Sequence-based detection of sleeping cell failures in mobile networks

Fedor Chernogorov1,2 • Sergey Chernov2 • Kimmo Brigatti2 • Tapani Ristaniemi2

Abstract This article presents an automatic malfunction detection framework based on data mining approach to analysis of network event sequences. The considered environment is long term evolution (LTE) of Universal Mobile Telecommunications System with sleeping cell caused by random access channel failure. Sleeping cell problem means unavailability of network service without triggered alarm. The proposed detection framework uses N-gram analysis for identification of abnormal behavior in sequences of network events. These events are collected with minimization of drive tests functionality standardized in LTE. Further processing applies dimensionality reduction, anomaly detection with K-Nearest Neighbors, cross-validation, postprocessing techniques and efficiency evaluation. Different anomaly detection approaches proposed in this paper are compared against each other with both classic data mining metrics, such as F-score and receiver operating characteristic curves, and a newly proposed heuristic approach. Achieved results demonstrate that the suggested method can be used in modern performance monitoring systems for reliable, timely and automatic detection of random access channel sleeping cells.

Keywords Data mining • Sleeping cell problem • Anomaly detection • Performance monitoring • Self-healing • LTE networks

1 Introduction

Modern cellular mobile networks are becoming increasingly diverse and complex, due to coexistence of multiple Radio Access Technologies (RATs), and their corresponding releases. Additionally, small cells are actively deployed to complement the macro layer coverage, and this trend will only grow. In the future this situation is going to evolve towards even higher complexity, as in 5th Generation (5G) networks there will be much more end-user devices, served by different technologies, and connected to cells of different types [21, 33, 59, 62, 63]. New applications and user behavior patterns are daily coming into play. In such environment, network performance and robustness are becoming critical values for mobile operators. In order to achieve these goals, efficient flow of Quality and Performance Management (QPM) [36], which is a sequence of fault detection, diagnosis and healing, should be developed and applied in the network in addition to other optimization functions.

The concept of self-organizing network (SON) [57, 58] has been proposed to automate and optimize the most tedious manual tasks in mobile networks, including QPM. Automation is the key idea in SON and it has been proposed for self-configuration, self-optimization and self-healing in LTE and UMTS networks [27, 36, 68]. In traditional systems detection, diagnosis and recovery of
network failures is mostly manual task, and it is heavily based on pre-defined thresholds, aggregation and averaging of large amounts of performance data—so called Key Performance Indicators (KPIs). Self-healing [32, 67] automates the functions of QPM process to improve reliability of network operation. Though, self-healing is still among the least studied functions of SON at the moment, and the developed solutions and use cases require improvement prior to application in the real networks. This is especially important for non-trivial network failures such as sleeping cell problem [13, 14, 36]. This is a special term used to denote a breakdown, which causes partial or complete degradation of network performance, and which is hard to detect with conventional QPM within reasonable time. Thus, in the research and standardization community automatic fault detection and diagnosis functions, enhanced with the most recent advancements in data analysis, are seen as the future of self-healing. Thus, development of improved self-healing functions for detection of sleeping cell problems, through application of anomaly detection techniques is of high importance nowadays. This article presents a novel framework based on N-gram analysis of MDT event sequences for detection of random access channel sleeping cells.

The rest of this paper is organized as follows. Section 2 describes common practices of quality and performance management in mobile networks, including MDT functionality, and advanced methods based on knowledge mining algorithms. Section 3 defines the concept of sleeping cell and its possible root cause failures. In Sect. 4 simulation environment, assumptions and random access channel problem are presented. Also Sect. 4 describes the generated and analyzed performance MDT data. Sect. 5 concentrates on the suggested sleeping cell detection knowledge mining framework. It includes overview of the applied anomaly detection methods: KNN anomaly outlier scores, N-gram, minor component analyses, postprocessing and data mining performance evaluation techniques. Section 6 is devoted to the actual research results. Data structures at different stages of analysis are shown, and efficiency of different postprocessing methods is compared. In Sect. 7 the concluding remarks regarding the findings of the presented research are given.

2 Quality and performance management in cellular mobile networks

Performance management in wireless networks includes three main components: data collection, analysis and results interpretation. Data gathering can be done either by aggregation of cell-level statistics—collection of KPIs, or collection of detailed performance data with drive tests. The main weaknesses in analysis of KPIs are that a lot of statistics is left out at the aggregation stage, due to averaging over time, element and because fixed threshold values are applied. Even though drive test campaigns provide far more elaborate information regarding network performance, they are expensive to carry out and do not cover overall area of network operation. Root cause analysis is done manually in majority of cases, and because of that there is a room for more intelligent approaches to detection and diagnosis of network failures, e.g. with data mining and anomaly detection techniques. This would provide possibility to automate performance monitoring task furthermore.

2.1 Minimization of drive tests

Yet another way to improve network QPM is to collect a detailed performance database. This is enabled with MDT functionality standardized in 3rd Generation Partnership Project (3GPP) [28]. MDT is designed for automatic collection and reporting of user measurements, where possible complemented with location information. Collected data is then reported to the serving cell, which in turn sends it to MDT server [39]. Thus, large amount of network and user performance is available for analysis. This is where the power of data mining and anomaly detection can be applied.

Specification describes several use cases for MDT: improvement of network coverage, capacity, mobility robustness and end user quality of service [36]. According to the standard, MDT measurements and reporting can be done both in idle and connected Radio Resource Control (RRC) modes. In logged MDT, User Equipment (UE) stores measurements in memory, and reporting is done at the next transition from idle to connected state. In immediate MDT, measurements are reported as soon as they are done through existing connection. In turn, there are two measurement modes in immediate MDT: periodic and event-triggered [39]. Periodic measurements are very useful for initial network deployment coverage and capacity verification as they provide detailed map of network performance, say in terms of signal propagation or throughput. The main disadvantage of periodic measurements is that they consume a lot of network and user resources. In contrast, event-triggered approach provides less information regarding the network status, but can be very efficient for mobility robustness and resource savings. In our study, immediate event-triggered MDT is used for collection of performance database. Table 1 presents the list of network events which triggered MDT measurements and reporting.

2.1.1 Location estimation in MDT

One of the important features of MDT is collection of geo-location information at the measurement time.
moments. Whenever UE location is provided in MDT report there are several ways to associated it with particular cell, such as: serving cell ID, dominance maps and a new approach based on target cell ID information.

Serving cell ID is available with MDT event-triggered report, even for early releases of LTE. However, in case of coverage hole or problems with new connection establishment, this approach can lead to mistakes in UE location association, because the faulty cell would never become serving in the worst case scenario. This limits the usage of serving cell method for sleeping cell detection. To overcome the problem presented above, a dominance maps method can be used. This is a map, which demonstrates the E-UTRAN Node B (eNB)\(^1\) with the strongest radio signal in each point of the network, see Fig. 1. The main advantage of dominance maps is that mapping of cell ID to location coordinate of UE MDT measurement is very precise, and this results in higher accuracy of sleeping cell detection. The downside of dominance maps approach is that it requires a lot of detailed input measurement information. Though, MDT functionality is one of the efficient ways to create such maps [25].

