networks (SONs) and Big Data Empowered SONs (BES). In cellular networks, due to large number of users, there is a possibility of unusual behaviour, for example anomalies in user calling activities. CADM, Call detail record based Anomaly Detection Method, is used to detect these anomalies. As this technology affects the performance and security of mobile networks, we can exploit a combination of other technologies such as baseline, K-means Clustering and PSO (Particle Swarm Optimization). The results obtained from carrying out these technologies are promising in terms of detection and false alarm rate.

Keywords: *Big data, Self-Organized Networks, anomalies, cellular networks, 5G.*

I. INTRODUCTION

Cellular networks induce a large amount of data and 5G networks further increases the data and it is expected to become even more complex. In order to address upcoming challenges, SONs approach is used which leads to a number of limitations and the effectiveness in a 5G scenario would not be sufficient to achieve the global network optimization and operational cost reduction goals. Thus, in the future we can opt for Big Data Empowered SONs (BES) as the most effective solution to autonomic network management for 5G systems. In this enormous data there could be anomalies during calling activities due to large amount of users. These anomalies can be detected using advanced technologies like CADM using CDR. The solution to this problem is to apply knowledge based anomaly detection methods and set rule policies depending on network behaviour. Here, 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 flexible as well as robust for the detection of anomalies. We use an approach for anomaly detection by analysing call-detail records in combination with recent Big Data analytical tools (Hadoop (HDFS, Map-Reduce)).

Since this algorithm affects the performance and security of the users, we can use other wide range of algorithms that can be applied for detecting anomalies with the most efficient one being clustering techniques. The most important, unsupervised learning processes such as K-means clustering algorithm which is a very simple algorithm for finding useful patterns. Its inability to escape from local optima can be overcome by combining with the PSO algorithm. PSO is a high efficient heuristic technique having the capability to escape from local optima and with low computational complexity. The anomaly detection technique combines the K-means and the PSO algorithm and it is performed by comparing real traffic and clusters centroids thus enhancing the performance and securing the privacy of the users.

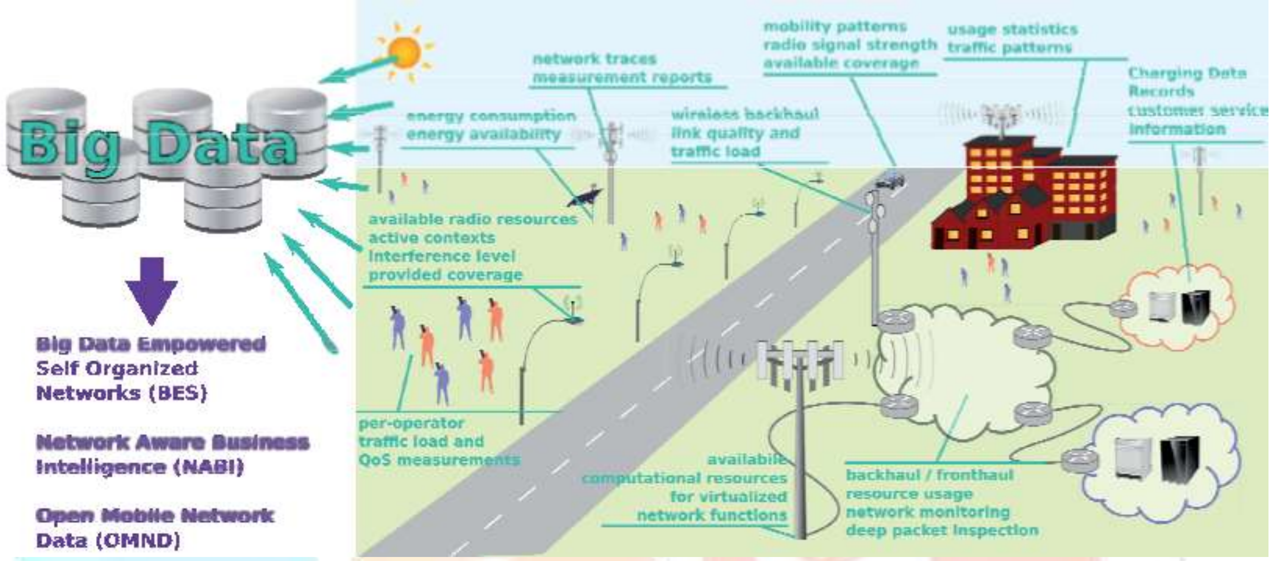
II. RELATED WORK:

SOURCES OF INFORMATION WITHIN A MOBILE NETWORK:

The information that we mentioned in the introduction comes from various network elements, such as base stations, mobile terminals, gateways and management entities, and these can be categorised as follows:

1. Control information related to regular short-term network operation, covering functionalities such as call/session set-up, release and maintenance, security, QoS, idle and connected mode mobility, and radio resource control.

2. Control information related to SON functions associated with optimization and maintenance of cellular systems, and which cover aspects such as radio link failure statistics, inter-cell interference and cell load signalling, etc.
3. Management information covering long-term network operation functionalities, such as Fault, Configuration, Accounting, Performance and Security management (FCAPS), as well as customer and terminal management. An example of such information is that defined for Operation and Management (OAM), which consists of aggregated statistics on network performance, such as number of active users, active bearers, successful/failed handover events, etc. per base station, as well as information gathered by means of active probing
4. Authentication, Authorization and Accounting (AAA) information, including for example Charging Data Records (CDRs).
5. Customer relationship information, e.g., complaints about bad service quality, churn information, etc.



The Big Data approach to 5G Networks

Big-data Empowered SONs (BES):

By leveraging all the information available within a mobile network, a Big-data Empowered SONs (BES) approach can bring mobile networks to improve the operational processes and culture of the organisation to meet the expectations of the customers. Some of the practical examples are mentioned below:

- A BES would autonomously recognize the cause of the problem on basis of data which was recorded previously and restore it with little or no engineer interference, where as in simple threshold based network fault detection systems that needs an engineer interference to run measurements to identify the source of the problem.
- By analysing the correlation among conflicting performance goals of various SON functions, and dynamically specifying operating point that gives the best performance trade-off, BES would solve the coordination between SON functions, which is the current open challenge.
- Traditionally data is inaccessible for network management purposes, such as CDRs, it could be leveraged by a BES, for example to determine the users typical mobility patterns , and proactively get the network optimization actions, such as taking informed handover decision, concentrating spectrum resources in location with larger number of users, and when base stations are foreseen to be not needed, turn off the base stations.

Using the above BES technology we collect CDRs data and use an anomaly detection scheme based on user's call-detail record (CDR) activities that is collected in a big data platform. The main design of this algorithm is to detect suspicious CITY ID and CELL ID pairs which are anomalies by monitoring the user's call detail activities. The categorization can be either normal or anomalous.

A user j's city-arrival time (a^{j}_{tk}) and city-departure time (d^{j}_{tk}) 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. We can thus define R_j = {P1,P2,P3,...,PN} as the sequence of cities travelled by the user, j. We determine a quintuple definition in R_j to represent the travel activities of a mobile user j for a specific city k: P_k =(ID_j, $l_k(a^{j}_{tk})$, $l_k(d^{j}_{tk})$, a^{j}_{tk}, d^{j}_{tk})where ID_k is the MSISDN of user j, $l_k(a^j_{tk})$ and $l_k(d^j_{tk})$ define the (CITY ID, CELL ID) pairs for arrival and departure location time information respectively. Definition of these attributes simplifies the tracking of users' inter-city traveling activities.

• Let TD represents the inter-city travel for a user between P_k and P_{k+1} in his traveling route R_j. The k^{th} TD is defined as TD_k =($a^j{}_{tk+1}$ −d^j_{tk}).

• Let CD_k represent the average distance between cities of l_k and l_{k+1}.

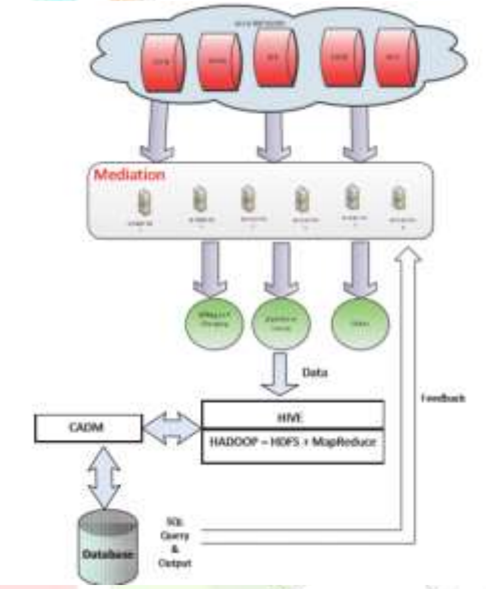
• Let TV represents the inter-city travel velocity for a user between P_k and P_{k+1} in his traveling route R_j. TV is defined as TV_k = CD_k/TD_k

• Let $MPTV_k$ represents maximum possible travel velocity between two consecutive cities l_k and l_{k+1} in a given route $R_{j.}$

The following algorithm is the pseudo code of the proposed CDR based Anomaly Detection Method (CADM). It is originally taken from a rule based approach 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. We also abolish the non-traveling users (i.e. users where Rj = P1) during analysis and obtained a total of M traveling users.