The last method for cell ID and UE report location association uses target cell ID feature. This information is available in the network events A3 RSRP, HO COMMAND, HO COMPLETE and RLF REESTABLISHMENT. The strong side of this method is that detection of sleeping cell becomes possible with a very limited amount of information, as it is shown in Sect. 6.

The key aspects which should be taken into account when selecting a location association method are accuracy and amount of information to create mapping between cell and user location.

---

\(^1\) Evolved Universal Terrestrial Radio Access Network (E-UTRAN)
the existing methods fail to detect network anomalies in an efficient and timely manner. This becomes problematic, especially in networks where user demand is high and network operators need to ensure optimal performance.

To address these challenges, the authors propose a new approach that leverages machine learning techniques to detect network anomalies in real-time. The proposed method involves collecting network data, processing it using a sequence-based analysis method, and then applying advanced data mining techniques to identify anomalies. This approach is designed to be scalable and adaptable to different network environments.

The method is implemented using a combination of statistical analysis and machine learning algorithms. The data collected is first processed to extract relevant features, and then these features are used to train the machine learning models. The models are then used to predict network anomalies, which can be used to trigger interventions to resolve issues in a timely manner.

The effectiveness of the proposed method is demonstrated through extensive simulations and real-world case studies. The results show that the proposed method can accurately detect network anomalies, even in complex and highly dynamic network environments. This makes it a valuable tool for network operators looking to improve the reliability and performance of their networks.

In summary, the proposed method provides a promising solution to the problem of network anomaly detection. By leveraging machine learning techniques, it offers a flexible and scalable approach that can be tailored to the specific needs of different network operators. The results also highlight the need for continued research and development in this area to further improve network performance and reliability.
experience. Classification of sleeping cells, depending on the extent of performance degradation from the lightest, to the most severe \[13, 15\]: impaired or deteriorated—smallest negative impact on the provided service, crippled—characterized by a severely decreased capacity, and catatonic—kind of outage which leads to complete absence of service in the faulty area, such cell does not carry any traffic.

Degradation can be caused by malfunction of different hardware or software components of the network. Depending on the failure type, different extent of performance degradation can be induced. In this study the considered sleeping problem is caused by Random Access Channel (RACH) failure. This kind of problem can appear due to RACH misconfiguration, excessive load or software/firmware problem at the eNB side \[1, 73\]. RACH malfunction leads to inability of the affected cell to serve any new users, while earlier connected UEs still get served. This problem can be classified to crippled sleeping cell type, and with time the affected cell tends to become catatonic. In many cases RACH problem becomes visible for the operator only after a long observation time or even due to user complains. For this reason, it is very important to timely detect such cells and apply recovery actions.

### 3.1 Random access sleeping cell

Malfunction of RACH can lead to severe problems in network operation as it is used for connection establishment in the beginning of a call, during handover to another cell, connection re-establishment after handover failure or Radio Link Failure (RLF) \[69\]. Malfunction of random access in cell with ID 1, is caused by erroneous behavior of T304 timer \[30\], which expires before random access procedure is finished. Modeling of this failure is done so that at certain moment of network operation cell 1 loses capability to successfully go through random access procedure. Thus, whenever UE tries to initiate random access to this cell, this attempt fails. Malfunction area covers around 5% of the overall network (1 out of total 21 cells).

### 4 Experimental setup

#### 4.1 Simulation environment

Experimental environment is dynamic system level simulator of LTE network, designed according to 3GPP Releases 8, 9, 10 and partly 11. Throughput, spectral efficiency and mobility-related behavior of this simulator are validated against results from other simulators of several companies in 3GPP \[31, 50, 52\]. Step resolution of the simulator is one Orthogonal Frequency-Division Multiplexing (OFDM) symbol. Methodology for mapping link level Signal to Interference plus Noise Ratio (SINR) to the system level is presented in \[7\]. Simulation scenario is an improved 3GPP macro case 1 \[29\] with wrap-around layout, 21 cells (7 base stations with 3-sector antennas), and inter-site distance of 500 meters. Modeling of propagation and radio link conditions includes slow and fast fading. Users are spread randomly around the network, so that on average there are 15 dynamically moving UEs per cell. The main configuration parameters of the simulated network are shown in Table 2.

#### 4.2 Generated performance data

Generated performance data includes dominance map information and MDT log, which contains the following fields:

- MDT triggering event ID. The list of possible events is presented in Table 1. This is a categorical (nominal) and sequential data, i.e. sequences of events are meaningful from data mining perspective;
- UE ID. This is also categorical data;
- UE location coordinates [m]. It is numerical, spatial data;
- Serving and target cell ID – spatial, categorical data.

It is important to know the type of the analyzed data to construct efficient knowledge mining framework \[9, 37\].

Simulations done for this study cover three types of network behavior: “normal” – network operation without random access sleeping cell; “problematic” – network with RACH failure in cell 1; “reference” – no sleeping cell, but different slow and fast fading maps, i.e. if compared to “normal” case, propagation-wise it is a different network. The latter case is used for validation purposes. All three of these cases have different mobility random seeds, i.e. call start locations and UE traveling paths are not the same. Each of these 3 cases is represented with 6 data chunks. The training and testing phases of sleeping cell detection are done with pairs of MDT logs by means of K-fold approach \[37\]. For example, “normal”-“problematic”, or “normal”-“reference” cases are considered. Thus, in total there are 72 unique combinations of analyzed MDT log pairs, which is rather statistically reliable data base.

### 5 Sleeping cell detection framework

The core of the presented study is sleeping cell detection framework based on knowledge mining, Fig. 2. Both training and testing phases are done in accordance to the process of Knowledge Discovery in Databases (KDD), which includes the following steps \[24, 37\]: data cleaning, integration from different sources, feature selection and
The constructed data analysis framework for sleeping cell detection is semi-supervised, because unlabeled error-free data is used for training of the data mining algorithms. The analysis can be logically separated to two parts: identification of the anomalous data points in MDT data and localization of these points in the real network and assignment of the real sleeping cell score to each cell (can be treated as extent of cell performance abnormality). The first problem is solved with preprocessing and pattern recognition, while the latter is more a task of pattern evaluation and postprocessing. In testing phase problematic data is analyzed to detect abnormal behavior. Reference data is used for testing in