In step 2 of the algorithm, a rule-based technique is proposed to flag anomalous activities of users traveling between two different cities by calculating the travel velocities and then comparing this velocity with $MPTV_k$. We could then decide whether there exists any anomaly by just comparing these two velocities with MPTV. Then, this information is stored in a relational database such as MySQL.



Algorithm: CDR based Anomaly Detection Method (CADM)

Inputs: ANALY SIS 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.

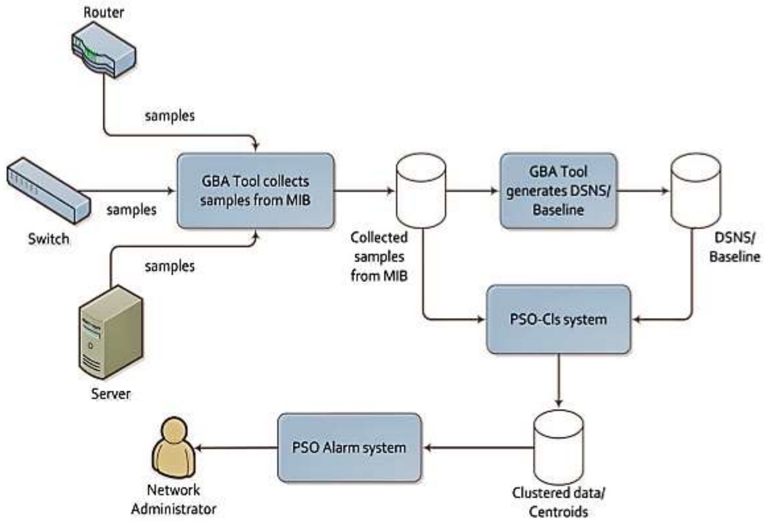
Outputs: List of anomalous CELL ID and CITY ID pair records

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 TY PE. Include users only who have done call record activities in at least two distinct CITY IDs during ANALY SIS DURATION. Group the records by MSISDN and sort them by CALL DATE, CALL TIME, then store the result in CDR ANALY SIS TABLE.
        2) for all j= 1,....,M do
        Obtain the route Rj for each user j in CDR ANALY SIS TABLE where $R_j$ = {P1,P2,P3,...,PN} and $P_k$ = ($ID_j$, $l_k(a^j_{tk})$, $l_k(d^j_{tk})$, $a^j_{tk}$,$d^j_{tk}$) by traversing over each row record of CDR ANALY SIS 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 $TV_{k−1}$ >MPTVk−1, then flag the $_{lk−1}$ 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.

In recent years, many anomaly detection techniques have been developed which is an open research area and most of them focus on maximizing the detection rate and minimizing the false alarm rate.


Router
samples
GBA Tool collects samples from MIB
samples
Switch
samples
Server
Collected samples from MIB
GBA Tool generates DSNS/ Baseline
DSNS/ Baseline
PSO-Cls system
Clustered data/ Centroids
PSO Alarm system
Network Administrator

Anomaly detection model

The initial stage to detect anomalies is to adapt to a model that characterizes the network traffic efficiently, which represents a significant challenge due to the non-stationary nature of network traffic. Large networks traffic behaviour is composed by daily routines where traffic levels are usually higher and distinct during working hours and are also different for weekends. In this work the GBA tool is used to generate different profiles of normal behaviour of the network.

K-means is a well-known clustering algorithm created by J. MacQueen. It can be used for unsupervised learning of neural networks, pattern recognitions, clustering analysis and more. The algorithm classifies data sets based on attributes into K groups. The grouping is performed by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The K-means algorithm suffers from the absence of diversity mechanism to escape from local optimum. Thus, in order to overcome this drawback and simultaneously keeps computational complexity under control, mainly because for high-dimensional problems complexity is a concern, the Kmeans algorithm can be associated to PSO.

The PSO is an evolutionary computation technique based on swarm intelligence which is very powerful since it is able to escape from global optima and also it tries to keep a simple structure.