Table 2 General simulation configuration parameters

| Parameter                        | Value                     | Parameter                        | Value  |
|----------------------------------|---------------------------|----------------------------------|--------|
| Cellular layout                  | Macro 21 Wrap-around      | Number of cells                  | 21     |
| UEs per cell                     | 17                        | Inter-Site Distance              | 500 m  |
| Link direction                   | Downlink                  | RRC IDLE mode                    | Disabled|
| User distribution in the network | Uniform                   | Maximum BS TX power              | 46 dBm |
| Initial cell selection criterion | Strongest RSRP value      | Handover margin (A3 margin)      | 3 dB   |
| Handover time to trigger         | 256 ms                    | Hybrid Adaptive Repeat and reQuest (HARQ) | Enabled |
| Slow fading standard deviation   | 8 dB                      | Slow fading resolution            | 5 m    |
| Simulation length                | 572 s (9.5 min)           | Simulation resolution             | 1 time step = 71.43 μs |
| Network synchronicity mode       | Asynchronous              | Max number of UEs/cell           | 20     |
| UE velocity                      | 30 km/h                   | Call duration                     | 90 s   |
| Traffic model                    | Constant Bit Rate 320 kbps| Normal and Reference cases       |        |
| Problematic case                 | Simulation with RACH problem in cell 1 | A2 RSRP Hysteresis | 3       |
| A2 RSRP Threshold                | -110                      | A2 RSRP Hysteresis               | 2       |
| A2 RSRQ Threshold                | -10                       |                                   |        |

Fig. 2 Sleeping cell detection framework
order to verify how much the designed framework is prone to make false alarms.

5.1 Feature selection and extraction

Feature selection and extraction is the first step of sleeping cell detection. At this stage, input data is prepared for further analysis. Preprocessing is needed as reported UEs MDT event sequences have variable lengths, depending on the user call duration, velocity, traffic distribution and network layout.

5.1.1 Sliding window preprocessing

Sliding window approach [64] allows to divide calls to sub-calls of constant length, and by that to unify input data. There are two parameters in sliding window algorithm: window size \( m \) and step \( n \). After transformation, one sequence of \( N \) events (a call) is represented by several overlapping (in case if \( n < m \)) sequences of equal sizes, except for the last sub-call, which is the remainder from \( N \) modulo \( n \).

In the presented results overlapping sliding window size is 15, and the step is 10 events. Such setup allows to maintain the context of the data after processing [49]. The number of calls and sub-calls for all three data sets are shown in Table 3.

5.1.2 N-gram analysis

When input user-specific MDT log entries are standardized with sliding window method, the data is transformed from sequential to numeric format. It is done with N-gram analysis method, widely used e.g. for natural language processing and text analysis applications such as speech recognition, parsing, spelling, etc. [6, 8, 35, 45, 56]. In addition, N-gram is applied for whole-genome protein sequences [26] and for computer virus detection [16, 23].

N-gram is a sub-sequence of \( N \) overlapping items or units from a given original sequence. The items can be characters, letters, words or anything else. The idea of the method is to count how many times each sub-sequence occurs. This is the transformation from sequential to numerical space.

Here is an example of \( N \)-gram analysis application for two words: ‘performance’ and ‘performer’, \( N = 2 \), and a single unit is a character. The resulting frequency matrix after \( N \)-gram processing is shown in Table 4. In case of sequence analysis of MDT data, a letter from this example corresponds to an MDT event given in Table 1. Thus, for 2-gram analysis pairs of network events are considered, such “PL PROBLEM - RADIO LINK FAILURE”, or “A3 RSRP—HO COMMAND”.

5.2 Dimensionality reduction with minor component analysis

Dimensionality reduction is applied to convert high-dimensional data to a smaller set of derived variables. In the presented study Minor Component Analysis (MCA) method is applied [54]. This algorithm has been selected on the basis of comparison with other dimensionality reduction methods such as Principal Component Analysis (PCA) [47] and diffusion maps [22]. MCA extracts components of covariance matrix of the input data set and uses minor components (eigenvectors with the smallest eigenvalues of covariance matrix). 6 minor components are used as a basis of the embedded space. This number is defined by means of Second ORder sTatistic of the Eigenvalues (SORTÉ) method [42, 43].

5.3 Pattern recognition: K-NN anomaly score outlier detection

In order to extract abnormal instances from the testing dataset K-NN anomaly outlier score algorithm is applied. In contrast with K-NN classification, method is not supervised, but semi-supervised, as the training data does not contain any abnormal labels. In general, there are two approaches concerning the implementation of this algorithm; anomaly score assigned to each point is either the sum of distances to \( k \) nearest neighbors [2] or distance to \( k \)-th neighbor [66]. The first method is employed in the presented sleeping cell detection framework, as it is more statistically robust. Thus, the algorithm assigns an anomaly score to every sample in the analyzed data based on the sum of distances to \( k \) nearest neighbors in the embedded

| Table 3 | Number of calls and sub-calls in analyzed data |
|---------|-----------------------------------------------|
| Amount / Dataset | Normal | Problem | Reference |
| Calls (all) | 2530 | 1940 | 2540 |
| Sub-calls (all) | 7230 | 7134 | 7201 |
| Normal sub-calls | 6869 | 5932 | 6821 |
| Abnormal sub-calls | 361 | 1202 | 380 |

| Table 4 | Example of N-gram analysis per character, \( N = 2 \). |
|---------|------------------------------------------------------|
| Analyzed word | pe | er | rf | fo | rm | ma | me | an | nc | ce |
| performance | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| performer | 1 | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
space. Euclidean metric is applied as similarity measure. Points with the largest anomaly scores are called outliers. Separation to normal and abnormal classes is defined by threshold parameter $T$, equal to 95th percentile of anomaly scores in the training data.

Configuration parameters of data analysis algorithms in the presented sleeping cell detection framework are summarized in Table 5.

### 5.4 Pattern evaluation

The main goal of pattern evaluation is conversion of output information from K-NN anomaly score algorithm to knowledge about location of the network malfunction, i.e. RACH sleeping cell. This is achieved with postprocessing of the anomalous data samples through analysis of their correspondence to particular network elements, such as UEs and cells. For this purpose we developed 4 post-processing methods: Dominance Cell Sub-Call Deviation, Dominance Cell 2-Gram Deviation, Dominance Cell 2-Gram Symmetry Deviation, and Target Cell Sub-Calls. The essence of these methods, discussed throughout this section, is reflected in their names. The first part describes which geo-location information is used for mapping data samples to cells, e.g. dominance map information, target or serving cell ID. The second part denotes what is used as feature space for postprocessing. It can be either “sub-calls”, when rows of the dataset are used as features or “2-gram”, when individual event pair combinations, i.e. columns of the dataset are used as features. The last, third part of the method name describes analysis considers the difference between training and testing data (“deviation” keyword), or whether only information about testing set is used to build sleeping cell detection histogram.

Output from the postprocessing methods described above is a set of values—sleeping cell scores, which correspond to each cell in the analyzed network. High value of this score means higher abnormality, and hence probability of failure. To achieve clearer indication of problematic cell presence, additional non-linear transformation is applied. It is called amplification, as it allows to emphasize problematic areas in the sleeping cell histogram. Sleeping cell score of each cell is divided by the sum of Sleeping Cell (SC) scores of all non-neighboring cells. Sleeping cell scores, received after postprocessing and amplification are then normalized by the cumulative SC score of all cells in the network. Normalization is necessary to get rid of dependency on the size of the dataset, i.e. number of calls and users.

### 5.5 Knowledge interpretation and presentation

The final step of the data analysis framework is visualization of the fault detection results. It is done with construction of a sleeping cell detection histogram and network heat map. However, sleeping cell histogram does not show how cells are related to each other: are they neighbors or not, and which area of the network is causing problems. Heat map method shows more anomalous network regions with darker and larger spots, while normally operating regions are in light grey color. The main benefit of network heat map is that mobile network topology and neighbor relations between cells are illustrated.

#### 5.5.1 Performance evaluation

To apply data mining performance evaluation metrics labels of data points must be known. Cell is labeled as abnormal if its SC score deviates more than $3\sigma$ (standard deviation of sleeping cell scores) from the mean of SC score in the network. Mean value and standard deviation of the sleeping cell scores are calculated altogether from 72 runs produced by K-fold method for “normal”-“problematic”, and “normal”-“reference” dataset pairs. Availability of the labels and the outcomes of different postprocessing methods enables application of such data mining performance metrics as accuracy, precision, recall, F-score, True Negative Rate (TNR) and False Positive Rate (FPR) [34]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$ \hfill (1)

$$Precision = \frac{TP}{TP + FP}$$ \hfill (2)

$$Recall = \frac{TP}{TP + FN} = TP_{rate}$$ \hfill (3)

$$FP_{rate} = \frac{FP}{FP + TN}$$ \hfill (4)

$$Fscore = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$ \hfill (5)

In these equations $TP$, $TN$, $FP$, $FN$ denote elements of confusion matrix [37, 38, 41], and represent correspondingly the number of true positive, true negative, false
Fig. 3 Normal dataset used for training of the sleeping cell detection framework. a Normal training dataset in the embedded space. b Sorted outlier scores of normal training dataset

Fig. 4 Problematic dataset used at the testing phase of the sleeping cell detection framework. a Problem testing dataset in the embedded space. b Sorted outlier scores of problem testing dataset

Fig. 5 Reference dataset used at the testing phase of the sleeping cell detection framework. a Reference testing dataset in the embedded space. b Sorted outlier scores of reference testing dataset
positive and false negative points. On the basis of these scores ROC curves are plotted.

In addition to the conventional performance evaluation metrics described above, a heuristic method is applied to complement the analysis. This approach measures how far is the achieved performance from the a priori known ideal solution. Performance of the sleeping cell detection algorithm can be described by a point in the space “cumulative standard deviation”-“sleeping cell magnitude”. “Sleeping cell magnitude” is the highest SC score, which can reach value 100 due to normalization. “Cumulative standard deviation” coordinate equals to the standard deviation of SC scores of all other cells. This plane contains two points of interest: \([0; 100]\) and \(\frac{N_{\text{cellsinthenetwork}}}{N_{\text{cellsinthenetwork}}} = \frac{1}{100}\). In case of malfunctioning network, the ideal sleeping cell detection algorithm assigns 100 value of SC score to the broken cell and zero values to the rest cells. Thus, the corresponding point \([0; 100]\) is calculated. In case of error-free network, the ideal performance is mapped to the point \(\frac{1}{N_{\text{cellsinthenetwork}}} = \frac{1}{100}\), because all the cells have even SC scores equal to 100. Thus, the smaller the Euclidean distance between the achieved and ideal sleeping cell histograms, the better the performance of the sleeping cell detection algorithm.

### 6 Results of sleeping cell detection

This section presents the results of sleeping cell detection for different post-processing algorithms. In addition, the data at different stages of the detection process is illustrated. Then performance metrics are used to compare effectiveness of the developed SC identification algorithms.

#### 6.1 Preprocessing and K-NN anomaly score calculations

After preprocessing with sliding window and N-gram methods we get a so-called 2-gram popularity matrix. The size of this matrix equals to data chunk size and has 32 features—the number of non-zero 2-grams. This popularity matrix is transformed with MCA. The output of dimensionality reduction with MCA has 6 features—coordinates of points in 6-dimensional embedded space based on eigenvectors with the smallest eigenvalues. Then training MDT data is processed with K-NN anomaly score algorithm. As it is discussed in Sect. 5.3, the anomaly score threshold, used for separation of data points to normal and abnormal classes, is selected to be 95th percentile of outlier score in training data. Shape of normal training dataset in

![Fig. 6](image-url)
the embedded space is shown in Fig. 3a. In this Fig. 3 dimensions are selected on the basis of visual inspection to demonstrate best the distribution of data points. Sorted anomaly outlier scores are presented in Fig. 3b. It can be seen that data points are very compact in the embedded space, and because of that there is no big difference in the anomaly score values. The main goals of analyzing testing dataset are to find anomalies, detect sleeping cell, and keep the false alarm rate as low as possible. At the testing phase either problematic or reference data are analyzed. After the same preprocessing stages as for training, the testing data is represented in the embedded space. When testing data is problematic dataset some of the samples are significantly further away from the main dense group of points, Fig. 4. These abnormal points are labeled as outliers, and the corresponding anomaly scores for these samples are much higher, as it can be seen from Fig. 4b. On the other hand, some of the points with relatively low anomaly score are above the abnormality threshold. This means that there is still certain percentage of false alarms, i.e. some “good” points are treated as “bad”. The extent of negative effect caused by false alarms is discussed further in Sect. 6.4. Though, there is no opposite behavior referred to as “miss-detection”—none of the anomalous points are treated as normal.

Validation of the data mining framework is done by using error-free reference dataset as testing data. No real anomalies are present in the network behavior. Reference testing data in the embedded space and corresponding anomaly outlier scores are shown in Fig. 5. Only few points can be treated as outliers, and in general the shapes of normal (Fig. 3a) and reference (Fig. 5a) datasets in the embedded space are very similar. Anomaly outlier scores of the reference testing data are low for all points, except 2 outliers.

6.2 Application of postprocessing methods for sleeping cell detection

After training and testing phases certain sub-calls are marked as anomalies. The next step is conversion of this information to knowledge about location of malfunctioning cell or cells, and this is done through postprocessing described in Sect. 5.4.