In the PSO strategy, each candidate-vector at nth iteration, defined as $\mathrm{p}i[n]$ with $D\times1$ dimension, is used for the velocity calculation of next iteration as:

$$\mathrm{v}_i\,[n + 1] = \omega\cdot\mathrm{v}_i[n]+\phi+ \cdot\mathrm{U}_{i1}[n](\mathrm{p}^{\mathrm{best}}{}_i - \mathrm{p}_i[n]) + \phi_2 \cdot\mathrm{U}_{i2}[n](\mathrm{p}^{\mathrm{best}}{}_g - \mathrm{p}_i[n]) \quad (1)$$

where ω is the inertia weight, adopted as an unitary value in this work, for simplicity; $\mathrm{U}_{i1}\,[n]$ and $\mathrm{U}_{i2}\,[n]$ are diagonal matrices with dimension D, and elements are random variables with uniform distribution $\sim\mathcal{U}\in[0, 1]$, generated for the i^{th} particle at iteration $n = 1, 2,...,N$; $\mathrm{p}^{\mathrm{best}}{}_g$ and $\mathrm{p}^{\mathrm{best}}{}_i$ are the best global position and the best local positions found until the n^{th} iteration, respectively; ϕ_1 and ϕ_2 are acceleration coefficients regarding the best particles and the best global positions influences in the velocity updating, respectively. The i^{th} particle's position at iteration n is a clustering candidate-vector $\mathrm{p}[n]$ of size $D\times 1$.

The position of each particle is updated using the new velocity vector (1) for that particle, according to:

$$\mathrm{p}_i[n + 1] = \mathrm{p}_i[n] + \mathrm{v}_i[n + 1],\ i = 1,...,M \quad (2)$$

The PSO algorithm consists of repeated application of the velocity and position updating equations until a stopping criteria is found. The stop criteria can be a fixed number of iteration or determined by the non-improvement in the solution when the algorithm evolves. In order to reduce the likelihood that the particle might leave the search universe, maximum velocity $V\mathrm{m}$ factor is added to the PSO model (1), which will be responsible for limiting the velocity in the range $[\pm V\mathrm{m}]$. The adjustment of velocity allows the particle to move in a continuous but constrained subspace, been simply accomplished by:

$v_i[n] = \min\{V_\mathrm{m}; \max\{-V_\mathrm{m}; v_i[n]\}$ (3)

From (3) it is clear that if $|vi[n]|$ exceeds a positive constant value Vm specified by the user, the i^{th} particle' velocity is assigned to be sign $(v_i[n])\ V_\mathrm{m}$, i.e. particles velocity on each of $D-$dimension is clamped to a maximum magnitude V_m. If we could define the search space by the bounds $[P_\mathrm{min}; P_\mathrm{max}]$, then the value of Vm will be typically set to $V_\mathrm{m} = \tau\ (P_\mathrm{max} - P_\mathrm{min})$, where $0.1 \leq \tau \leq 1.0$. In this work, the objective function to be minimized by PSO is the sum of Euclidean distances of the candidate-vector regarding each data point of the K^{th} cluster generated by K-means, given by:

$J(\mathrm{p}) = {}^{K}\sum_{k=1} {}^{S}\sum_{s=1} \sqrt{}|\mathrm{p}k\ s - \mathrm{c}k|^2$ (4)

Where K is the number of clusters, S is the number of traffic samples and ck is the kth cluster centroid.

This clustering-based anomaly detection algorithm showed robustness against false alarm while held good anomaly detection rates, achieving 82.92% detection rate with 2.85% false alarm rate for the test. Thus, the proposed approach can be extended to other types of anomalies while improving the detection and false alarm rate.

III. CONCLUSION AND FUTURE WORK

The concept of big data and its empowered organized networks were introduced along with the sources of information from various fields like wireless mobile networks. Also, there is a brief discussion on anomaly detection in wireless networks using CADM technology with the help of CDRs. To enhance the security and management of the networks, we discussed the evolved technologies like K-means clustering and PSO technology. All of the above techniques mainly aims to detect the anomalies effectively by increasing the detection rate and minimizing the false alarm rate. In future work, we can avoid the location tracking of the users which affects the security of users that exists in the above technology.

IV. REFERENCES

[1] Nicola Baldo, Lorenza Giupponi, Josep Mangues-Bafalluy, "Big Data Empowered Self Organized Networks", Centre Tecnològic de Telecomunicacions de Catalunya (CTTC) 08860 Castelldefels, Barcelona, Spain

[2] Ilyas Alper Karatepe, Engin Zeydan, "Anomaly Detection In Cellular Network Data Using Big Data Analytics"

[3] Mois´es F. Lima∗, Bruno B. Zarpel˜ao†, Lucas D. H. Sampaio∗, Joel J. P. C. Rodrigues‡, Taufik Abr˜ao∗ and Mario Lemes Proenc¸a Jr.∗, "Anomaly detection using baseline and K-means clustering", Computing Science Department, State University of Londrina (UEL), Londrina, Brazil †School of Elect. & Comp. Engineering, University of Campinas (UNICAMP), Campinas, Brazil ‡Instituto de Telecomunicac¸˜oes, University of Beira Interior, Covilh˜a, Portugal E-mails: {moisesflima, brunozarpelao, lucas.dias.sampaio}@gmail.com, joeljr@ieee.org, {taufik, proenca}@uel.br