![Fig. 7 Results of sleeping cell detection for dominance cell 2-gram deviation method](image-url)
6.2.1 Detection based on dominance cell sub-call deviation

In our earlier study [12] postprocessing based on dominance cells and call deviation for sleeping cell detection is presented. One problem of using calls as samples is that, in case if the duration of the analyzed user call is long, the corresponding number of visited cells is large, especially for fast UEs. Hence, even if certain call is classified as abnormal, it is very hard to say which cell has anomalous behavior. To overcome this problem, analysis is done for sub-calls, derived with sliding window method, see Sect. 5.1.1. Sub-calls contain the same number of network events, and the length of the analyzed sequence is short enough to identify the exact cell, with problematic behavior. Deviation measures the difference between training and testing data, and it is used to sleeping cell detection histogram, presented in Fig. 6a. From this figure, it can be seen that abnormal sub-calls are encountered more frequently in the area of dominance of cell 1, which has the highest deviation. One can see that there are 2 types of bars—colorful (in this case blue) and grey. The second variant implies additional postprocessing step—amplification, described in Sect. 5.4. In addition to cell 1, its neighboring cells 8, 9, 11 and 12 also have increased deviation values, as it can be seen from the network heat map in Fig. 6c. Sleeping cell detection histogram and network heat map for reference dataset used as testing are shown in Fig. 6b, d correspondingly. Even though cells 6 and 17 have higher SC scores than other cells, they are not marked as abnormal, because their abnormality does not reach mean + 3σ level.

6.2.2 Detection based on dominance cell 2-gram deviation

In this method problematic network elements are found through the comparison of 2-gram frequencies in different areas of dominance map. For this purpose we consider all sub-calls from training data set against sub-calls assigned to abnormal class from testing dataset. In case there is a big increase or decrease, the cell associated with these changes is marked as abnormal. From sleeping cell detection histogram in Fig. 7a it can be that cell 1 has a clear difference in number of 2-gram occurrences in testing data, if compared to training data. This happens because handovers toward this cell fail. Due to this fact 2-gram sequence with events related to handovers become imbalanced in testing data if compared to training data. For instance, 2-grams

Fig. 8 Results of sleeping cell detection for dominance cell 2-gram symmetry deviation method. a Problematic dataset sleeping cell detection histogram. b Reference dataset sleeping cell detection histogram. c Problematic dataset heat map. d Reference dataset heat map
like Handover (HO) Command—HO Complete and HO Complete—A2 RSRP ENTER, become very rare. On the other hand, 2-gram HO Command—A2 RSRP ENTER, which can be treated as indication of unsuccessful handovers, in opposite becomes very popular in testing data, while in training data it does not exist at all. Among the neighbors of problematic cell 1, only cell 11 has slightly increased sleeping cell score. Testing sleeping cell detection framework with reference data and postprocessing with Dominance Cell 2-Gram Deviation method demonstrates lower false-alarm rate than Dominance Cell Sub-Call Deviation, as it can be seen from Fig. 7b, d.

6.2.3 Detection based on dominance cell 2-gram symmetry deviation

This postprocessing method analyzes the symmetry imbalance of network event 2-grams. The symmetry imbalance is evaluated based on all sub-calls from training data set and sub-calls assigned to abnormal class from testing dataset. Information about the number of 2-grams directed to the cell, and from the cell is extracted from the training set. The considered 2-grams consist of events which sequentially occur in the dominance areas of 2 cells. It means that if in the training data, the number of handovers from Cell A to Cell B, and from Cell B to Cell A, is roughly the same, and in the testing set it is not, it can be concluded that symmetry of this particular 2-gram is skewed. Most common types of 2-grams which are analyzed with this method are related to handovers, e.g. A3—HO COMMAND sequences.

From Fig. 8 it can be seen that Dominance Cell 2-Gram Symmetry Deviation finds sleeping cell 1, while its neighboring cells 8, 9, 11 and 12 have suspiciously high sleeping cell score, if compared to other cells in the network.

Comparison of symmetry analysis method with two previously described postprocessing approaches shows that this method is very efficient in detecting sleeping cell and its neighbors. At the same time stability, i.e. false alarm rate, of this method is also very good, as it can be seen from Fig. 8b.

6.2.4 Detection based on target cell sub-calls

As it is discussed in Sect. 5.4, deviation between training and testing data is not calculated in this method. Extensive location information, like dominance map information, is
not required for sleeping cell detection with target cell sub-call method. The sleeping cell detection histogram, presented in Fig. 9, is constructed by counting all unique target cell IDs for each anomalous sub-call. It can be clearly seen that cell 1 is successfully detected. Neighboring cells 8, 9, 11 and 12 also contain indication of malfunction in this area, as it can be noticed from heat map, shown in Fig. 9b. For this method, the SC score of cell 1 is slightly lower than for the postprocessing methods, based on dominance cell deviation. On the other hand, target cell sub-call method is much simpler, and requires significantly less information about user event occurrence location.

6.3 Combined method of sleeping cell detection

The idea of this method is to create a cumulative sleeping cell detection histogram based on the results from all 4 postprocessing methods described above. The resulting amplified SC histogram is shown in Fig. 10. Cell 1 has sleeping cell score well over $\mu + 3 \sigma$ threshold. Neighboring cells 8, 9, 11, 12 also have increased sleeping cell scores comparing to other cells, though they do not exceed the $\mu + 3 \sigma$ threshold. Reference data used as testing also demonstrates stability of the combined approach – no false alarms are triggered. Though, it can be seen that usage of target cell sub-call method introduces some noise. It is important to note that postprocessing methods are applied with equal weights. However, it is possible to emphasize more accurate method by increasing its weight, and penalize the unreliable, by reducing its weight. Though, selection of optimal weights is a matter of a separate study and is not discussed in this article.

6.4 Comparison of algorithms and performance evaluation

The postprocessing methods discussed above have their own advantages and disadvantages. Traditional data mining metrics, discussed in Sect. 5.5.1, are applied for quantitative comparison of sleeping cell detection methods, Fig. 11a. Ideal performance is presented with the solid double black line, and corresponds to the maximum area of the hexagon. K-fold cross validation method is utilized to obtain statistically significant results. Figures 6, 7, 8, 9 and 10 show averaged values of the sleeping cell scores from all runs of K-fold separation. In some of the runs certain neighbors of cell 1 demonstrated sleeping cell scores...
higher than $3\sigma$ threshold. This results in non-zero false positive rates. Formally, according to the values of the metrics, Dominance Cell 2-gram Deviation and Dominance Cell Sub-call Deviation methods, demonstrate better performance than other postprocessing techniques. However, high false positive rate for Dominance Cell 2-gram Symmetry Deviation and Target Cell Sub-call methods does not necessarily mean that these methods are worse. The reason is that neighbors of cell 1 exceed the $3\sigma$ threshold. This happens because adjacent cells are not completely independent, and are affected by malfunction in one of the neighbors. Thus, Dominance Cell 2-gram Symmetry Deviation and Target Cell Sub-call methods can be treated as more sensitive than the others. The ROC curves of the designed sleeping cell detection algorithms are presented in Fig. 11b. True positive rate equals 1 for all postprocessing methods.

7 Conclusions

This article presents a novel sleeping cell detection framework based on knowledge mining paradigm. MDT reports are used for the detection of a random access channel malfunction in one of the network cells. Experimental setup implements a simulated LTE network, used to generate a diverse statistics base with several thousands of user calls and tens of thousands of MDT samples. Investigated failure case is a sleeping cell caused by RACH malfunction. Even though the studied problem is rather specific, the proposed framework does not consider any properties or peculiarities of the random access failure for the detection. Moreover, analysis of event sequences makes the presented method applicable to data collected with MDT, TRACE functionality, mobile quality agents, and any other method, which is capable to gather the user specific sequences of network events. The studied type of sleeping cell problem is rather complex, and detection of this problem has never been done before. The applicability of our sequential analysis to other network failures might be beneficial, but it has to be studied.
The designed knowledge mining framework is semi-supervised. From the perspective of SONs the proposed system has centralized architecture, but it can also be hybrid, with preprocessing and transformation stages done in distributed manner. The heart of the developed detection framework is the analysis of sequences with N-gram method in the series of user event-triggered measurement MDT reports. Data preprocessing with sliding window transformation method allows to make the statistics base more reliable through standardization of the input event sequences. 2-gram analysis is used to convert sequential data to numeric format in the new feature space. To simplify analysis of the data in the new space, dimensionality reduction with minor component analysis method is applied. K-NN anomaly score detection algorithm is used to find the outliers in the data. Using this information, anomalous data points are converted with postprocessing to the knowledge about location of the problematic regions in the network. Comparison of different location mapping postprocessing methods is done. Additionally, so called amplification is used to take into account neighbor relations between cells and network topology, for improvement of

Fig. 12 Heuristic performance comparison of algorithms a Distances in original—non amplified approach. b Distances in amplified approach. c Heuristic performance distances for problematic case. d Heuristic performance distances for reference case.
sleeping cell detection performance. As it can be seen, amplified sleeping cell score of truly problematic cell is higher than corresponding non-amplified score.

Results demonstrate that the developed suggested framework, based on sequence analysis, allows for efficient detection of the random access sleeping cell problem in the network. The projection of the data in the new space is such that accurate separation of normal and abnormal data points becomes possible. Evaluation shows that post-processing method named Dominance Cell 2-Gram Symmetry Deviation demonstrates the best combination of results, with respect to heuristic performance measure. According to the same metric, the proposed amplification approach, improves the detection quality of postprocessing methods. However, this approach is an additional element of the developed nontrivial framework and is not the most important outcome of our research.

Results of this work lay grounds and suggest exact methods for building advanced performance monitoring systems in modern mobile networks. One of the possible directions in this area is extensive usage of data mining techniques in general, and anomaly detection in particular. New systems of network maintenance would allow to address growing complexity and heterogeneity of modern mobile networks, and will help to meet the requirements of 5G.

Future work in this field includes validation of the developed system in more complex scenarios, detection of several or different types of malfunctions, and substitution of semi-supervised approach with unsupervised. The ultimate goal is to achieve accurate and timely detection of different sleeping cell types in highly dynamic mobile network environments. Obviously, low level of false alarms must be supported, and at the same time significant increase of computational complexity should be avoided.

Acknowledgments Authors would like to thank colleagues from Magister Solutions, Nokia and University of Jyväskylä for collaboration, their valuable feedback regarding this research, and peer reviews. Work on this study has been partly funded by MIPCOM project, Graduate School in Electronics, Telecommunications and Automation (GETA), and Doctoral Program in Computing and Mathematical Sciences (COMAS).

References

1. Amirijoo, M., Frenger, P., Gunnarsson, F., Moe, J., & Zetterberg, K. (2009). On self-optimization of the random access procedure in 3g long term evolution. In Integrated network management-workshops, 2009. IM ’09. IFIP/IEEE international symposium on, pp 177–184.
2. Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. In Proceedings of the 6th European conference on principles of data mining and knowledge discovery, Springer-Verlag, London, UK, UK, PKDD ’02, pp 15–26.
3. Barco, R., Lazaro, P., Diez, L., & Wille, V. (2008). Continuous versus discrete model in autodiagnosis systems for wireless networks. Mobile Computing, IEEE Transactions on, 7(6), 673–681. doi:10.1109/TMC.2008.23.
4. Barco, R., Lazaro, P., Wille, V., Diez, L., & Patel, S. (2009). Knowledge acquisition for diagnosis model in wireless networks. Expert Systems with Applications, 36(3, Part 1), 4745–4752.
5. Barco, R., Wille, V., Diez, L., & Toril, M. (2010). Learning of model parameters for fault diagnosis in wireless networks. Wireless Networks, 16(1), 255–271. doi:10.1007/s11276-008-0128-z.
6. Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18, 467–479.
7. Brueninghaus, K., Astely, D., Salzer, T., Visuri, S., Alexiou, A., Karger, S., & Seraji, G. A. (2005). Link performance models for system level simulations of broadband radio access systems. In IEEE 16th international symposium on personal, indoor and mobile radio communications, 2005. PIMRC 2005., vol 4, pp 2306–2311 Vol. 4, doi:10.1109/PIMRC.2005.1651855.
8. Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information Retrieval, pp 161–175.
9. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Survey, 41(3), 15:1–15:58.
10. Chernogorov, F. (2010). Detection of sleeping cells in long term evolution mobile networks. Master’s thesis, University of Jyväskylä, Finland.
11. Chernogorov, F., Turkka, J., Ristaniemi, T., & Averbuch, A. (2011). Detection of sleeping cells in LTE networks using diffusion maps. In Vehicular technology conference (VTC Spring), 2011 IEEE 73rd, pp 1–5.
12. Chernogorov, F., Brigatti, K., Ristaniemi, T., & Chernov, S. (2013). N-gram analysis for sleeping cell detection in LTE networks. In Proceedings of the 38th international conference on acoustics, speech, and signal processing (ICASSP).
13. Cheung, B., Kumar, G. N., & Rao, S. A. (2005). Statistical algorithms in fault detection and prediction: Toward a healthier network. Bell Labs Technical Journal, 9(4), 171–185.
14. Cheung, B., Fishkin, S. G., Kumar, G. N., & Rao, S.A. (2006a). Method of monitoring wireless network performance. Tech. rep., Los Angeles, CA, uS Patent 2006/0063521 A1, CN1755341A, EP1638253A1.
15. Cheung, B., Fishkin, S. G., Kumar G.N., & Rao, S. A. (2006b). Method of monitoring wireless network performance. US Patent 2006/0063521 A1, CN1755341A, EP1638253A1.
16. Choi, J., Kim, H., Choi, C., & Kim, P. (2011). Efficient malicious code detection using n-gram analysis and svm. In L. Barolli, F. Xhafa, & M. Takizawa (Eds.), NSbS (pp. 618–621). : IEEE Computer Society.
17. Ciocarlie, G., Lindqvist, U., Novaczki, S., & Sanneck, H. (2013). Detecting anomalies in cellular networks using an ensemble method. In Network and service management (CNSM), 2013 9th international conference on, pp 171–174. doi:10.1109/CNSM.2013.6727831.
18. Ciocarlie, G., Cheng, C. C., Connolly, C., Lindqvist, U., Nitz, K., Novaczki, S., Sannec, H., & Naseer-ul Islam, M. (2014a). Anomaly detection and diagnosis for automatic radio network verification. In 6th international conference on mobile networks and management, MONAMI 2014.
19. Ciocarlie, G., Cheng, C. C., Connolly, C., Lindqvist, U., Novaczki, S., Sannec, H., & Naseer-ul Islam, M. (2014b). Managing scope changes for cellular network-level anomaly detection. In Wireless communications systems (ISWCS), 2014 11th international symposium on, pp 375–379, doi:10.1109/ISWCS.2014.6933381.
20. Ciocarlie, G., Lindqvist, U., Nitz, K., Novaczki, S., & Sanneck, H. (2014c). On the feasibility of deploying cell anomaly detection in operational cellular networks. In Network operations and management symposium (NOMS), 2014 IEEE, pp 1–6, doi:10.1109/NOMS.2014.6838305.

21. Cisco Systems (2014) Cisco visual networking index: Global mobile data traffic forecast update 2014–2019 white paper. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

22. Coifman, R. R., & Lafon, S. (2006). Diffusion maps. Applied and Computational Harmonic Analysis, 21(1), 5–30.

23. David, G. (2009). Anomaly detection and classification via diffusion processes in hyper-networks. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel.

24. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Widener, T. (1996). The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM, 39, 27–34.

25. Federal Communications Commission (2011) Small entity compliance guide: Wireless E911 location accuracy requirements. Federal communications commission: Report and order FCC 10-176 PS docket No 07-114 p 3

26. Ganapathiraju, M., Weisser, D., Rosenfeld, R., Carbonell, J., Reddy, R., & Klein-Geven, J. (2002). Comparative n-gram analysis of whole-genome protein sequences. In Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA,HLT ’02, pp 76–81.

27. 3rd Generation Partnership Project. (2009a). Evolved universal terrestrial radio access network (e-utran); self-configuring and self-optimizing network (SON) use cases and solutions (release 9). Tech. Rep. TR 36.902, 3GPP

28. 3rd Generation Partnership Project. (2009b). Technical specification group radio access network: study on minimization of drive-tests in next generation networks (release 9). Tech. Rep. TR 36.805, 3GPP

29. 3rd Generation Partnership Project. (2010). 3GPP; TSG radio access network; further advancements for e-utra physical layer aspects (release 9). Tech. Rep. TR 36.814, 3GPP

30. 3rd Generation Partnership Project. (2011). Technical specification group radio access network; evolved universal terrestrial radio access (e-utra); radio resource control (rrc); protocol specification (release 10). Tech. Rep. TS 36.331, 3GPP

31. 3rd Generation Partnership Project. (2012). Technical report 3rd generation partnership project; technical specification group radio access network; evolved universal terrestrial radio access (e-utra); mobility enhancements in heterogeneous networks (release 11). Tech. rep., 3GPP

32. 3rd Generation Partnership Project. (2014). Self-organizing networks (SON); self-healing concepts and requirements (release 12). Tech. rep., 3GPP TS 32.541 V12.0.0

33. GSMA Intelligence (2014) Understanding 5G: Perspectives on future technological advancements in mobile. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking/index-vni-white_paper_c11-520862.html

34. Guillet, F., & Hamilton, H. J. (Eds.). (2007). Quality measures in data mining, studies in computational intelligence (Vol. 43). Berlin: Springer.

35. Haider, M., & O’Shaughnessy, D. (2012). Topic n-gram count language model adaptation for speech recognition. In Spoken language technology workshop (SLT), 2012 IEEE, pp 165–169, doi:10.1109/SLT.2012.6424216.

36. Hämäläinen, S., Sanneck, H., & Sartori, C. (2012). LTE self-organising networks (SON): Network management automation for operational efficiency (1st ed.). Hoboken: Wiley Publishing.

37. Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed., Vol. 54). Burlington: Morgan Kaufmann.

38. Hand, D. J., Smyth, P., & Mannila, H. (2001). Principles of data mining. Cambridge, MA, USA: MIT Press.

39. Hapsari, W., Umesh, A., Iwamura, M., Tomala, M., Gyula, B., & Sebire, B. (2012a). Minimization of drive tests solution in 3GPP. Communications Magazine, IEEE, 50(6), 28–36.

40. Hapsari, W., Umesh, A., Iwamura, M., Tomala, M., Gyula, B., & Sebire, B. (2012b). Minimization of drive tests solution in 3GPP. Communications Magazine, IEEE, 50(6), 28–36.

41. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. doi:10.1109/TKDE.2008.239.

42. He, Z., Cichocki, A., & Xie, S. (2009). Efficient method for tucker3 model selection. Electronics Letters, 45, 805.

43. He, Z., Cichocki, A., Xie, S., & Choi, K. (2010). Detecting the number of clusters in n-way probabilistic clustering. IEEE Transaction Pattern Analysis Machine Intelligence, 32(11), 2006–2021.

44. Holma, H., & Toskala, A. (2011). LTE for UMTS: Evolution to LTE-advanced (2nd ed.). Hoboken: Wiley Publishing.

45. Islam, A., & Inkpen, D. (2009). Real-word spelling correction using Google Web 1t n-gram with backoff. In Natural language processing and knowledge engineering, 2009. NLP-KE 2009. International conference on, pp 1 –8, doi:10.1109/NLPKE.2009.5313823

46. Johansson, J., Hapsari, W., Kelley, S., & Bodog, G. (2012). Minimization of drive tests in 3GPP release 11. Communications Magazine, IEEE, 50(11), 36–43.

47. Jolliffe, I. (2002). Principal component analysis. Springer series in statistics. Berlin: Springer.

48. Kac, M., Kiefer, J., & Wolfowitz, J. (1955). On tests of normality and other tests of goodness of fit based on distance methods. The Annals of Mathematical Statistics, 26(2), 189–211.

49. Kassis, E. (2010). Anomaly-based error detection in base station data. Master’s thesis, Tel-Aviv University, Israel

50. Kela, P. (2007). Downlink channel quality indication for evolved universal terrestrial radio access network. Master’s thesis, University of Jyväskylä, Finland.

51. Khanafer, R., Solana, B., Triola, J., Barco, R., Moltsen, L., Altman, Z., et al. (2008). Automated diagnosis for umts networks using bayesian network approach. Vehicular Technology, IEEE Transactions on, 57(4), 2451–2461. doi:10.1109/TVT.2007.912610.

52. Kolteinen, N. (2007). Downlink packet scheduling performance in evolved universal terrestrial radio access network. Master’s thesis, University of Jyväskylä, Finland.

53. Laiho, J., Raivio, K., Lehtimaki, P., Hatonen, K., & Simula, O. (2005). Advanced analysis methods for 3g cellular networks. Wireless Communications, IEEE Transactions on, 4(3), 930–942. doi:10.1109/TWC.2005.847088.

54. Luo, F. L., Unbehauen, R., & Cichocki, A. (1997). A minor component analysis algorithm. Neural Networks, 10(2), 291–297.

55. Mueller, C. M., Kaschub, M., Blankenhorn, C., & Wanke, S. (2008). A cell outage detection algorithm using neighbor cell list reports. In K. Hummel & J. Sterbenz (Eds.), Self-organizing systems (Vol. 5343, pp 218–229.), Lecture notes in computer science Berlin Heidelberg: Springer.

56. Nagao, Makoto, Mori, Shinsuke (1994) A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. In: Proceedings of the 15th conference on computational linguistics - volume 1, association for computational linguistics, Stroudsburg, PA, USA, COLING ’94, pp 611–615

57. Next Generation Mobile Networks (2008a) Recommendation on SON and O&M Requirements. Tech. rep., NGMN, URL: http://www.ngmn.org/
58. Next Generation Mobile Networks (2008b) Use Cases related to Self Organising Network, overall description. Tech. rep., NGMN, URL http://www.ngmn.org/
59. Next Generation Mobile Networks (2015) NGMN 5G Initiative White Paper. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1.0.pdf
60. Novaczki, S. (2013). An improved anomaly detection and diagnosis framework for mobile network operators. In Design of reliable communication networks (DRCN), 2013 9th international conference on the, pp 234–241.
61. Novaczki, S., & Szilagyi, P. (2011). Radio channel degradation detection and diagnosis based on statistical analysis. In Vehicular technology conference (VTC Spring), 2011 IEEE 73rd, pp 1–2.
62. NTT DOCOMO Inc (2014) Docomo 5g white paper: 5g radio access: Requirements, concept and technologies. https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/white_paper_5g/DOCOMO_5G_White_Paper.pdf
63. Osseiran, A., Braun, V., Hidekazu, T., Marsch, P., Schotten, H., Tullberg, H., Uusitalo, M., & Schellman, M. (2013). The foundation of the mobile and wireless communications system for 2020 and beyond: Challenges, enablers and technology solutions. In Vehicular technology conference (VTC Spring), 2013 IEEE 77th, pp 1–5, doi:10.1109/VTCSpring.2013.6692781.
64. Rabin, N. (2010). Data mining dynamically evolving systems via diffusion methodologies. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel.
65. Raivio, K., Simula, O., Laiho, J., & Lehtimaki, P. (2003). Analysis of mobile radio access network using the self-organizing map. In Integrated network management, 2003. IFIP/IEEE Eighth international symposium on, pp 439–451, doi:10.1109/INM.2003.1194197.
66. Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. SIGMOD Record, 29(2), 427–438.
67. Ramiro, J., & Hamied, K. (2012). Self-organizing networks (SON): Self-planning, self-optimization and self-healing for GSM, UMTS and LTE (1st ed.). Hoboken: Wiley Publishing.
68. Scully, N., et al. (2008). D2.1: Use cases for self-organising networks. URL http://www.fp7-socrates.eu
69. Sesia, S., Baker, M., & Toufik, I. (2011). LTE - The UMTS long term evolution: From theory to practice. Hoboken: Wiley.
70. Szilagyi, P., & Novaczki, S. (2012). An automatic detection and diagnosis framework for mobile communication systems. IEEE Transactions on Network and Service Management, 9(2), 184–197.
71. Turkka, J., Ristami, T., David, G., & Averbuch, A. (2011). Anomaly detection framework for tracing problems in radio networks. In The 10th international conference on networks, ICN 2011.
72. Turkka, J., Chernogorov, F., Brigatti, K., Ristami, T., & Lempiainen, J. (2012). An approach for network outage detection from drive-testing databases. Journal of Computer Networks and Communications.
73. Yilmaz, O. N. C., Hamalainen, J., & Hamalainen, S. (2011). Self-optimization of random access channel in 3rd generation partnership project long term evolution. Wireless Communications and Mobile Computing, 11(12), 1507–1517.

Sergey Chernov received his M.Sc. degree with honors in radio physics and electronics in 2008 in Yaroslavl State University, Russia. Since 2011 he is Ph.D. student at the faculty of informational technologies in the University of Jyvaskyla, Finland. His current research is focused on Self-Organizing Network concept of Long-Term Evolution mobile cellular networks and the application of data mining and machine learning algorithms to the various radio network data.

Fedor Chernogorov received his M.Sc. degree with honors in Telecommunications in 2009 from P.G. Demidov Yaroslavl State University, Yaroslavl, Russia. From the University of Jyvaskyla, Finland he received his M.Sc. in Mobile Technology in 2010, and his Ph.D. in Mathematical Information Technology in 2015. Since 2011 he is with Magister Solutions, Jyvaskyla, Finland, where he currently works at position of Senior Researcher. He has authored and co-authored 9 conference and 2 journal publications in the fields of cellular mobile communications and data mining, anomaly detection. Areas of his interest are self-organizing and cognitive mobile networks, advanced performance monitoring in cellular, knowledge mining, anomaly detection and data mining.

Kimmo Brigatti received his M.Sc. (Information Technology) in 2011 from University of Jyvaskyla, Jyvaskyla, Finland. He has been involved in different research groups on data mining, anomaly detection and wireless network technologies starting from 2009 in University of Jyvaskyla. Kimmo has co-authored two publications N-Gram Analysis For Sleeping Cell Detection in LTE Networks (ICASSP, 2013) and An Approach for Network Outage Detection from Drive-Testing Databases (Journal of Computer Networks and Communications, 2012).
Tapani Ristaniemi (SM’11) received his M.Sc. degree in mathematics in 1995, the Ph.Lic. degree in applied mathematics in 1997, and the Ph.D. degree in wireless communications in 2000, all from the University of Jyväskylä, Jyväskylä, Finland. In 2001, he was appointed as a Professor in the Department of Mathematical Information Technology, University of Jyväskylä. In 2004, he moved to the Department of Communications Engineering, Tampere University of Technology, Tampere, Finland, where he was appointed as a Professor of Wireless Communications. In 2006, he moved back to the University of Jyväskylä to take up his appointment as a Professor of Computer Science. He is an Adjunct Professor of Tampere University of Technology. In 2013, he was a Visiting Professor in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He has authored or co-authored over 200 publications in journals, conference proceedings, and invited sessions. He served as a Guest Editor of IEEE WIRELESS COMMUNICATIONS in 2011 and currently he is an Editorial Board Member of Wireless Networks and the International Journal of Communication Systems. His research interests are in the areas of brain and communication signal processing and wireless communication systems. Besides academic activities, Professor Ristaniemi is also active in the industry. In 2005, he co-founded a start-up, Magister Solutions, Ltd., in Finland, specializing in wireless systems (R&D) for telecom and space industries in Europe. Currently, he serves as a consultant and a Chairman of the Board.